Estratégia Concursos

Aula 00

TCE-RJ (Auditor de Controle Externo - Controle Externo) Análise de Dados e Informações (Parte Específica)

Autor:
Diego Carvalho, Equipe Informática 2 (Diego Carvalho)

19 de Novembro de 2024

Índice

APRESENTAÇÃO DO PROFESSOR

PROF. DIEGO CARVALHO

FORMADO EM CIÊNCIA DA COMPUTAÇÃO PELA UNIVERSIDADE DE BRASÍLIA (UNB), PÓS-GRADUADO EM GESTÃO DE TECNOLOGIA DA INFORMAÇÃO NA ADMINISTRAÇÃO PÚBLICA E, ATUALMENTE, AUDITOR FEDERAL DE FINANÇAS E CONTROLE DA SECRETARIA DO TESOURO NACIONAL.

ESTRATÉGIA CONCURSOS

PROFESSOR DIEGO CARVALHO - WWW.INSTAGRAM.COM/PROFESSORDIEGOCARVALHO

Sobre o curso: galera, todos os tópicos da aula possuem Faixas de Incidência, que indicam se o assunto cai muito ou pouco em prova. Diego, se cai pouco para que colocar em aula? Cair pouco não significa que não cairá justamente na sua prova! A ideia aqui é: se você está com pouco tempo e precisa ver somente aquilo que cai mais, você pode filtrar pelas incidências média, alta e altíssima; se você tem tempo sobrando e quer ver tudo, vejam também as incidências baixas e baixíssimas. *Fechado?*

INCIDÊNCIA EM PROVA: BAIXÍSSIMA

INCIDÊNCIA EM PROVA: BAIXA

INCIDÊNCIA EM PROVA: MÉDIA

INCIDÊNCIA EM PROVA: ALTA

INCIDÊNCIA EM PROVA: ALTÍSSIMA

Além disso, essas faixas não são por banca – é baseado tanto na quantidade de vezes que caiu em prova independentemente da banca quanto nas minhas próprias avaliações sobre cada assunto.

#ATENÇÃO

Avisos Importantes

O curso abrange todos os níveis de conhecimento...

Esse curso foi desenvolvido para ser acessível a **alunos com diversos níveis de conhecimento diferentes.** Temos alunos mais avançados que têm conhecimento prévio ou têm facilidade com o assunto. Por outro lado, temos alunos iniciantes, que nunca tiveram contato com a matéria ou até mesmo que têm trauma dessa disciplina. A ideia aqui é tentar atingir ambos os públicos - iniciantes e avançados - da melhor maneira possível..

Por que estou enfatizando isso?

O **material completo** é composto de muitas histórias pessoais, exemplos, metáforas, piadas, memes, questões, desafios, esquemas, diagramas, imagens, entre outros. Já o **material simplificado** possui exatamente o mesmo núcleo do material completo, mas ele é menor e mais objetivo. *Professor, eu devo estudar por qual material?* Se você quiser se aprofundar nos assuntos ou tem dificuldade com a matéria, necessitando de um material mais passo-a-passo, utilize o material completo. Se você não quer se aprofundar nos assuntos ou tem facilidade com a matéria, necessitando de um material mais direto ao ponto, utilize o material simplificado.

Por fim...

O curso contém diversas questões espalhadas em meio à teoria. Essas questões possuem um comentário mais simplificado porque **têm o único objetivo de apresentar ao aluno como bancas de concurso cobram o assunto previamente administrado**. A imensa maioria das questões para que o aluno avalie seus conhecimentos sobre a matéria estão dispostas ao final da aula na lista de exercícios e **possuem comentários bem mais abrangentes**.

APRESENTAÇÃO DO TÓPICO

Fala, galera! O assunto do nosso tópico é **Mineração de Dados (Data Mining)**! *Pessoal, o que um mineiro faz? Faz pão de queijo e feijão tropeiro, Diego!* Não estou me referindo a esse mineiro, zé mané! Estou me referindo àquele que trabalha com mineração! Ele escava grandes áreas de terra em busca de minérios, combustíveis ou até pedras preciosas. O mineiro do mundo moderno escava bases de dados em busca de dados relevantes para uma organização e esse é o tema de hoje :)

PROFESSOR DIEGO CARVALHO - WWW.INSTAGRAM.COM/PROFESSORDIEGOCARVALHO

Galera, todos os tópicos da aula possuem Faixas de Incidência, que indicam se o assunto cai muito ou pouco em prova. *Diego, se cai pouco para que colocar em aula?* Cair pouco não significa que não cairá justamente na sua prova! A ideia aqui é: se você está com pouco tempo e precisa ver somente aquilo que cai mais, você pode filtrar pelas incidências média, alta e altíssima; se você tem tempo sobrando e quer ver tudo, vejam também as incidências baixas e baixíssimas. *Fechado?*

INCIDÊNCIA EM PROVA: BAIXÍSSIMA
INCIDÊNCIA EM PROVA: BAIXA
INCIDÊNCIA EM PROVA: MÉDIA
INCIDÊNCIA EM PROVA: ALTA
INCIDÊNCIA EM PROVA: ALTÍSSIMA

Além disso, essas faixas não são por banca – é baseado tanto na quantidade de vezes que caiu em prova independentemente da banca e também em minhas avaliações sobre cada assunto...

#ATENÇÃO

Avisos Importantes

O curso abrange todos os níveis de conhecimento...

Esse curso foi desenvolvido para ser acessível a **alunos com diversos níveis de conhecimento diferentes.** Temos alunos mais avançados que têm conhecimento prévio ou têm facilidade com o assunto. Por outro lado, temos alunos iniciantes, que nunca tiveram contato com a matéria ou até mesmo que têm trauma dessa disciplina. A ideia aqui é tentar atingir ambos os públicos - iniciantes e avançados - da melhor maneira possível..

Por que estou enfatizando isso?

O **material completo** é composto de muitas histórias, exemplos, metáforas, piadas, memes, questões, desafios, esquemas, diagramas, imagens, entre outros. Já o **material simplificado** possui exatamente o mesmo núcleo do material completo, mas ele é menor e bem mais objetivo. *Professor, eu devo estudar por qual material?* Se você quiser se aprofundar nos assuntos ou tem dificuldade com a matéria, necessitando de um material mais passo-a-passo, utilize o material completo. Se você não quer se aprofundar nos assuntos ou tem facilidade com a matéria, necessitando de um material mais direto ao ponto, utilize o material simplificado.

Por fim...

O curso contém diversas questões espalhadas em meio à teoria. Essas questões possuem um comentário mais simplificado porque **têm o único objetivo de apresentar ao aluno como bancas de concurso cobram o assunto previamente administrado**. A imensa maioria das questões para que o aluno avalie seus conhecimentos sobre a matéria estão dispostas ao final da aula na lista de exercícios e **possuem comentários bem mais completos, abrangentes e direcionados.**

MINERAÇÃO DE DADOS

Contextualização

INCIDÊNCIA EM PROVA: ALTÍSSIMA

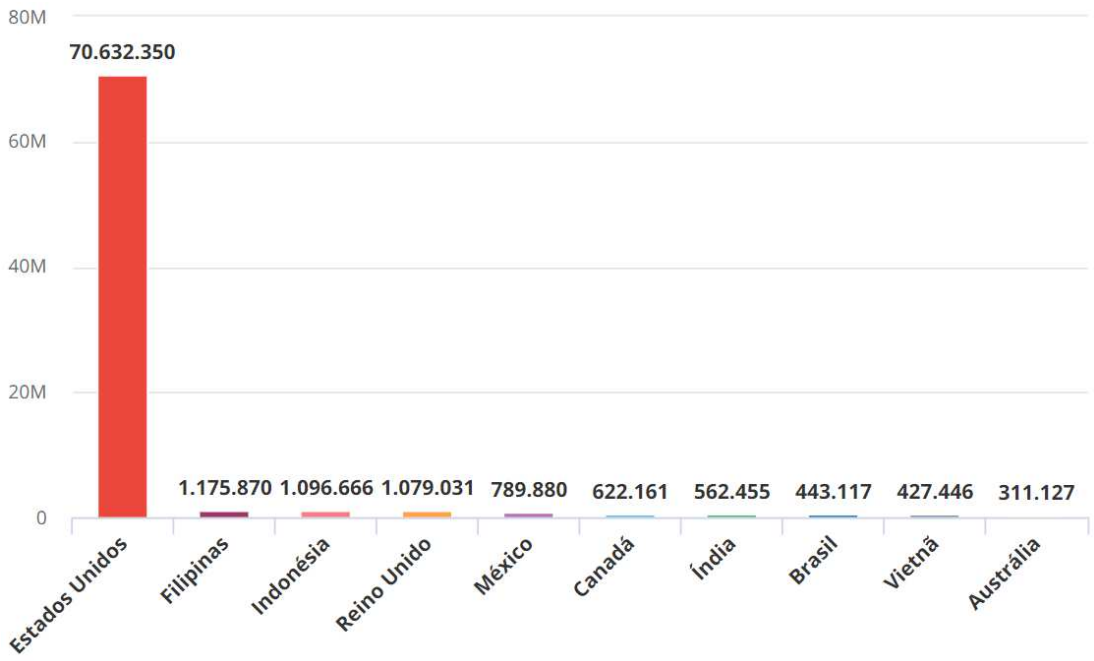
Galera, vamos falar agora sobre dados! *Professor, eu não aguento mais falar sobre dados!* Eu sei, mas fazer o que se esse assunto está na moda: *Data Warehouse*, *Data Mart*, *Big Data*, *Data Analytics*, *Data Privacy*, *Data Lake*, *Data Security*... e a estrela da nossa aula de hoje: Data Mining! **Também chamada de Mineração de Dados ou Prospecção de Dados, trata-se do processo de explorar grandes quantidades de dados à procura de padrões consistentes.**

Nada melhor para explicar um conceito do que por meio de exemplos. E o exemplo que eu vou utilizar é do escândalo da Cambridge Analytica. *Que fofoca é essa que eu não fiquei sabendo, Diego?* **Eu vou te contar: uma empresa britânica supostamente especializada em consultoria política conseguiu obter de forma ilegal dados sobre mais de oitenta milhões de usuários do Facebook sem consentimento**.

Esses dados foram utilizados em vários países para verificar o perfil de eleitores e influenciar suas opiniões no intuito de ajudar políticos a vencerem eleições. Por meio de consultas às páginas curtidas, data de nascimento, cidade, sexo, etc, uma aplicação de mineração de dados conseguiu traçar perfis psicológicos dessas pessoas e criar campanhas ou propagandas direcionadas de forma mais eficaz para influenciar em suas convicções políticas.

Essa imagem acima é do Mark Zuckerberg – proprietário do Facebook – prestando esclarecimentos em uma sessão de questionamentos no congresso americano. Ele foi ao Congresso para esclarecer como o Facebook reagiu ao vazamento de dados de 87 milhões de

pessoas pela consultoria política Cambridge Analytica e como sua empresa de tecnologia trabalha para proteger os dados de seus usuários. *Professor, o Brasil estava na lista?*



Olha lá... talvez seus dados tenham sido analisados pela Cambridge Analytica. *Professor, essa influência resultou em algo?* **Bem, há suspeitas de que a empresa britânica teria sido utilizada para ajudar Donald Trump a vencer as eleições presidenciais norte-americanas de 2016, mas isso é outro papo!** O que importa aqui é que técnicas similares são utilizadas por empresas como Twitter, Google ou Amazon para descobrir o que nós queremos ver ou comprar.

E isso não é utilizado apenas em anúncios e política: a mineração de dados possibilita que companhias aéreas prevejam quem perderá um voo; é capaz de informar a grandes lojas de departamento quem possivelmente está grávida; ajuda médicos a identificarem infecções fatais; e impressionantemente podem ser utilizadas até para prever – por meio de dados celulares – possíveis atentados terroristas em diversos países.

O poder da mineração de dados e todo o enfoque que está tendo atualmente podem fazer parecer que é uma varinha mágica capaz de salvar uma empresa ou destruir uma democracia. Não é nada disso: é simplesmente a aplicação de técnicas estatísticas capazes de fazer uma varredura em uma quantidade massiva de dados em busca de padrões impossíveis de serem detectados por seres humanos.

Esses padrões não são baseados na intuição humana, mas no que os dados nos sugerem. Isso pode parecer uma ideia revolucionária, mas nada mais é do que a previsão do tempo faz há décadas (aliás, a mineração de dados é muito parecida com a meteorologia). Os meteorologistas basicamente buscam duas coisas: primeiro, eles querem descrever padrões climáticos genéricos; segundo, eles querem prever a temperatura e umidade de um dia qualquer.

De forma similar, os cientistas de dados do Spotify podem estar interessados em descobrir o perfil de pessoas que curtem rock pesado e sugerir músicas de gêneros similares. O grande lance aqui é descrever e prever algo não através de um estudo sociológico realizado por grandes especialistas em pesquisas de mestrado, por exemplo, mas simplesmente analisar uma quantidade massiva de dados. *Bacana?*

E como isso poderia ser feito no Spotify? Poderiam ser analisados padrões de gravadoras, reviews publicados na Internet, parcerias entre músicos, idade, localização, entre outros. A mineração de dados é mais sobre identificar padrões do que sobre explicá-los. *O que isso significa?* **Isso significa que muitas vezes você encontrará um padrão específico, mas não fará ideia do que ele significa.** Vejam que legal...

Vou contar para vocês uma lenda que existe desde a década de noventa e que eu ouvi quando estava na faculdade: fraldas x cervejas! Você vai ao mercado comprar uma pasta de dente e do lado tem um... fio dental! Você aproveita e decide comprar também um espaguete e do lado tem um... molho de tomate! *Professor, não há nada de interessante nisso – você coloca produtos que geralmente são utilizados juntos um ao lado do outro.*

É verdade, mas e se eu chego em um supermercado e, ao lado da seção de fraldas, existe uma pilha de caixas de cerveja? Olhando assim não há nenhuma lógica, mas vamos imaginar o seguinte cenário: João é casado com Maria! Ambos passaram em um concurso público, adquiriram uma estabilidade financeira e tiver um filho, que é recém-nascido). Um dia, João vem voando do trabalho para fazer um esquenta antes do jogo do maior time da história: **Flamengo**.

Ele chega em casa, toma um banho e vai correndo para frente da televisão para assistir o Mengão dar aquela surra tradicional no adversário. Enquanto isso, Maria vai ao supermercado, depois busca o filho na creche e, por fim, volta para casa cheia de sacolas e o bebê! Só que ao retirar os itens das sacolas, ela percebe que esqueceu de comprar as fraldas. Ela diz para João: "*Amor, vá ao supermercado comprar fraldas porque eu esqueci de comprar e não temos mais, por favor*".

João argumenta dizendo que faltam vinte minutos para começar o jogo, mas não convence sua esposa. *O que ele faz?* Pega o carro e sai voado até o supermercado mais próximo. Só que ele chega na seção onde se encontram as fraldas e percebe que do lado tem uma pilha de caixas de cervejas. Ele pensa: "*Poxa! Hoje é quarta-feira, trabalhei o dia todo, estou cansado, meu time vai jogar, eu estou aqui comprando fralda... eu acho que mereço tomar uma cervejinha!*"

E João volta feliz e contente para assistir ao jogo! *Pessoal, por que essa história é, na verdade, uma lenda? Porque o Flamengo não ganha de ninguém, professor!* Que isso? Não façam piadas com meu Mengão! Galera, essa história é uma lenda porque ela não é totalmente verdade! **Uma farmácia[1] americana realmente identificou essa relação, mas não foi baseado na análise de dados genérica, foi apenas uma pessoa que ficou curiosa e decidiu pesquisar exatamente essa relação.**

[1] *Cerveja na farmácia, professor?* Sim, farmácias americanas parecem um shopping center – vocês podem encontrar de tudo lá! ;)

Por outro lado, a ideia aqui é mostrar que essa correlação seria possível por meio de ferramentas de mineração de dados. E, retornando a ideia anterior, em muitos casos você encontrará correlações que são acidentais e em outros casos você encontrará correlações, mas não conseguirá explicá-las – como vimos no caso das fraldas e cervejas. **Uma rede varejista, por exemplo, descobriu que a venda de colírios aumentava na véspera de feriados**. *Por que, professor?* Ninguém sabe! Ainda não encontraram uma resposta para isso, de todo modo essa rede passou a preparar seus estoques e promoções com base nesse cenário. A Sprint, uma das líderes no mercado americano de telefonia, desenvolveu – com base no seu armazém de dados – um método capaz de prever com 61% de confiança se um consumidor trocaria de companhia telefônica dentro de um período de dois meses.

Com um marketing agressivo, ela conseguiu evitar o cancelamento de 120.000 clientes e uma perda de 35 milhões de dólares em faturamento. *Professor, dá um exemplo brasileiro aí?* Claro, tem um exemplo interessantíssimo da SEFAZ-AM! **Com o uso de mineração de dados, foram verificadas correlações entre o estado civil e salários se servidores da Secretaria de Fazenda do Estado do Amazonas**. *Como assim, professor?*

Notou-se que cerca de 80% dos servidores de maior poder aquisitivo deste órgão eram divorciados, enquanto em outras instituições (Ex: Secretaria de Educação) esta média de divorciados era inferior a 30%. *Qual seria a explicação para esse fenômeno?* **Os dados parecem sugerir que servidores com maior poder aquisitivo se envolvem mais em relações extraconjugais, resultando geralmente no término do casamento.**

Como na Secretaria de Fazenda do Amazonas há servidores geralmente com maior poder aquisitivo do que na Secretaria de Educação (que é composta basicamente por professores), a explicação parece válida. **Outra interpretação possível seria a de que, com poder aquisitivo individual mais elevado, não haveria razão para manter um casamento que não estivesse feliz**. Logo, prevalece a máxima de que é melhor só do que mal acompanhado...

Definições e Terminologias

INCIDÊNCIA EM PROVA: ALTÍSSIMA

O termo Mineração de Dados remonta à década de 1980, quando seu objetivo era extrair conhecimento dos dados. **Nesse contexto, o conhecimento é definido como padrões interessantes que são geralmente válidos, novos, úteis e compreensíveis para os seres humanos.** Se os padrões extraídos são interessantes ou não, depende de cada aplicação específica e precisa ser verificado pelos especialistas dessas aplicações.

Já o termo Análise de Dados/Informações tornou-se popular no início dos anos 2000. **A análise de dados é definida como a aplicação de softwares para a análise de grandes conjuntos de dados para o suporte de decisões.** A análise de dados é um campo muito interdisciplinar que adotou aspectos de muitas outras disciplinas científicas, como estatística, teoria de sinais, reconhecimento de padrões, inteligência computacional, aprendizado de máquina e pesquisa operacional.

A Análise de Dados é uma ferramenta importante para entender tendências e projeções, permitindo identificar padrões, correlações e relacionamentos entre diferentes pontos de dados e fornecer informações sobre os dados que podem ser usados para fazer previsões e projeções. **A análise de dados pode ser usada para analisar tendências passadas e fazer projeções sobre tendências futuras, bem como para identificar áreas de melhoria potencial e áreas de risco.**

Ela também pode ser usada para avaliar a eficácia de uma estratégia ou para identificar áreas de melhoria. Ao analisar tendências e projeções, as empresas podem entender melhor seus mercados e clientes e tomar melhores decisões sobre como posicionar seus produtos e serviços. Mais recentemente ainda, surgiram com mais uma palavra da moda: ***Analytics***! Trata-se do processo sistemático de coletar, analisar e interpretar dados a fim de obter insights e tomar decisões.

Honestamente, isso é tudo jogada de marketing. De tempos em tempos, alguém cria uma palavra da moda para revolucionar o mundo de tecnologia da informação, mas é tudo quase a mesma coisa. **Logo, nós vamos nos ater ao que efetivamente mais cai em prova: Data Mining (Mineração de Dados). É importantíssimo saber esse conceito, visto que as bancas adoram cobrar maneiras diferentes de se definir mineração de dados**. Vamos lá...

DEFINIÇÕES DE DATA MINING

Data Mining é o processo de explorar grande quantidade de dados para extração não-trivial de informação implícita desconhecida.

Palavras-chave: exploração; informação implícita desconhecida.

Data Mining é uso de teorias, métodos, processos e tecnologias para organizar uma grande quantidade de dados brutos para identificar padrões de comportamentos em determinados públicos.

Palavras-chave: teorias; métodos; processos; tecnologias; organizar dados brutos; padrões de comportamentos.

Data Mining é a categoria de ferramentas de análise denominada open-end e que permite ao usuário avaliar tendências e padrões não conhecidos entre os dados.

Palavras-chave: ferramenta de análise; open-end; tendências e padrões.

Data Mining é o processo de descoberta de novas correlações, padrões e tendências entre as informações de uma empresa, por meio da análise de grandes quantidades de dados armazenados em bancos de dados usando técnicas de reconhecimento de padrões, estatísticas e matemáticas.

Palavras-chave: descoberta; correlações; padrões; tendências; reconhecimento de padrões; estatística; matemática.

Data Mining constitui em uma técnica para a exploração e análise de dados, visando descobrir padrões e regras, a princípio ocultos, importantes à aplicação.

Palavras-chave: exploração e análise de dados; padrões; regras; ocultos.

Data Mining é o conjunto de ferramentas que permitem ao usuário avaliar tendências e padrões não conhecidos entre os dados. Esses tipos de ferramentas podem utilizar técnicas avançadas de computação como redes neurais, algoritmos genéticos e lógica nebulosa (fuzzy), dentre outras.

Palavras-chave: tendências; padrões; redes neurais; algoritmos genéticos; lógica nebulosa.

Data Mining é o conjunto de ferramentas e técnicas de mineração de dados que têm por objetivo buscar a classificação e o agrupamento (clusterização) de dados, bem como identificar padrões.

Palavras-chave: classificação; agrupamento; clusterização; padrões.

Data Mining é o processo de explorar grandes quantidades de dados à procura de padrões consistentes com o intuito de detectar relacionamentos sistemáticos entre variáveis e novos subconjuntos de dados.

Palavras-chave: padrões; relacionamentos.

Data Mining consiste em explorar um conjunto de dados visando a extrair ou a ajudar a evidenciar padrões, como regras de associação ou sequências temporais, para detectar relacionamentos entre estes.

Palavras-chave: exploração; padrões; regras; associação; sequência temporal; detecção.

Data Mining são ferramentas que utilizam diversas técnicas de natureza estatística, como a análise de conglomerados (*cluster analysis*), que tem como objetivo agrupar, em diferentes conjuntos de dados, os elementos identificados como semelhantes entre si, com base nas características analisadas.

Palavras-chave: estatística; análise de conglomerados; agrupamento.

Data Mining é o conjunto de técnicas que, envolvendo métodos matemáticos e estatísticos, algoritmos e princípios de inteligência artificial, tem o objetivo de descobrir relacionamentos significativos entre dados armazenados em repositórios de grandes volumes e concluir sobre padrões de comportamento de clientes de uma organização.

Palavras-chave: métodos matemáticos e estatístico; inteligência artificial; relacionamentos; padrões; comportamentos.

Data Mining é o processo de explorar grandes quantidades de dados à procura de padrões consistentes, como regras de associação ou sequências temporais, para detectar relacionamentos sistemáticos entre variáveis, detectando assim novos subconjuntos de dados.

Palavras-chave: padrões; regras de associação; sequências temporais; relacionamentos.

Data Mining é o processo de identificar, em dados, padrões válidos, novos, potencialmente úteis e, ao final, compreensíveis.

Palavras-chave: padrões; utilidade.

Data Mining é um método computacional que permite extrair informações a partir de grande quantidade de dados.

Palavras-chave: extração.

Data Mining é o processo de explorar grandes quantidades de dados à procura de padrões consistentes, como regras de associação ou sequências temporais.

Palavras-chave: exploração; padrões consistentes; regras de associação; sequência temporal.

Data Mining é o processo de analisar de maneira semi-automática grandes bancos de dados para encontrar padrões úteis.

Palavras-chave: padrões.

Professor, não entendi a utilidade de ver tantas definições diferentes. Galera, eu coloquei todas as definições porque TODAS elas caíram em prova (sem exceção!). Eu retirei todas essas definições de provas anteriores. **Notem como é importante saber de forma abrangente as possíveis definições de um conceito teórico importante**. Dito isso, vamos tentar condensar todos esses conceitos em uma grande definição a seguir:

> *Data Mining – Mineração de Dados – é um conjunto de processos, métodos, teorias, ferramentas e tecnologias open-end utilizadas para explorar, organizar e analisar de forma automática ou semi-automática[2] uma grande quantidade de dados brutos com o intuito de identificar, descobrir, extrair, classificar e agrupar informações implícitas desconhecidas, além de avaliar correlações, tendências e padrões consistentes de comportamento potencialmente úteis – como regras de associação ou sequências temporais – de forma não-trivial por meio de técnicas estatísticas e matemáticas, como redes neurais, algoritmos genéticos, inteligência artificial, lógica nebulosa, análise de conglomerados (clusters), entre outros.*

Pronto! Essa é uma definição extremamente abrangente com diversas palavras-chave que remetem à mineração de dados. Sabendo isso, vocês já saberão responder a grande maioria das questões de prova. É importante mencionar também que – apesar de geralmente ser utilizada em conjunto com Data Warehouses[3] – não é obrigatório que o seja! Você pode aplicar técnicas de mineração em diversos outros contextos (inclusive bases de dados transacionais).

É importante mencionar que a mineração de dados necessita, por vezes, utilizar processamento paralelo para dar conta da imensa quantidade de dados a serem analisados. Aliás, as ferramentas de mineração geralmente utilizam uma arquitetura cliente/servidor ou até uma arquitetura web. Outra pergunta que sempre me fazem é se é necessário ser um grande analista de dados para utilizar essas ferramentas de mineração. A resposta é: Não!

Usuários podem fazer pesquisas sem necessariamente saber detalhes da tecnologia, programação ou estatística. Por fim, a mineração de dados pode ser aplicada a uma grande variedade de contextos de tomada de decisão de negócios a fim de obter vantagens competitivas estratégicas. **Em particular, algumas áreas de ganhos significativos devem incluir as seguintes: marketing, finanças, manufatura e saúde**.

---

[2] Sobre esse ponto, há uma polêmica: alguns examinadores consideram que é de forma automática e outros consideram que é de forma semiautomática. Por vezes, examinadores de uma mesma banca possuem entendimentos diferentes. *E o que levar para a prova, professor?* Pessoalmente, eu entendo que seja semiautomática, mas eu levaria para a prova que pode ser automática ou semiautomática e – caso essa questão venha a ser cobrada – já se preparar para possíveis recursos.

[3] Data Warehouse (Armazém de Dados) é um repositório central de dados de várias fontes, projetado para relatórios e análises. Ele armazena dados atuais e históricos em um único local, criado pela integração de dados de várias fontes e processos e são projetados para dar suporte às decisões de negócios, permitindo que os dados sejam analisados de várias perspectivas. Eles também são usados para fornecer aos usuários uma visão abrangente dos dados, facilitando a análise de conjuntos complexos de informações.

CARACTERÍSTICAS DE MINERAÇÃO DE DADOS
Há diferentes tipos de mineração de dados: (1) diagnóstica, utilizada para entender os dados e/ou encontrar causas de problemas; (2) preditiva, utilizada para antecipar comportamentos futuros.
As provas vão insistir em afirmar que a mineração de dados só pode ocorrer em bancos de dados muito grades como Data Warehouses, mas isso é falso – apesar de comum, não é obrigatório.
Em geral, ferramentas de mineração de dados utilizam uma arquitetura web cliente/servidor, sendo possível realizar inclusive a mineração de dados de bases de dados não estruturadas.
Não é necessário ter conhecimentos de programação para realizar consultas, visto que existem ferramentas especializadas que auxiliam o usuário final de negócio.

(TCU – 2015) No ambiente organizacional, devido à grande quantidade de dados, não é recomendado o emprego de data mining para atividades ligadas a marketing.

_____

Comentários: é recomendável – sim – o emprego de Data Mining para atividades ligadas a marketing (Errado).

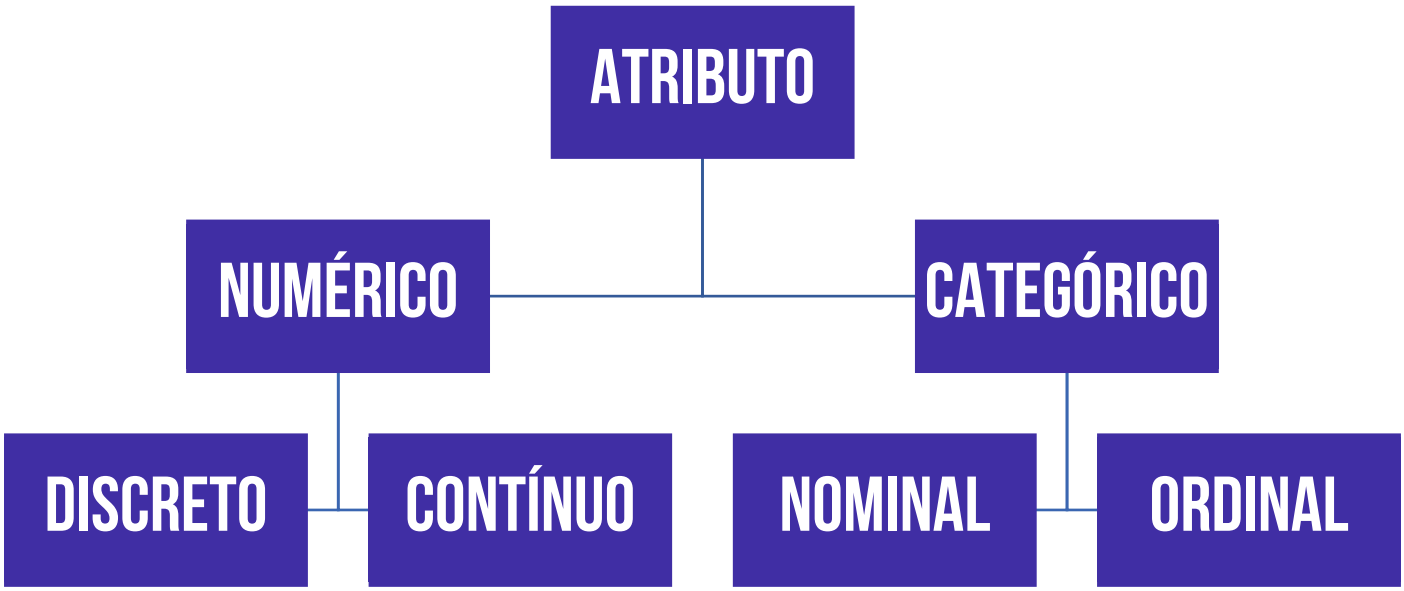
Antes de seguir, vamos ver algumas definições que serão importantes para o entendimento da aula. Primeiro, vamos tratar da classificação de dados quanto à estrutura:

TIPOS DE DADOS	DESCRIÇÃO
DADOS ESTRUTURADOS	São aqueles que residem em campos fixos de um arquivo (Ex: tabela, planilha ou banco de dados) e que dependem da criação de um modelo de dados, isto é, uma descrição dos objetos juntamente com as suas propriedades e relações.
DADOS SEMIESTRUTURADOS	São aqueles que não possuem uma estrutura completa de um modelo de dados, mas também não é totalmente desestruturado. Em geral, são utilizados marcadores (*tags*) para identificar certos elementos dos dados, mas a estrutura não é rígida.
DADOS NÃO ESTRUTURADOS	São aqueles que não possuem um modelo de dados, que não está organizado de uma maneira predefinida ou que não reside em locais definidos. Eles costumam ser de difícil indexação, acesso e análise (Ex: imagens, vídeos, sons, textos livres, etc).

Nessa aula, vamos nos focar mais nos dados estruturados, isto é, aqueles em que cada linha de uma tabela geralmente corresponde a um objeto (também chamado de exemplo, instância, registro ou vetor de entrada) e cada coluna corresponde a um atributo (também chamado de característica, variável ou *feature*). Nesse contexto, podemos ver na tabela seguinte uma classificação de atributos quanto à sua dependência:

ATRIBUTO DEPENDENTE	Representa um atributo de saída que desejamos manipular em um experimento de dados (também chamado de variável alvo ou variável target).
ATRIBUTO INDEPENDENTE	Representa um atributo de entrada que desejamos registrar ou medir em um experimento de dados.

É muito parecido com uma função matemática, por exemplo: y = ax + b. Nesse caso, x seria a variável de entrada (independente) e y seria a variável de saída (dependente).



Por fim, podemos ver no diagrama acima e na tabela abaixo como os atributos se classificam em relação ao seu valor:

TIPOS DE ATRIBUTOS	DESCRIÇÃO
ATRIBUTO NUMÉRICO	Também chamado de atributo quantitativo, é aquele que pode ser medido em uma escala quantitativa, ou seja, apresenta valores numéricos que fazem sentido.
DISCRETO	Os valores representam um conjunto finito ou enumerável de números, e que resultam de uma contagem (Ex: número de filhos, número de bactérias por amostra, número de logins em uma página web, entre outros).
CONTÍNUO	Os valores pertencem a um intervalo de números reais e representam uma mensuração (Ex: altura de uma pessoa, peso de uma marmita, salário de um servidor público, entre outros).

TIPOS DE ATRIBUTOS	DESCRIÇÃO
ATRIBUTO CATEGÓRICO	Também chamado de atributo qualitativo, é aquele que pode assumir valores categóricos, isto é, representam uma classificação.
NOMINAL	São aquelas em que não existe uma ordenação própria entre as categorias (Ex: sexo, cor dos olhos, fumante/não fumante, país de origem, profissão, religião, raça, time de futebol, entre outros).
ORDINAL	São aquelas em que existe uma ordenação própria entre as categorias (Ex: Escolaridade (1º, 2º, 3º Graus), Estágio de Doença (Inicial, Intermediário, Terminal), Classe Social (Classe Baixa, Classe Média, Classe Alta), entre outros)

Principais Objetivos

INCIDÊNCIA EM PROVA: MÉDIA

Segundo Navathe, a Mineração de Dados costuma ser executada com alguns objetivos finais ou aplicações. De um modo geral, esses objetivos se encontram nas seguintes classes: Previsão, Identificação, Classificação ou Otimização. *Como assim, Diego?* Cara, isso significa que você pode utilizar a mineração de dados com o objetivo que podem ser divididos nas seguintes classes: previsão, identificação, classificação e otimização. Fiquem com o mnemônico para memorizar:

PREVISÃO IDENTIFICAÇÃO CLASSIFICAÇÃO OTIMIZAÇÃO

OBJETIVO	DESCRIÇÃO
PREVISÃO	A mineração de dados pode mostrar como certos atributos dos dados se comportarão no futuro. Um de seus objetivos é prever comportamentos futuros baseado em comportamentos passados. Exemplo: análise de transações de compras passadas para prever o que os consumidores comprarão futuramente sob certos descontos, quanto volume de vendas uma loja gerará em determinado período e se a exclusão de uma linha de produtos gerará mais lucros. Em tais aplicações, a lógica de negócios é usada junto com a mineração de dados. Em um contexto científico, certos padrões de onda sísmica podem prever um terremoto com alta probabilidade.
IDENTIFICAÇÃO	Padrões de dados podem ser usados para identificar a existência de um item, um evento ou uma atividade. Por exemplo: intrusos tentando quebrar um sistema podem ser identificados pelos programas por eles executados, arquivos por eles acessados ou pelo tempo de CPU por sessão aberta. Em aplicações biológicas, a existência de um gene pode ser identificada por sequências específicas de nucleotídeos em uma cadeia de DNA. A área conhecida como autenticação é uma forma de identificação. Ela confirma se um usuário é realmente um usuário específico ou de uma classe autorizada, e envolve uma comparação de parâmetros, imagens ou sinais contra um banco de dados.
CLASSIFICAÇÃO	A mineração de dados pode particionar os dados de modo que diferentes classes ou categorias possam ser identificadas com base em combinações de parâmetros. Por exemplo: os clientes em um supermercado podem ser categorizados em compradores que buscam desconto, compradores com pressa, compradores regulares leais, compradores ligados a marcas conhecidas e compradores eventuais. Essa classificação pode ser usada em diferentes análises de

	transações de compra de cliente como uma atividade pós-mineração. Às vezes, a classificação baseada em conhecimento de domínio comum é utilizada como uma entrada para decompor o problema de mineração e torná-lo mais simples (Ex: alimentos saudáveis, alimentos de festa ou alimentos de lanche escolar são categorias distintas nos negócios do supermercado. Faz sentido analisar o relacionamento dentro e entre categorias como problemas separados). Essa categorização pode servir para codificar os dados corretamente antes de submetê-los a mais mineração de dados.
OTIMIZAÇÃO	Um objetivo 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 determinado conjunto de restrições. Como tal, esse objetivo da mineração de dados é semelhante à função objetiva, usada em problemas de pesquisa operacional, que lida com otimização sob restrições.

(TCE/SC – 2016) Para a realização de prognósticos por meio de técnicas de mineração de dados, parte-se de uma série de valores existentes obtidos de dados históricos bem como de suposições controladas a respeito das condições futuras, para prever outros valores e situações que ocorrerão e, assim, planejar e preparar as ações organizacionais.

_____

Comentários: prognóstico ou previsão partem dados históricos para prever situações futuras (Correto).

(ANAC – 2016) São objetivos da Mineração de Dados:

a) Distribuição, Identificação, Organização e Otimização.
b) Previsão, Priorização, Classificação e Alocação.
c) Previsão, Identificação, Classificação e Otimização.
d) Mapeamento, Identificação, Classificação e Atribuição.
e) Planejamento, Redirecionamento, Classificação e Otimização.

_____

Comentários: trata-se da Previsão, Identificação, Classificação e Otimização (Letra C).

(MDA – 2014) A mineração de dados costuma ser executada com alguns objetivos finais ou aplicações. Em geral, esses objetivos se encontram nas seguintes classes:

a) levantamento, previsão, classificação e otimização.
b) requisito, identificação, classificação e otimização.
c) previsão, identificação, levantamento e requisito
d) levantamento, requisito, classificação e otimização.
e) previsão, identificação, classificação e otimização

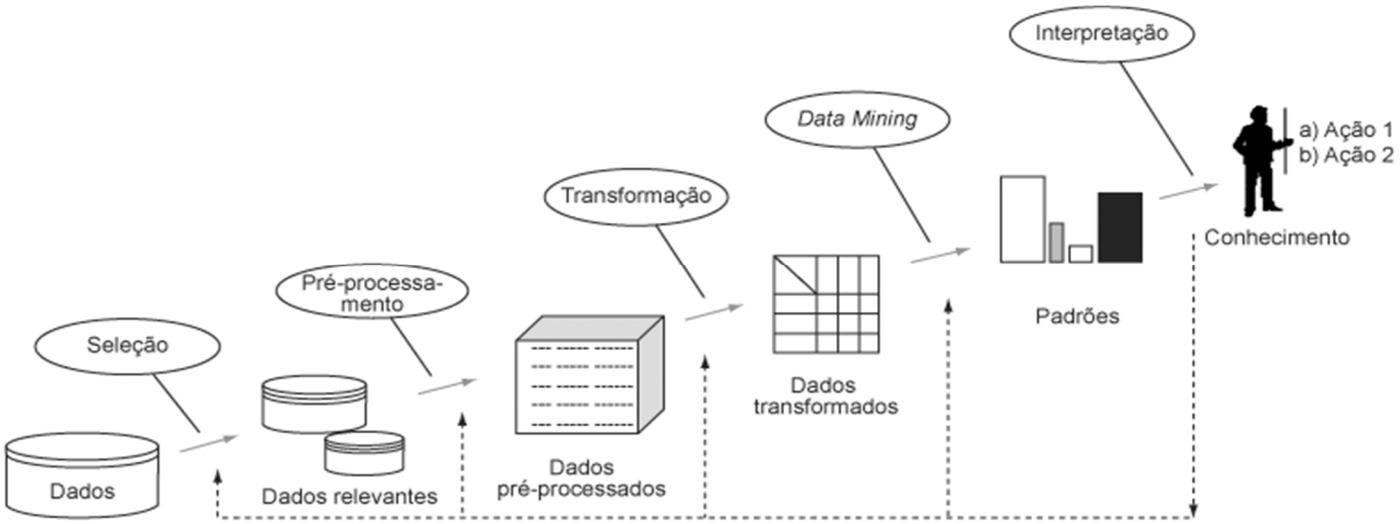
_____

Comentários: é a Previsão, Identificação, Classificação e Otimização (Letra E).

Processo de Descoberta de Conhecimento

INCIDÊNCIA EM PROVA: ALTA

A Mineração de Dados faz parte de um processo muito maior de descoberta de conhecimento chamada KDD (Knowledge Discovery in Databases – Descoberta de Conhecimento em Bancos de Dados). O processo de descoberta de conhecimento compreende cinco fases: (1) Seleção; (2) Pré-processamento; (3) Transformação; **(4) Data Mining**; (5) Interpretação e Avaliação – alguns autores possuem uma classificação um pouco diferente como veremos adiante.



O processo de descoberta de conhecimento é interativo e iterativo, envolvendo várias etapas com muitas decisões tomadas pelo usuário. É necessário desenvolver uma compreensão do domínio de aplicação e os conhecimentos anteriores relevantes. Dessa forma, a primeira etapa é selecionar um conjunto de dados de diversas bases – ou se concentrar em um subconjunto de variáveis ou amostras de dados – no qual a descoberta será realizada.

Com a seleção de dados relevantes, a segunda etapa será o pré-processamento dos dados. **Operações básicas incluem limpeza, remoção de erros, eliminação de redundância, decidir estratégias para lidar com campos de dados ausentes, entre outros**. Com os dados pré-processados, passamos à etapa de transformação, em que os dados são enriquecidos e consolidados em formas apropriadas à mineração, sumarizando-os ou agregando-os.

Com os dados transformados, passamos à etapa de mineração de dados. Utilizam-se algoritmos e técnicas para extrair possíveis padrões úteis de dados. Por fim, teremos a descoberta de diversos padrões que serão interpretados e avaliados em busca de padrões realmente interessantes e úteis, além de suas possíveis explicações ou interpretações. Vamos detalhar agora um pouco mais a fase de pré-processamento...

Quando temos bases de dados muito grandes e heterogêneas, é comum termos registros que comprometem a qualidade dos dados, como – por exemplo – registros inconsistentes, falta de informação, registros duplicados, valores discrepantes, entre outros. **No entanto, existem**

diversas técnicas para pré-processar dados com o intuito de preparar os dados brutos para serem analisados sem erros de incompletudes, inconsistências, ruídos, entre outros[4].

A técnica de pré-processamento possui alguns objetivos principais: **melhorar a qualidade dos dados; diminuir a ambiguidade das expressões linguísticas; diminuir a quantidade de dados a ser processado; estruturar as informações como tuplas; e melhorar a eficiência da mineração de dados.** Eslmari Navathe – renomado autor de bancos de dados – considera que o KDD contempla seis etapas:

1	SELEÇÃO DE DADOS	Dados sobre itens ou categorias são selecionados.
2	LIMPEZA DE DADOS	Dados são corrigidos ou eliminados dados incorretos.
3	ENRIQUECIMENTO DE DADOS	Dados são melhorados com fontes de informações adicionais.
4	TRANSFORMAÇÃO DE DADOS	Dados são reduzidos por meio de sumarizações, agregações e discretizações[5].
5	MINERAÇÃO DE DADOS	Padrões úteis são descobertos.
6	EXIBIÇÃO DE DADOS	Informações descobertas são exibidas ou relatórios são construídos.

Dito isso, esse autor considera que as quatro primeiras etapas podem ser agrupadas em uma única etapa de pré-processamento. *Fechou?* Fechado...

(ME – 2020) O objetivo das técnicas de pré-processamento de dados é preparar os dados brutos para serem analisados sem erros de incompletudes, inconsistências e ruídos.

_____

Comentários: o objetivo das técnicas de pré-processamento de dados realmente é preparar os dados brutos para serem analisados sem erros de incompletudes, inconsistências e ruídos (Correto).

(IPLANFOR – 2015) O processo de extração de informações de bases de dados é conhecido como descoberta de conhecimento em banco de fados, em inglês Knowledge Discovery in Databases (KDD). Segundo Fayyad, tal processo é constituído pelas etapas na seguinte ordem:

a) seleção – pré-processamento – transformação – mineração de dados – avaliação.
b) limpeza dos dados – seleção – transformação – mineração de dados – conhecimento.
c) mineração de dados –interpretação – avaliação.
d) projeção – seleção – mineração de dados – avaliação – conhecimento.

_____

Comentários: seleção – pré-processamento – transformação – mineração de dados – avaliação (Letra A).

[4] No aprendizado de máquina, ruídos são flutuações aleatórias ou erros em conjuntos de dados que podem afetar a precisão de um modelo. O ruído pode ser causado por uma variedade de fatores, como erros de sensores, outliers, entrada incorreta de dados, etc.

[5] Variáveis numéricas são convertidas em classes ou categorias.

Pré-Processamento de Dados

RELEVÂNCIA EM PROVA: BAIXA

PRÉ-PROCESSAMENTO DE DADOS

Trata-se do processo de preparação de dados para análise e modelagem adicionais. Envolve a transformação de dados brutos em um formato mais adequado para algoritmos de aprendizado de máquina por meio de tarefas como limpeza, normalização e organização dos dados para que possam ser analisados mais facilmente e usados para fazer previsões. É também uma etapa essencial em qualquer projeto de aprendizado de máquina, pois garante que os dados estejam em um formato adequado para o processo de modelagem.

Ao construir uma casa nova, antes de pensar em aspectos estéticos ou nos móveis planejados, é necessário construir uma base sólida sobre a qual serão construídas as paredes. Além disso, quanto mais difícil for o terreno em que você construirá a casa, mais tempo e esforço podem ser necessários. Se você deixar de criar uma base robusta, nada construído sobre ela resistirá ao tempo e à natureza por muito tempo.

O mesmo problema ocorre no aprendizado de máquina. Não importa o nível de sofisticação do algoritmo de aprendizado, se você não preparar bem sua base – ou seja, seus dados – seu algoritmo não durará muito quando testado em situações reais de dados. Mesmo os modelos mais sofisticados e avançados atingem um limite e têm desempenho inferior quando os dados não estão devidamente preparados.

Tem um ditado no meio de tecnologia da informação que diz: "*Garbage In, Garbage Out*". Em outras palavras, se entrar dados ruins para serem processados por algoritmos de aprendizado de máquina, dados ruins sairão das predições dos algoritmos. Não há mágica! *E o que seria um dado ruim, Diego?* São dados ausentes, anômalos, redundantes, não padronizados, entre outros. O tratamento dos dados brutos é útil para facilitar a compreensão, visualização e identificação de padrões.

O tempo gasto no tratamento de dados pode levar cerca de 80% do tempo total que você dedica a um projeto de aprendizado de máquina. Vamos ver os principais problemas e soluções:

Valores Ausentes

Um valor ausente costuma ser representado por um código de ausência, que pode ser um valor específico, um espaço em branco ou um símbolo (por exemplo, "?"). Um valor ausente caracteriza um valor ignorado ou que não foi observado, e, nesse sentido, a substituição de valores ausentes, também conhecida como imputação, tem como objetivo estimar os valores ausentes com base nas informações disponíveis no conjunto de dados.

A imputação de valores ausentes é uma técnica de pré-processamento que assume que a ausência de valor implica a perda de informação relevante de algum atributo. Consequentemente, o valor a

ser imputado não deve somar nem subtrair informação à base, ou seja, ele não deve enviesar a base. Associado a isso está o fato de que muitos algoritmos de mineração não conseguem trabalhar com os dados na ausência de valores e, portanto, a imputação é necessária para a análise.

Além disso, o tratamento incorreto ou a eliminação de objetos com valores ausentes pode promover erros das ferramentas de análise. Para resolver o problema de valores ausentes, temos algumas opções: ignorar os objetos que possuem um ou mais valores ausentes; imputar manualmente os valores ausentes; usar uma constante global para imputar o valor ausente; utilizar a média ou moda de um atributo para imputar o valor ausente; entre outros.

Dados Inconsistentes

No contexto de aprendizado de máquina, a consistência de um dado está relacionada à sua discrepância em relação a outros dados ou a um atributo, e tal consistência influencia na validade, na utilidade e na integridade da aplicação de mineração de dados. Um problema fundamental é que diferentes atributos podem ser representados pelo mesmo nome em diferentes sistemas, e o mesmo atributo pode ter diferentes nomes em diferentes sistemas.

Sempre gosto de lembrar do exemplo de bases de dados que utilizam nomes "Caixa Econômica Federal", "Caixa Econômica", "Caixa", "CEF" para representar o mesmo atributo. Outros problemas típicos de inconsistência ocorrem quando o valor apresentado na tabela não corresponde aos valores possíveis de um atributo ou o valor está fora do domínio correspondente; problemas de capitalização de textos, caracteres especiais, padronização de formatos (Ex: Data, CPF), etc.

Uma das formas de resolver esse problema é por meio de uma análise manual auxiliada por rotinas de verificação que percorrem a base de dados em busca de inconsistências.

Redução de Dimensionalidade

É intuitivo pensar que, quanto maior a quantidade de objetos e atributos, mais informações estão disponíveis para o algoritmo de mineração de dados. Entretanto, o aumento do número de objetos e da dimensão do espaço (número de atributos/variáveis na base) pode fazer com que os dados disponíveis se tornem esparsos e as medidas matemáticas usadas na análise tornem-se numericamente instáveis.

Além disso, uma quantidade muito grande de objetos e atributos pode tornar o processamento dos algoritmos de aprendizado de máquina muito complexo, assim como os modelos gerados. O ideal é utilizar técnicas de redução de dimensionalidade para reduzir a quantidade de atributos que descrevem os objetos. Não vamos nos aprofundar nesse tema agora porque teremos um tópico dedicado a esse tema mais à frente.

Normalização Numérica

A normalização é uma técnica geralmente aplicada como parte da preparação de dados para o aprendizado de máquina. O objetivo da normalização é mudar os valores das colunas numéricas no conjunto de dados para usar uma escala comum, sem distorcer as diferenças nos intervalos de valores nem perder informações. A normalização também é necessária para alguns algoritmos para modelar os dados corretamente.

Por exemplo: vamos supor que o conjunto de dados de entrada contenha uma coluna com valores variando de 0 a 1 e outra coluna com valores variando de 10.000 a 100.000. A grande diferença na escala dos números pode causar problemas ao tentar combinar os valores durante a modelagem. A normalização criando novos valores que mantêm a distribuição geral e as proporções nos dados de origem, mantendo os valores em uma escala aplicada em todas as colunas numéricas do modelo.

E como resolve isso? Você pode mudar todos os valores para uma escala de 0 a 1. Por exemplo: se você tiver um conjunto de dados de idades, poderá normalizá-lo para que todas as idades fiquem entre 0 e 1, subtraindo a idade mínima de cada valor e dividindo pela diferença entre as idades máxima e mínima. Vamos considerar o seguinte conjunto de dados de idade: [12, 17, 21, 29, 54, 89]. Agora utilizamos a seguinte fórmula:

$$x_i' = \frac{x_i - x_{min}}{x_{max} - x_{min}}$$

,em que x é a idade e i ∈ {1, ..., n}. Logo, vamos normalizar os valores do nosso conjunto original de dados para a escala entre 0 e 1:

$$\mathbf{IDADE}_1\text{'} = \frac{IDADE1 - IDADEmin}{IDADEmáx - IDADEmin} = \frac{12-12}{89-12} = \frac{12-12}{89-12} = \frac{0}{77} = 0{,}000$$

$$\mathbf{IDADE}_2\text{'} = \frac{IDADE2 - IDADEmin}{IDADEmáx - IDADEmin} = \frac{17-12}{89-12} = \frac{17-12}{89-12} = \frac{5}{77} = 0{,}064$$

$$\mathbf{IDADE}_3\text{'} = \frac{IDADE3 - IDADEmin}{IDADEmáx - IDADEmin} = \frac{21-12}{89-12} = \frac{21-12}{89-12} = \frac{9}{77} = 0{,}116$$

$$\mathbf{IDADE}_4\text{'} = \frac{IDADE4 - IDADEmin}{IDADEmáx - IDADEmin} = \frac{29-12}{89-12} = \frac{29-12}{89-12} = \frac{17}{77} = 0{,}220$$

$$\mathbf{IDADE}_5\text{'} = \frac{IDADE5 - IDADEmin}{IDADEmáx - IDADEmin} = \frac{54-12}{89-12} = \frac{54-12}{89-12} = \frac{42}{77} = 0{,}545$$

$$\mathbf{IDADE}_6\text{'} = \frac{IDADE6 - IDADEmin}{IDADEmáx - IDADEmin} = \frac{89-12}{89-12} = \frac{89-12}{89-12} = \frac{77}{77} = 1{,}000$$

Pronto! Agora o conjunto de dados original foi normalizado em um novo conjunto de dados contemplando valores que variam entre 0 e 1: [0, 0,064, 0,116, 0,220, 0,545, 1]. Esse tipo de normalização se chama Normalização Max-Min, mas existem outros tipos. Por fim não confundam

a normalização numérica de dados com a normalização de bases de dados relacionais, que é um conceito totalmente diferente visto dentro do contexto de bancos de dados.

Existe também a normalização por padronização – também conhecida como z-score ou padronização de variáveis –, que é uma técnica de pré-processamento de dados aplicada a variáveis contínuas e tem como objetivo colocar essas variáveis em uma escala com média zero (0) e desvio padrão um (1). Isso significa que, após a normalização, os dados terão uma média igual a zero e um desvio padrão igual a um.

Discretização

Alguns algoritmos de mineração operam apenas com atributos categóricos e, portanto, não podem ser aplicados a dados numéricos. Nesses casos, atributos numéricos podem ser discretizados, dividindo o domínio do atributo em intervalos e ampliando a quantidade de métodos de análise disponíveis para aplicação. Além disso, a discretização reduz a quantidade de valores de um dado atributo contínuo, facilitando, em muitos casos, o processo de mineração.

A maneira mais óbvia de discretizar um certo atributo é dividindo seu domínio em um número predeterminado de intervalos iguais, o que normalmente é feito no momento da coleta dos dados. *Vamos ver um exemplo?* Imagine que tenhamos um conjunto de dados numéricos que representam o peso de um grupo de pessoas. Nesse caso, podemos dividi-los em três faixas: 50 a 75 kg, 76 a 100 kg e 101 a 150 kgs.

Anomalias (Outliers)

Também chamado de valores ruidosos, referem-se a modificações dos valores originais e que, portanto, consistem em erros de medidas ou em valores consideravelmente diferentes da maioria dos outros valores do conjunto de dados. Casos típicos percebidos quando se conhece o domínio esperado para os valores em um atributo ou a distribuição esperada para os valores de um atributo, mas, no conjunto de dados, aparecem alguns valores fora desse domínio ou dessa distribuição.

Alguns exemplos são: valores que naturalmente deveriam ser positivos, mas, no conjunto de dados, aparecem como negativos, como seria o caso de observar um valor negativo para a quantidade de vendas de um restaurante em um período de um mês; ou ainda, observar que o valor de vendas desse mesmo restaurante ultrapassa, com uma grande margem, o valor total de vendas de todos os anos anteriores.

No primeiro caso, diz-se que o valor está errado por não fazer sentido no contexto dos dados; no segundo caso, tem-se um exemplo de *outlier*, que pode representar um valor errado ou uma mudança não esperada, porém real, do comportamento dos valores de um atributo, e, nesse caso, há a necessidade de identificar qual é a explicação que se adéqua à situação. Para resolver esse problema, pode-se fazer inspeções/correções manuais ou identificação/limpeza automáticas.

Dados Categóricos

Dados categóricos são comuns em problemas aprendizado de máquina e, na maioria das vezes, podem ser mais desafiadores de extrair informações do que dados numéricos. Esses dados deverão ser transformados em dados numéricos para que possam ser incluídos na etapa de treinamento. Assim, uma melhor representação dos dados categóricos afeta diretamente a performance do treinamento como também introduz melhores formas de explicar sua contribuição para a predição.

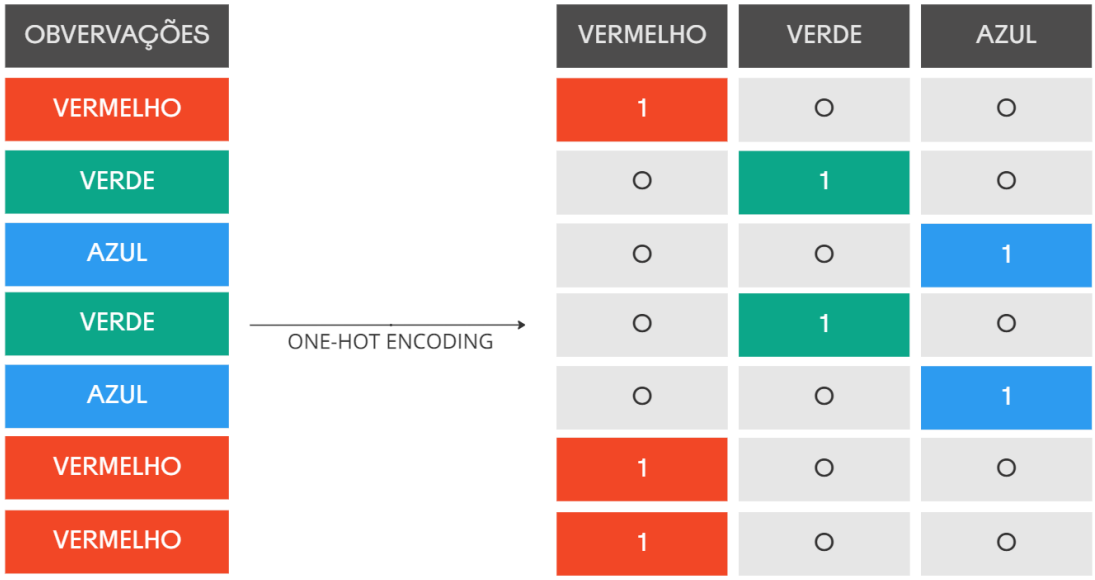
A maioria dos algoritmos de aprendizado de máquina trabalha com variáveis numéricas, mas e quando temos dados categóricos? Aí temos que fazer uma conversão! Lembrando que dados categóricos podem ser ordenados (Ex: baixo, médio, alto) ou não ordenados (Ex: vermelho, azul, verde). No primeiro caso, pode-se representar por meio de uma codificação chamada *Ordinal Encoding*, em que cada valor varia de 1 até n classes (Ex: baixo = 1, médio = 2 e alto = 3).

Professor, essa atribuição de valores não pode dar problema? Pode, sim! Nesse exemplo, meu modelo pode entender que "alto" tem três vezes o peso de "baixo" – o que pode não ser verdadeiro em meu modelo de negócio. O nome desse problema é ponderação arbitrária, dado que estamos dando pesos arbitrários às classes. No caso de dados categóricos não ordenados, é mais complicado ainda e nem a solução anterior funciona.

Para resolver esse tipo de problema, o ideal é utilizar uma codificação chamada *One-Hot Enconding* ou *Dummy Encoding*. Vamos ver um exemplo rapidamente: imagine uma classificação de dados nas categorias Vermelho, Verde e Azul. Note que não existe uma ordem natural dessas categorias e atribuir valores aleatórios poderia recair no problema da ponderação arbitrária. Podemos utilizar, portanto, variáveis *dummy*. *O que seria isso, Diego?*

Variáveis *dummy* são variáveis que assumem os valores binários (0 ou 1) para representar a ausência ou presença de um determinado atributo, característica, categoria, evento ou critério.

Agora vamos criar uma tabelinha com uma coluna para cada categoria (vermelho, verde e azul). Em seguida, para cada observação do nosso conjunto de dados, vamos atribuir o valor 1 (um) – quando correspondente à categoria – e o (zero) – quando não correspondente à categoria. Logo, a primeira observação é "vermelho", logo inserimos 1 (um) na coluna "vermelho" e o (zero) nas outras colunas; e fazemos assim para cada uma das observações.



Note que essa operação é bidirecional: é possível sair do valor original para o *one-hot enconding* quanto do *one-hot encoding* para o valor original. Dessa forma, não temos uma perda nem um acréscimo de informação! Existem variações dessa técnica para a codificação de dados categóricos. A vantagem dessas técnicas é permitir que o algoritmo de aprendizado de máquina entenda melhor as relações entre os dados e faça previsões mais precisas sem perda/acréscimo de dados.

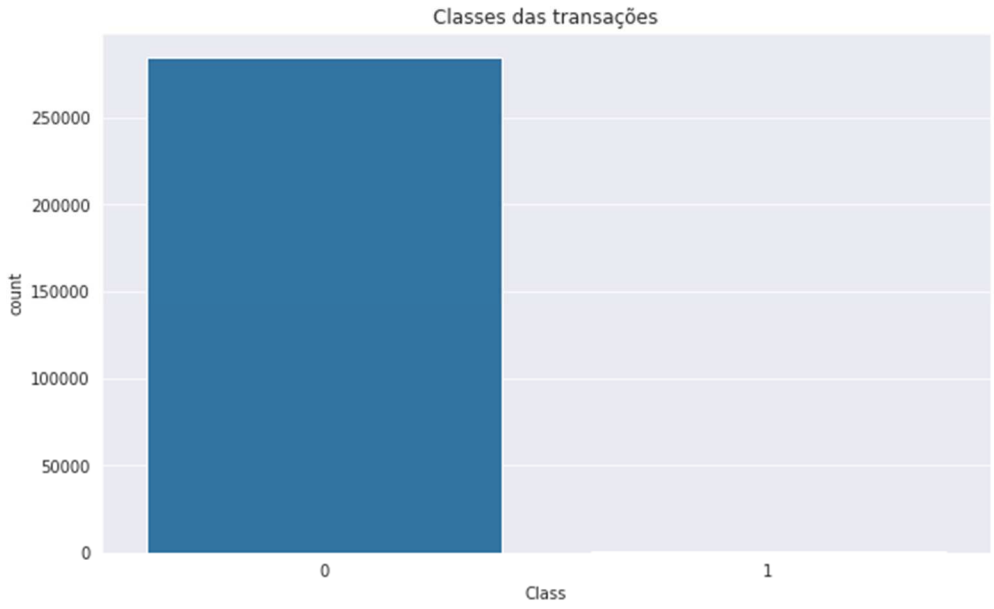
Para finalizar, note que o *one-hot encoding* tem uma redundância: se eu tenho *n* categorias, eu não preciso de *n* colunas na tabela para representá-las. No caso das categorias de cores, se uma observação não é vermelha nem verde, ela é necessariamente azul; se não é vermelha nem azul, ela é necessariamente verde; e se não é verde nem azul, ela é necessariamente vermelha. Logo, para representar essas categorias, bastava n-1 colunas na tabela. Vejamos:

Note que, com apenas duas colunas, foi possível representar a mesma informação da tabela anterior. É bem simples: tudo que não for vermelho ou verde, será azul!

Classes Desbalanceadas

Um problema muito comum de classificação ocorre quando existe uma desproporção nos dados, isto é, temos muito mais dados de uma classe do que de outra (pode ocorrer também para problemas com mais de duas classes). Então, quando temos um desequilíbrio de classe, o classificador de aprendizado de máquina tende a ser mais tendencioso em relação à classe majoritária, causando má classificação da classe minoritária. *Como assim, Diego?*

Imagine os dados sobre análise de fraudes em transações de cartões de crédito. *Sabe quando aparece no seu celular uma notificação sobre uma compra que você sabe que você não fez?* É aquela correria para ligar no banco e bloquear o cartão. Pois é, tem sido cada vez mais comum! Quando analisamos um conjunto de dados sobre transações de cartões de crédito, vemos basicamente duas classes: transações normais e transações fraudulentas. *E onde entra o desbalanceamento?*



Ora, eu já tive cartões bloqueados em duas oportunidades por conta de transações fraudulentas, mas é a exceção da exceção da exceção: a regra é que a imensa maioria das transações sejam normais. Vejam na imagem anterior um exemplo de um conjunto de dados de transações de uma operadora de cartão de crédito: a Classe 0 representa as transações normais e a Classe 1 representa as transações fraudulentas.

Note que nem é possível ver nada azul da Classe 1 porque é tão pouco que é quase irrelevante, mas esse gráfico representa 284.315 transações normais e 492 transações fraudulentas. Em outras palavras, temos 99,82% de transações normais e 0,17% de transações fraudulentas. *E por que isso é um problema, Diego?* Porque esse desequilíbrio pode levar a modelos tendenciosos que não são representativos da população como um todo.

Isso pode ocasionar previsões imprecisas e baixo desempenho do modelo. Além disso, pode levar a classificadores excessivamente sensíveis à classe majoritária, ignorando a classe minoritária.

Galera, imagine que eu faça a pior modelagem possível de um modelo de classificação! Por pior que ele seja, se ele simplesmente "chutar" sempre que uma transação foi normal, ele acertará na maioria das vezes porque raramente uma transação é fraudulenta.

Nesse caso, a acurácia do nosso modelo será próxima de 100%, mesmo ele tendo sido pessimamente modelado. A máquina não aprendeu nada – ela apenas foi guiada a minimizar o erro no conjunto de dados de treinamento (que tem ampla maioria de transações normais) e ignorou os padrões de transações fraudulentas. O nome disso é Paradoxo da Acurácia, em que parâmetros de um algoritmo não diferenciam a classe minoritária das demais categorias.

Ora, mas o principal objetivo de um modelo de aprendizado de máquina que trata de transações de cartão de crédito é justamente identificar padrões de transações fraudulentas a fim de impedi-las. Bacana! *E como resolve isso?* Bem, uma alternativa é conseguir mais dados de treinamento – preferencialmente da classe minoritária. Outra alternativa é alterar a métrica de desempenho: em vez de usar a acurácia.

A acurácia é uma métrica muito sensível ao acerto médio, logo não tem um bom desempenho na medição da qualidade de modelos quando o conjunto de dados tem uma desproporção muito grande entre as classes. Para resolver esse problema, podemos utilizar outras métricas de desempenho, tais como F-Score ou Curva ROC. Outra alternativa é fazer uma reamostragem e, para isso, temos duas opções...

Podemos fazer um *undersampling* majoritário (ou subamostragem), que consiste em eliminar aleatoriamente dados da classe majoritária até que ambas as classes tenham a mesma quantidade de dados. Inversamente, podemos fazer um *oversampling* minoritário (ou superamostragem), que consiste em criar novos dados baseados nos dados da classe minoritária de forma aleatória até que ambas as classes tenham a mesma quantidade de dados.

Trata-se de uma forma de garantir que os dados da classe minoritária apareçam várias vezes – é como se eu replicasse aleatoriamente dados da classe minoritária para eliminar a desproporção. É claro que essas soluções também possuem problemas: a subamostragem perde informações úteis, o que pode fazer o modelo gerar *underfitting*; já a superamostragem pode induzir o modelo a encontrar padrões de dados que não refletem a realidade da classe minoritária.

Existem outras técnicas, tais como: utilização de amostras sintéticas, atribuição de pesos diferentes às classes, utilização de algoritmos específicos para detecção de anomalias, avaliação de algoritmos menos sensíveis ao desbalanceamento (como as árvores de decisão). Enfim... existem diversas soluções e essas soluções podem ser combinadas de diversas formas para minimizar o problema do desbalanceamento (também chamado de amostras enviesadas).

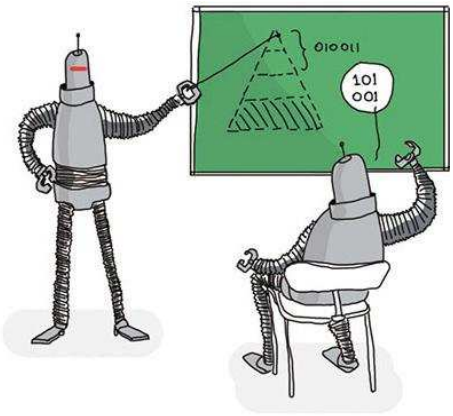
Esse foi basicamente o tema de uma das questões da prova discursiva do TCU! Ele mostrava um caso em que um classificador de peças boas ou defeituosas atingiu altíssima precisão. Ocorre que o modelo basicamente que todas as peças eram boas e nenhuma peça era defeituosa – caso típico de

desbalanceamento de classes. Como o modelo tinha o objetivo principal justamente de indicar peças defeituosas, ele fez um péssimo serviço!

Pedia-se também para dar exemplos de soluções, tais como *oversampling*, *undersampling*, atribuição de pesos diferentes às classes, entre outras – tudo que acabamos de estudar!

Aprendizado de Máquina

INCIDÊNCIA EM PROVA: BAIXÍSSIMA



Aprendizado de Máquina (do inglês, *Machine Learning*) é a área da inteligência artificial que busca desenvolver técnicas computacionais sobre aprendizado assim como a construção de sistemas capazes de adquirir conhecimento de forma autônoma que tome decisões baseado em experiências acumuladas por meio da solução bem-sucedida de problemas anteriores. **Trata-se de uma ferramenta poderosa para a aquisição automática de conhecimento por meio da imitação do comportamento de aprendizagem humano com foco em aprender a reconhecer padrões complexos e tomar decisões.**

Vamos ver exemplo bem simples: imagine que você deseja criar um software para diferenciar maçãs e laranjas. Você tem dados que indicam que as laranjas pesam entre 150-200g e as maçãs entre 100-130g. Além disso, as laranjas têm cascas ásperas e as maçãs têm cascas suaves (que você pode representar como 0 ou 1). **Ora, se você tem uma fruta que pesa 115g e possui casca suave, seu programa pode determinar que provavelmente se trata de uma maçã.**

De outra forma, se a fruta tem 175g e possui casca áspera, provavelmente se trata de uma laranja. Qualquer coisa fora desses limites também não será nem um nem outro. *E se sua fruta tem casca suave, mas pesa apenas 99g?* Você sabe que provavelmente é uma maçã, mas o seu programa não sabe disso. Logo, quanto mais dados você tiver, mais precisos serão os seus resultados. **A parte legal é que seu algoritmo pode ir aprendendo sozinho o que é uma laranja ou uma maçã.**

Cada vez que o usuário confirma que o programa acertou ou errou, ele é capaz de aprender e melhorar! Vejam no exemplo abaixo uma tecnologia que está ficando cada vez mais comum: reconhecimento fácil! Hoje em dia, a China a possui mais de 200 milhões de câmeras de vigilância com essa tecnologia e, quanto mais ela for testada, mais a máquina aprenderá e ficará mais precisa. *Sério, tem coisa mais linda que a área de Tecnologia da Informação!* :)

(SERPRO – 2013) Algumas das principais técnicas utilizadas para a realização de Datamining são: estatística, aprendizado de máquina, datawarehouse e recuperação de informações.

Comentários: todas essas são técnicas de Data Mining – inclusive Aprendizado de Máquina. *Data Warehouse também?* Galera, isso é bastante impreciso! O ideal seria mencionar *Data Warehousing,* mas a banca ignorou solenemente (Correto).

Mineração de Texto (Text Mining)

INCIDÊNCIA EM PROVA: BAIXÍSSIMA

A Mineração de Texto é um meio para encontrar padrões interessantes/úteis em um contexto de informações textuais não estruturadas, combinado com alguma tecnologia de extração e de recuperação da informação, processo de linguagem natural e de sumarização ou indexação de documentos. A internet está cheia de informações e processá-las pode ser uma tarefa e tanto, mas essa tarefa pode ser facilmente executada por meio de ferramentas de *Text Mining*.

Para ser bem honesto, qualquer informação é inútil se não pudermos interpretá-la da maneira adequada. As tecnologias de aprendizado de máquina são usadas para extrair informações ocultas ou abstratas de um pedaço de texto utilizando um computador. **O resultado é inovador porque descobrimos tendências e opiniões inexploradas que têm implicações inacreditáveis em diversos campos do conhecimento.**

Há alguns anos, nós estamos testemunhando uma enorme reviravolta em todos os setores devido à aceitação das mídias sociais como uma plataforma para compartilhar opiniões e comentários sobre qualquer coisa, incluindo grandes marcas, pessoas, produtos, etc. **Os usuários agora utilizam plataformas como o Facebook e o Twitter para compartilhar seus sentimentos. Hoje mesmo eu estava navegando pela minha linha do tempo e vi o seguinte tweet:**

Observem que a empresa não é boba e já foi atuar pela própria rede social!

Pois é, as mídias sociais agora estão preenchidas com milhões de avaliações sobre diversos produtos, imagine então sobre uma marca. Dessa forma, torna-se imperativo que qualquer marca utilize alguma ferramenta de mineração de textos para conhecer os sentimentos de seus clientes a fim de atendê-los bem. Com essas ferramentas, um profissional de marketing pode conhecer todas as tendências e opiniões relacionadas aos seus concorrentes em potencial.

O que as pessoas sentem e dizem está espalhado por toda a Internet, mas quem se daria ao luxo de avaliar e estudar tudo isso? **Vamos ser sinceros: as ferramentas de mineração de texto neste contexto podem ser um potencial enorme de fonte de renda para esse sujeito de marketing!** Enfim, esse é só um panorama geral para que vocês entendam a abrangência da aplicação dessa tecnologia no mundo hoje. Voltemos à teoria...

Em suma, a mineração de texto tem como objetivo a busca de informações relevantes e a descoberta de conhecimentos significativos a partir de documentos textuais não estruturados ou semiestruturados. **Este processo envolve um grau de dificuldade significativo considerando que as informações normalmente estão disponíveis em linguagem natural, sem a preocupação com a padronização ou com a estruturação dos dados – sua matéria prima é a palavra!**

Por fim, um bom exemplo de Mineração de Texto é o Processamento de Linguagem Natural (PLN). Trata-se de uma área dentro da Inteligência Artificial que busca fazer com que os computadores entendam e simulem uma linguagem humana. **É utilizado em diversas ferramentas como Google Tradutor, Sistemas de Reconhecimento de Falas e Nuvem de Palavras – esse último é um dos que eu acho mais interessantes.**

Uma vem de palavras é um recurso gráfico (usado principalmente na internet) para descrever os termos mais frequentes de um determinado texto. O tamanho da fonte em que a palavra é apresentada é uma função da frequência da palavra no texto: palavras mais frequentes são desenhadas em fontes de tamanho maior, palavras menos frequentes são desenhadas em fontes de tamanho menor. Eu gosto muito de brincar com essas ferramentas, vejam só...

HTTPS://MONKEYLEARN.COM/WORD-CLOUD/

Você pode inserir qualquer texto e ele devolverá uma nuvem de palavras com tamanho proporcional a frequência. Por curiosidade, eu inseri a letra de Faroeste Caboclo do Legião Urbana e abaixo está o resultado. A palavra que mais aparece é "Santo Cristo", depois "Maria Lúcia", e assim por diante. Claro que se trata de uma ferramenta de inteligência artificial capaz de eliminar alguns pronomes, artigos, entre outros para que seja mais útil ao usuário.

Achei tão divertido brincar com isso que fiz o download de uma bíblia digital e inseri o capítulo do Gênesis na ferramenta. Vejam as palavras que mais aparecem:

(IF/GO – 2019) É um meio de encontrar padrões interessantes ou úteis em um contexto de informações textuais não estruturadas, combinado com alguma tecnologia de extração e de recuperação da informação, processo de linguagem natural e de sumarização/indexação de documentos. (Dixson, 1997, apud TRYBULA, 1999).

O conceito apresentado pelo autor se refere ao processo de:

a) mineração de dados.
b) ontologia.
c) redes semânticas.
d) mineração de texto.

_____

Comentários: o meio de encontrar padrões úteis em textos não estruturados se refere à mineração de texto (Letra D).

(ME – 2020) Mecanismos de busca utilizam mineração de textos para apresentar ao usuário os resultados de suas pesquisas, de modo que ambos os conceitos se equivalem.

_____

Comentários: opa... cuidado! A Mineração de Texto busca descobrir informações previamente desconhecidas; já Mecanismos de Busca partem já do que o usuário deseja procurar e apenas recupera (Errado).

(ME – 2020) A mineração de textos utiliza técnicas diferentes da mineração de dados, tendo em vista que os textos representam um tipo específico de dado.

_____

Comentários: na verdade, as técnicas utilizadas são as mesmas – muda apenas o tipo de dado (Errado).

Técnicas e Tarefas

INCIDÊNCIA EM PROVA: ALTÍSSIMA

Sendo rigoroso, tarefas consistem na especificação do que estamos querendo buscar nos dados (O que?) e as técnicas consistem na especificação de como descobrir os padrões (Como?). Em geral, as bancas não cobram essa diferenciação e tratam os conceitos como similares. Conforme é apresentado na imagem a seguir, podemos dividi-las em duas categorias: Preditivas e Descritivas. Vejamos na imagem seguinte...

Técnicas Preditivas buscam prever os valores dos dados usando resultados conhecidos coletados de diferentes conjuntos de dados, isto é, busca-se prever o futuro com base dos dados passados. Já **Técnicas Descritivas** buscam descrever relacionamentos entre variáveis e resumir grandes quantidades de dados. Eles usam técnicas estatísticas para encontrar relações entre variáveis, como correlações e associações.

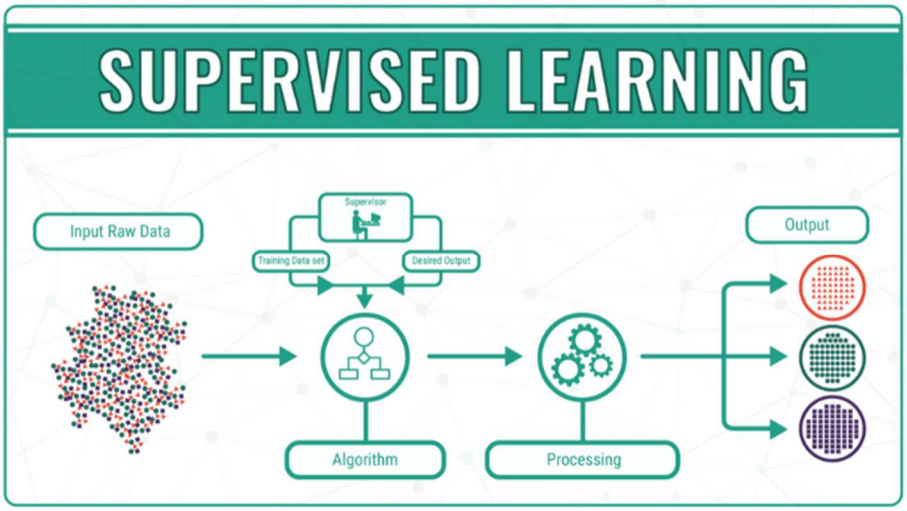
Antes de seguir para o detalhamento de cada uma das técnicas, vamos falar sobre o conceito de aprendizado supervisionado, aprendizado não supervisionado e aprendizado por reforço[6]. É bem simples: uma técnica de aprendizado supervisionado é aquela que necessita de supervisão ou interação com um ser humano, enquanto uma técnica de aprendizado não supervisionado não necessita desse tipo de supervisão ou interação. *Calma que ficará mais claro...*

No aprendizado supervisionado, um ser humano alimenta o algoritmo com categorias de dados de saída de tal forma que o algoritmo aprenda como classificar os dados de entrada nas categorias de dados de saída pré-definidas. Vejam na imagem seguinte que há um conjunto de pontinhos e eu desejo classificá-los de acordo com as suas cores em vermelho, verde e roxo (é só um exemplo, as categorias poderiam ser pontinhos cujo nome da cor começa com a letra V ou R).

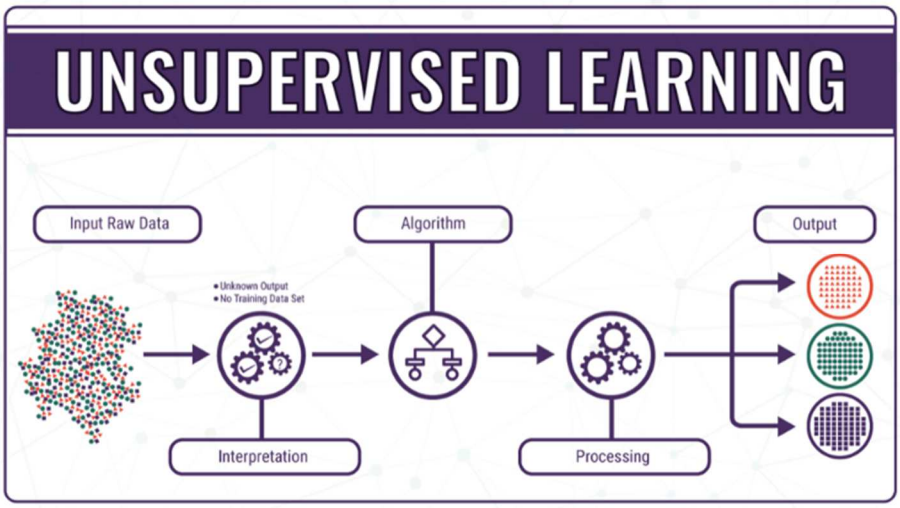
Note que – de antemão – eu já defini quais serão as categorias de saída que desejo, logo o algoritmo irá receber os dados brutos de entrada, irá processá-los e aprenderá a classificá-los em cada uma

[6] Em tese, existe também o aprendizado semi-supervisionado em que parte dos dados são rotulados e parte não são rotulados (isso ajudar a reduzir custos do processo supervisionado).

das categorias de saída que eu defini inicialmente. **Se um ser humano interferiu no algoritmo pré-definindo as categorias de saída do algoritmo, o algoritmo utilizou um aprendizado supervisionado porque ele aprendeu a categoria/rótulo, mas com o auxílio de um ser humano.**



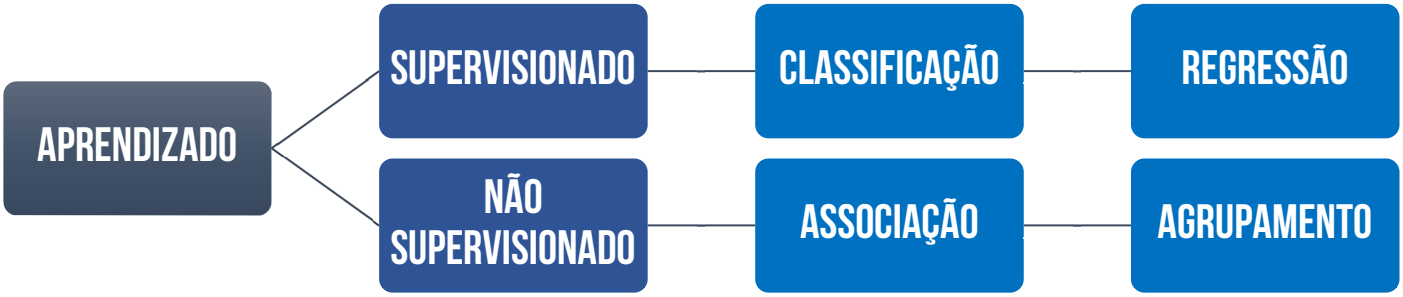
Já no aprendizado não supervisionado, um ser humano não alimenta o algoritmo com categorias de dados de saída pré-definidas. Vejam na imagem anterior que há um conjunto de pontinhos e o algoritmo – por si só – interpreta esses dados de entrada em busca de similaridades, padrões e características em comum, realiza o processamento e ele mesmo os categoriza da maneira que ele acha mais interessante sem nenhuma interferência humana durante o processo.



Existe também uma terceira categoria chamada **Aprendizado por Reforço**, que usa recompensas e punições para ajudar os algoritmos a aprenderem. O princípio é simples: quando um agente de aprendizado realiza uma ação que leva a uma recompensa, ele irá tentar repetir essa ação para conseguir a recompensa novamente. Por outro lado, quando uma ação leva a uma punição, o agente evitará essa ação no futuro. O objetivo aqui é maximizar a recompensa ao longo do tempo!

Imagine um programa de computador que joga xadrez com um humano. O programa receberia um conjunto de regras e recompensas e, então, tentaria encontrar o melhor movimento a ser feito para maximizar suas recompensas (vencer o adversário). Ao jogar e tomar decisões na tentativa e erro, aprenderia com seus sucessos e insucessos e, eventualmente, se tornaria um jogador de xadrez experiente (tanto que hoje em dia é quase impossível ganhar de um computador).

As técnicas de Classificação e Regressão são consideradas de aprendizado supervisionado; já as técnicas de Associação, Agrupamento são de aprendizado não-supervisionado[7].



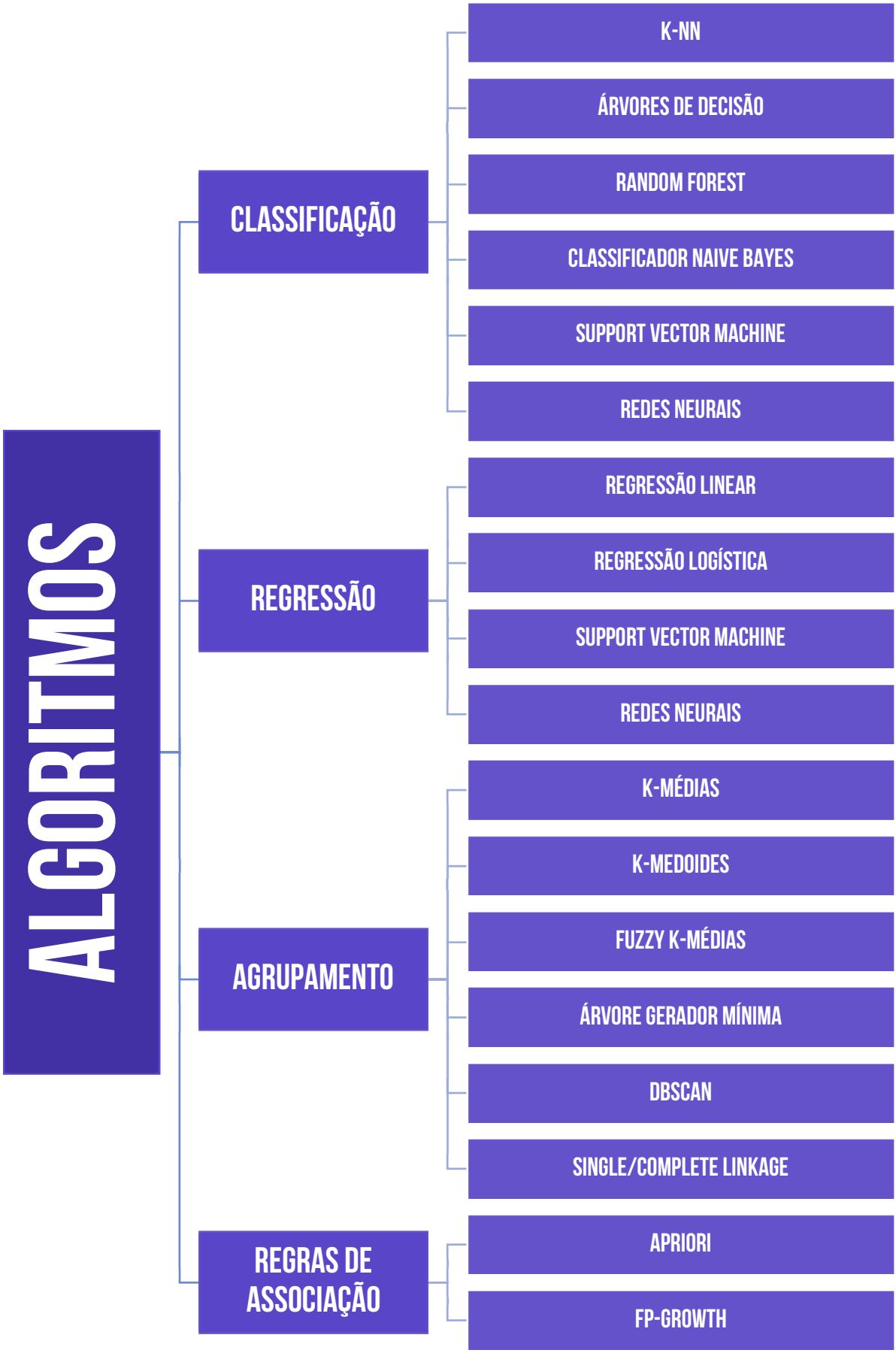
ATENÇÃO MÁXIMA

Galera, agora um ponto importante: nas próximas 50 páginas (mais ou menos), nós veremos essas quatro técnicas/tarefas de mineração de dados e aprendizado de máquina (que possuem um **excelente** custo/benefício) e dentro de cada uma das quatro nós estudaremos diversos algoritmos que implementam de maneiras diferentes essas tarefas (que têm um **péssimo** custo/benefício). Então, eu recomendo que vocês analisem bem o tipo de estudo que desejam...

Existem dezenas de questões sobre as técnicas/tarefas e pouquíssimas questões sobre os algoritmos que as implementam. *Ué, Diego! Então por que você não deixa logo essa parte de fora da aula?* Porque a aula tem que atender tanto aos alunos que desejam fazer um estudo mais aprofundado quanto aos alunos que desejam fazer um estudo mais superficial. Além disso, como esse é um conteúdo relativamente novo, é possível que as provas aprofundem mais com o tempo.

Na próxima página, há um diagrama com os principais algoritmos de implementação das tarefas de mineração de dados e aprendizado de máquina (nem todas serão explicadas nessa aula).

[7] Dica que eu encontrei recentemente: começou com A é uma técnica de aprendizado não supervisionado (**A**ssociação, **A**grupamento).


ALGORITMOS
CLASSIFICAÇÃO
K-NN
ÁRVORES DE DECISÃO
RANDOM FOREST
CLASSIFICADOR NAIVE BAYES
SUPPORT VECTOR MACHINE
REDES NEURAIS
REGRESSÃO
REGRESSÃO LINEAR
REGRESSÃO LOGÍSTICA
SUPPORT VECTOR MACHINE
REDES NEURAIS
AGRUPAMENTO
K-MÉDIAS
K-MEDOIDES
FUZZY K-MÉDIAS
ÁRVORE GERADOR MÍNIMA
DBSCAN
SINGLE/COMPLETE LINKAGE
REGRAS DE ASSOCIAÇÃO
APRIORI
FP-GROWTH

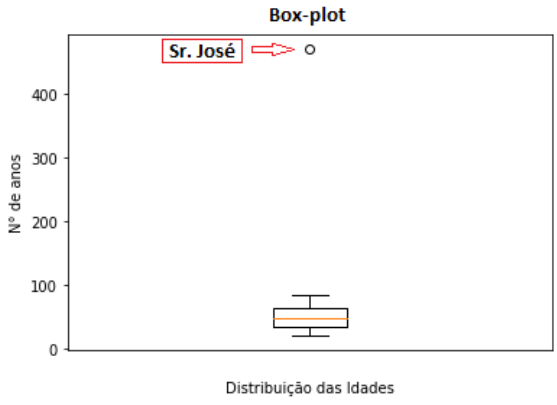
Detecção de Anomalia

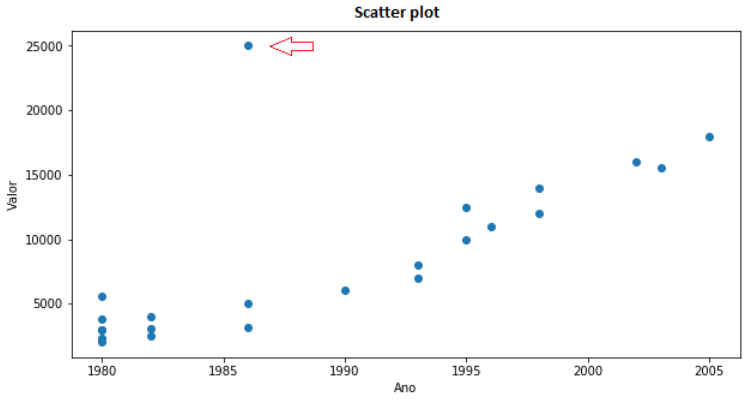
INCIDÊNCIA EM PROVA: BAIXÍSSIMA

Antes de falarmos especificamente das quatro principais técnicas de mineração, vamos falar sobre a Detecção de Anomalia! Ela também é uma tarefa de mineração, porém pode ser utilizada em conjunto com as outras. **O padrão é que ela funcione com algoritmos de aprendizado não supervisionados, mas ela também pode funcionar com algoritmos de aprendizado supervisionados.**

O que é uma anomalia? Em nosso contexto, é aquele famoso ponto "fora da curva" – também conhecido em inglês como *outlier*. *Como assim, professor?* Galera, a média de altura para um homem no mundo é de 1,74m! Eu, por exemplo, tenho 1,77m e estou bem próximo da média, no entanto algumas vezes nós encontramos exceções. Na imagem ao lado, nós temos o nepalês Chandra Dangi e o turco Sultan Kosen – o primeiro é o menor homem do mundo atualmente, medindo 0,54m e o segundo é o maior homem do mundo atualmente, medindo 2,51m. *É comum ver pessoas desses tamanhos?* Não! Por essa razão, esses dados de altura são considerados anomalias, exceções, aberrações justamente porque eles fogem drasticamente da normalidade e do padrão esperado. *E qual é o problema disso?* **O problema é que isso pode causar anomalias nos resultados obtidos por meio de algoritmos e sistemas de análise.** *E o que causa uma anomalia?*

Bem, a causa pode ter uma origem natural ou artificial. **Uma pessoa que há trinta anos declara ter crescimento de patrimônio anual de 1% e, de repente, declara ter tido um crescimento de 1000% pode ter ganhado na megasena – essa é uma anomalia natural.** Por outro lado, essa mesma pessoa pode simplesmente ter errado na hora de digitar e declarou patrimônio de R$10.000.000 em vez de R$100.000 – essa é uma anomalia artificial.





Anomalias artificiais podem partir de erros de amostragem, erros de processamento de dados, erros na entrada de dados, erros de medida ou erros intencionais. **Bem, uma forma de detectar anomalias é por meio de técnicas de análise de *outlier*, usando – por exemplo – gráficos como *Box-plot* ou *Scatter-plot*.** Fora as ferramentas gráficas, podemos utilizar cálculos que podem ser acrescentados às nossas rotinas, tornando o tratamento dos outliers mais eficiente.

A Detecção de Anomalias[8] é basicamente um valor discrepante, ou seja, um valor que se localiza significativamente distante dos valores considerados normais. É importante notar que uma anomalia não necessariamente é um erro ou um ruído, ela pode caracterizar um valor ou uma classe bem definida, porém de baixa ocorrência, às vezes indesejada, ou que reside fora de agrupamentos ou classes típicas.

Quase todas as bases de dados reais apresentam algum tipo de anomalia, que pode ser causada por fatores como **atividades maliciosas** (por exemplo, furtos, fraudes, intrusões, etc), **erros humanos** (por exemplo, erros de digitação ou de leitura), **mudanças ambientais** (por exemplo, no clima, no comportamento de usuários, nas regras ou nas leis do sistema etc), **falhas em componentes** (por exemplo, peças, motores, sensores, atuadores etc), etc.
As principais aplicações de detecção de anomalias incluem: detecção de fraudes, análise de crédito, detecção de intrusão, monitoramento de atividades, desempenho de rede, diagnóstico de faltas, análise de imagens e vídeos, monitoramento de séries temporais, análise de textos, etc. **A detecção de anomalias em bases de dados é essencialmente um problema de classificação binária, no qual se deseja determinar se um ou mais objetos pertencem à classe normal ou à anômala.**

(MF – 2011) Funcionalidade cujo objetivo é encontrar conjuntos de dados que não obedecem ao comportamento ou modelo dos dados. Uma vez encontrados, podem ser tratados ou descartados para utilização em mining. Trata-se de:

a) descrição.
b) agrupamento.
c) visualização.
d) análise de outliers.
e) análise de associações.

_____
Comentários: a análise de outliers busca encontrar conjuntos de dados que não obedecem ao comportamento ou modelo dos dados (Letra D).

[8] Também chamada de Detecção de Novidades, Detecção de Ruídos, Detecção de Desvios, Detecção de Falhas, Detecção de Exceções ou Mineração de Exceções.

Classificação

INCIDÊNCIA EM PROVA: MÉDIA

Trata-se de uma técnica de mineração de dados que designa itens de dados a uma determinada classe ou categoria previamente definida a fim de prever a classe alvo para cada item de dado.

Galera, a ideia da técnica de classificação é categorizar coisas. *Posso dar um exemplo?* Existe uma loja americana chamada Target – ela vende de tudo, mas principalmente roupas. Ela ficou famosa em 2002 por conseguir adivinhar mulheres que estavam grávidas e lhes enviar cupons de desconto relacionados a bebês. *Diego, como eles fizeram isso?* Cara, isso virou exemplo de problemas de classificação de livros didáticos de mineração de dados.

A Target precisava classificar cada cliente em uma de duas categorias: provavelmente grávida ou provavelmente não grávida. A classificação é um processo que geralmente funciona em vários estágios. Primeiro, cada instância tem que ser dividida em uma coleção de atributos – também chamados de rótulo ou etiqueta. Para uma loja como a Target, uma instância poderia ser uma mulher qualquer. *Calma que vocês vão entender...*

Eu disse que cada instância deve ser dividida em uma coleção de atributos. *Que atributos seriam relevantes para descobrir se uma mulher está grávida?* Bem, a Target possuía um banco de dados com informações sobre todas as suas clientes como nome, data de nascimento, endereço, e-mail e o principal: histórico de compras. Ademais, é comum a compra ou compartilhamento de bases de dados entre empresas.

Logo, ela possuía todo o histórico de compras de diversas clientes em várias empresas. Além disso, a Target possuía um cadastro em seu site para oferecer descontos para mulheres que se registrassem como grávidas. **Com base em tudo isso, ela analisou uma pequena amostra de dados e, em pouco tempo, surgiram alguns padrões úteis.** Um exemplo interessantíssimo foi a compra de creme hidratante! *Como assim, professor?*

Analisando a base de mulheres grávidas, um analista descobriu que elas estavam comprando quantidades maiores de creme hidratante sem cheiro por volta do começo do segundo trimestre de gravidez. Outro analista observou que, em algum momento das vinte primeiras semanas, mulheres grávidas compraram muitos suplementos como cálcio, magnésio e zinco. **Cada informação dessas pode ser considerada um atributo que ajuda a encaixar nas duas categorias.**

Analistas também descobriram que mulheres na base de registro de grávidas bem próximas de ganhar o bebê estavam comprando de forma diferenciada uma quantidade maior de sabão neutro e desinfetantes para mãos. **Após interpretação, foi descoberto que isso significava que**

provavelmente o bebê estava próximo de nascer! *Por que isso é importante?* Porque cada informação dessa é um atributo fundamental para a classificação de gravidez ou não-gravidez.

Claro que, até agora, isso é uma mera expectativa – ainda não é possível dizer que o algoritmo vai acertar na maioria das vezes só porque nós identificamos algumas correlações em uma amostra pequena. Mas, então, foram identificados 25 produtos que – quando analisados em conjunto – permitissem a atribuição de um peso ou uma pontuação. *Como assim, professor?* Basicamente, cada produto (atributo) analisado possui um peso diferente. *Concordam?*

Comprar fraldas é um atributo com um peso/pontuação muito maior do que comprar creme hidratante! Bem, à medida que os computadores começaram a processar os dados, chegou-se a uma pontuação em relação à probabilidade de gravidez de cada compradora! **Mais que isso: foi também possível estimar sua data de nascimento dentro de uma pequena janela, para que a Target pudesse enviar cupons cronometrados para fases mais específicas da gravidez.**

Para testar, criou-se uma pessoa fictícia no banco de dados chamada Jennifer Simpson que tinha 23 anos, morava em Atlanta e em março havia comprado um creme hidratante de cacau, uma bolsa grande o suficiente para caber um pacote de fraldas, suplementos de zinco e magnésio e um tapete azul brilhante. **O algoritmo estimou que havia uma chance de 87% de que ela estivesse grávida e que o bebê nasceria em algum momento no final de agosto**.

Chegou o momento então de colocar o algoritmo para rodar na base histórica inteira! Isso foi feito e a Target começou a enviar cupons de desconto para itens de bebê pelos correios para clientes que nunca haviam se cadastrado como grávidas – tudo baseado na pontuação de seus algoritmos em relação à classificação realizada. **Galera, houve um caso em que um homem irritado entrou em uma Target de Minneapolis exigindo falar com o gerente**.

"*Minha filha recebeu isso pelo correio*", disse ele. "*Ela ainda está no ensino médio e vocês estão enviando cupons para roupas de bebê e berços? Vocês estão tentando incentivá-la a engravidar?".* O gerente não tinha ideia do que o homem estava falando. **Ele olhou o cupom e viu que o endereço estava correto, realmente era para a filha daquele senhor e continha anúncios de roupas de maternidade, móveis de criança e fotos de bebês sorridentes**.

O gerente pediu mil desculpas e se despediu do pai da menina. **No dia seguinte, ele – não satisfeito – fez questão de ligar para aquele senhor e pedir desculpas mais uma vez.** Conta-se a história que ao receber o telefone de desculpas, o pai ficou envergonhado: "*Eu tive uma conversa com minha filha*", disse ele. "*Acontece que tem havido algumas atividades em minha casa que eu não conhecia completamente. Ela está grávida, prevista para agosto e eu que te devo desculpas*"[9].

[9] *Quem aí já ouviu falar no livro O Poder do Hábito?* Pois bem... o capítulo 7 desse livro explica com mais detalhes toda essa história que eu acabei de contar. Fica aqui a recomendação de leitura :)

Voltando ao tema de classificação! Companhias de Seguro utilizam a classificação para adivinhar quais pacientes idosos morrerão em breve; médicos usam para verificar se bebês prematuros estão desenvolvendo infecções perigosas (já que o classificador pode colocar indicadores sutis de doenças antes que os humanos notem quaisquer sinais); **enfim, há infinitos exemplos de utilização da classificação como técnica de mineração de dados.**

Em suma, podemos dizer que a técnica de classificação utiliza um algoritmo de aprendizado supervisionado a fim de distribuir um conjunto de dados de entrada em categorias ou classes pré-definidas de saída para realizar a análise de dados. **Constroem-se modelos de classificação a partir de um conjunto de dados de entrada, identificando cada classe por meio de múltiplos atributos e os rotulando/etiquetando – sendo essa técnica possível de ser utilizada com outras técnicas!**

(DPE/AM – 2018) Dentre os algoritmos utilizados em data mining, há um algoritmo que visa o estabelecimento de categorias, a partir do conjunto de dados, bem como a distribuição dos dados nas categorias estabelecidas. Essa descrição corresponde aos algoritmos de:

a) classificação.
b) sumarização.
c) visualização.
d) evolução.
e) detecção de desvios.

_____

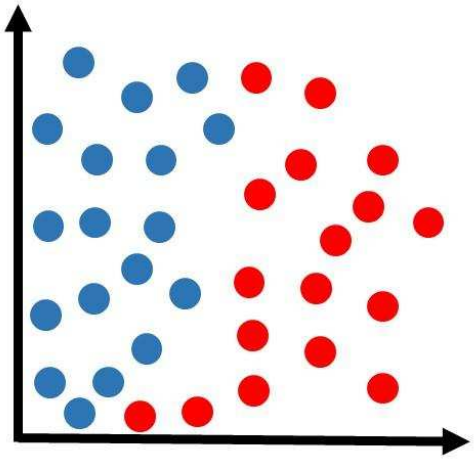
Comentários: o algoritmo que visa o estabelecimento de categorias é o algoritmo de classificação (Letra A).

Classificador k-NN

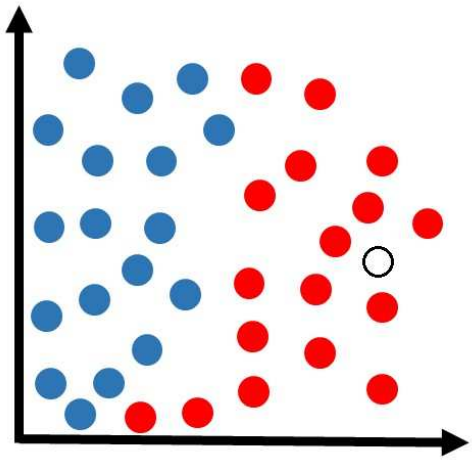
INCIDÊNCIA EM PROVA: BAIXÍSSIMA

Vamos começar falando sobre um dos mais simples algoritmos de classificação: k-NN (k-Nearest Neighbours) – também chamado de k-Vizinhos Mais Próximos. **Trata-se de uma técnica que basicamente separa dados conhecidos em diversas classes a fim de prever a classificação de um novo dado baseado em valores conhecidos mais similares ou mais próximos em termos de distância entre as características.** Calma que ficará claro...

Imagine que recebemos um conjunto de dados (*dataset*) dividido em duas classes distintas: azul e vermelha. Note que já sabemos – de antemão – a classificação das bolinhas dessa amostra:



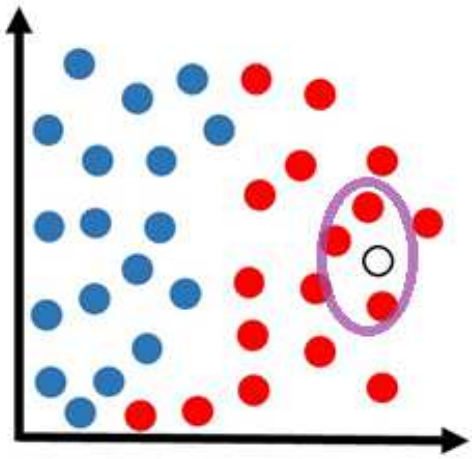
Só que agora eu vou adicionar uma bolinha nova nesse conjunto de dados que eu não sei ainda qual será sua classificação. Aliás, meu objetivo é descobrir qual será a classificação dessa bolinha nova:



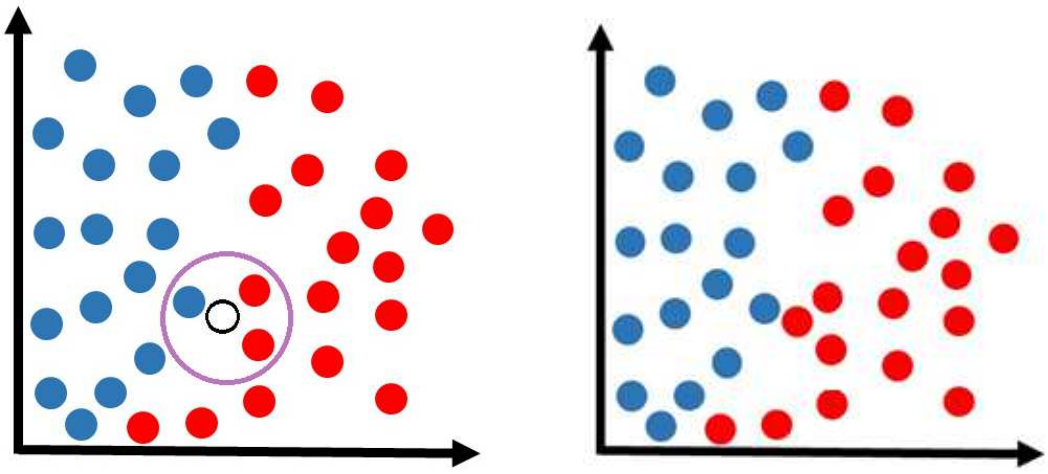
Vocês hão de concordar comigo que – de forma intuitiva – já dá para imaginar que ela tende a ser vermelha. Ora, ela está rodeada de bolinhas vizinhas vermelhas e está mais distante das bolinhas azuis. De toda forma, vamos utilizar o k-NN para fazer essa previsão de forma matemática! *Vocês*

já sabem que a tradução do nome do algoritmo é k-Vizinhos Mais Próximos, mas o que seria o k? K é um valor arbitrário escolhido pelo supervisor que indica a quantidade de vizinhos.

Vamos supor que eu seja o supervisor e escolha o valor k = 2, logo o algoritmo procurará os 2 vizinhos mais próximos do dado que queremos prever a classificação; se eu escolher o valor o valor k = 3, o algoritmo procurará os 3 vizinhos mais próximos do dado que queremos prever a classificação; e assim por diante. No exemplo abaixo, nós procuramos os três vizinhos mais próximos. Vejam só:

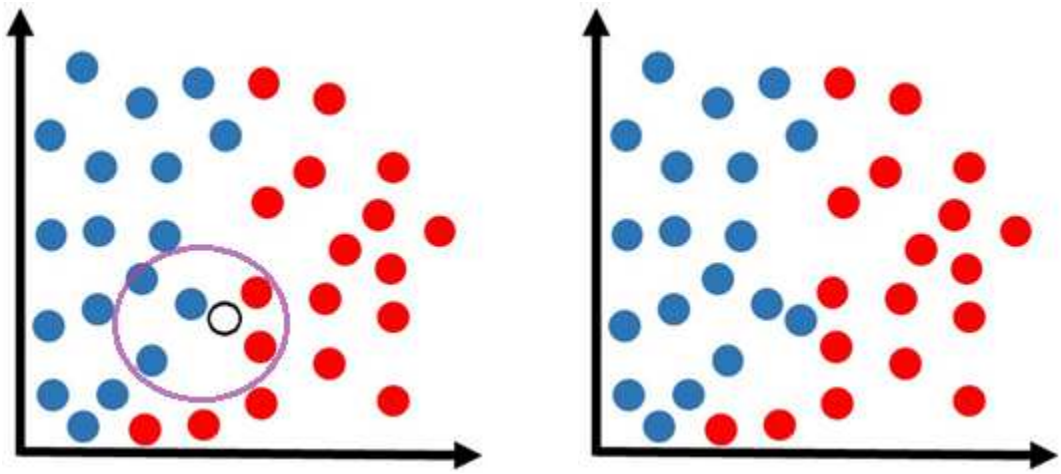


Agora nós faremos uma contagem: nós vamos olhar os três vizinhos mais próximos e vamos contar quantas bolinhas são azuis e quantas bolinhas são vermelhas. Ora, esse exemplo é muito fácil: as três bolinhas ao redor do dado que queremos classificar são vermelhas e nenhuma é azul, então o placar ficou 3x0, logo o algoritmo vai estimar que o dado que queremos classificar será também uma bolinha vermelha. Agora vamos ver um exemplo um pouquinho mais difícil:



Note que também temos k = 3-Vizinhos Mais Próximos. Se fizermos a contagem das classes dos vizinhos mais próximos, o placar ficará em 2x1 porque temos duas bolinhas vermelhas e uma bolinha azul. Logo, o algoritmo vai estimar que o dado que queremos classificar será também uma

bolinha vermelha. *Simples, né?* Então, vamos complicar um pouquinho mais nosso exemplo alterando o valor de k para k = 5:



Opa! Agora o placar mudou: ficou 3x2 para o azul, logo o algoritmo vai estimar que o dado que queremos classificar será também uma bolinha azul. *Você deve estar se perguntando: e se o valor de k for par, não corre o risco de acontecer um empate?* Ocorre, sim! Aí temos algumas opções, quais sejam: a classe da instância mais próxima é atribuída ao novo dado ou simplesmente se atribui uma classe aleatória. *Entendido?*

Agora vamos falar sobre alguns conceitos mais técnicos: k-NN é um classificador de aprendizado supervisionado não paramétrico baseado em distância que pode ser utilizado tanto para classificação quanto para regressão[10]. *Por que é baseado em distância?* Porque o algoritmo se baseia no grau de similaridade entre diversas observações em função de suas características ou variáveis, isto é, calcula a distância de cada um dos pontos em relação ao novo dado que desejamos classificar.

No exemplo, ficou fácil de ver quais são as bolinhas mais próximas, mas o algoritmo utiliza fórmulas de distância euclidiana, distância de cossenos, entre outros para realizar essas medições. *E por que ele é chamado de paramétrico, Diego?* Porque o algoritmo não pressupõe que a relação entre as entradas e as saídas de dados sigam uma função matemática específica, isto é, para cada novo dado que se deseja classificar, utiliza-se o conjunto de dados de treinamento original para calcular.

Complicou um pouco né? Veja só: se eu digo que um conjunto de dados possui uma relação linear entre dados de entrada e saída, eu não preciso armazenar os dados de treinamento para realizar a previsão de classificação de um novo dado. *Por que?* Porque se o conjunto de dados de treinamento segue uma função específica (Ex: linear), basta eu utilizar a função para realizar a previsão da classificação do novo dado – a isso, chamamos de métodos paramétricos.

Já os métodos não paramétricos não seguem uma função matemática específica. Toda vez que eu quiser realizar a previsão de classificação de um novo dado, eu vou ter que olhar para todos os dados

[10] Quando o k-NN é utilizado para classificação, contabiliza-se a classificação mais comum dos vizinhos mais próximos para estimar a classificação do novo dado; quando ele é utilizado para regressão, contabiliza-se a média dos valores dos vizinhos mais próximos para estimar a classificação.

de treinamento – a isso, chamamos de métodos não paramétricos. *O k-NN é um método paramétrico ou não paramétrico?* Como ele não segue nenhuma função matemática específica, ele é considerado um método não paramétrico! Prosseguindo...

Ele também utiliza o que chamamos de aprendizado preguiçoso (*lazy learning*) justamente porque não é necessária uma etapa de treinamento do modelo para somente depois realizar a previsão. Todo o trabalho ocorre no momento em que uma nova previsão é solicitada. É importante falar também sobre o valor de k. Um valor de k pequeno demais gera baixo viés e alta variância (*overfitting*); e um valor k grande demais gera alto viés e baixa variância (*underfitting*).

Ainda não veremos esses conceitos com profundidade. Tudo que você precisa saber por enquanto é que, embora não exista uma regra para se definir k, valores grandes reduzem o efeito dos ruídos na classificação, mas tornam fronteiras de classe menos definidas; já valores pequenos geram modelos mais complexos, o que também traz consequências. Em geral, o valor de k é definido por alguma heurística ou simplesmente por tentativa e erro.

Em outras palavras, o tamanho da vizinhança utilizada para realizar a previsão é arbitrário e definido experimentalmente. Testa-se o modelo com diversos valores de k até encontrar o tamanho ótimo de vizinhança que leva ao menor erro médio. O k-NN é um algoritmo simples com boa acurácia e versátil (pode ser utilizando tanto para classificação quanto para regressão), por outro lado é computacionalmente caro e necessita armazenar os dados de treinamento.

Em suma: o k-NN opera de forma que, dado um objeto x_0 cuja classe se deseja inferir, encontram-se os k objetos x_i, i = 1, ..., *k* da base que estejam mais próximos a x_0 e, depois, se classifica o objeto x_0 como pertencente à classe da maioria dos *k* vizinhos. Logo, ele determina a classe de um objeto com base na classe de outros objetos (instâncias), sendo chamado – portanto – de algoritmo baseado em instância. Vejamos uma questão para solidificar o entendimento...

(TCU – 2022) Um analista do TCU recebe o conjunto de dados com covariáveis e a classe a que cada amostra pertence na tabela a seguir.

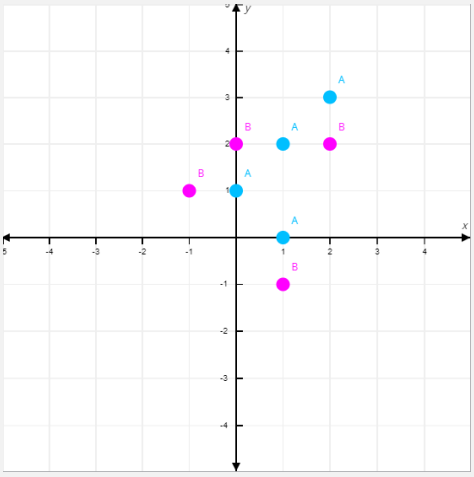
X_1	X_2	Classe
0	1	A
0	2	B
1	0	A
1	-1	B
2	2	B
1	2	A
-1	1	B
2	3	A

Esse analista gostaria de prever a classe dos pontos (1,1), (0,0) e (-1,2) usando o algoritmo de k-vizinhos mais próximos com k=3 e usando a distância euclidiana usual. Suas classes previstas são, respectivamente:

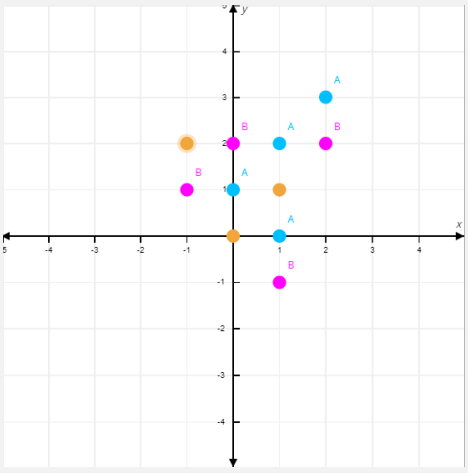
a) A, B, A;
b) B, A, A;
c) A, B, B;
d) A, A, B;
e) A, A, A.

Comentários:

Vamos lá! Essa questão parece difícil, mas é relativamente tranquila. O que a questão quer dizer é: eu vou te passar um conjunto de coordenadas de pontos de um plano cartesiano com suas respectivas classes. Baseado nele, eu gostaria de saber quais são as classes previstas para três outros pontos: (1,1), (0,0) e (-1,2). Para tal, eu quero que você veja a frequência de classes dos três pontos mais próximos desses pontos apresentados. Para resolver, nós vamos – em primeiro lugar – plotar os pontos exibidos na tabela do enunciado em um plano cartesiano:



Note que nós simplesmente plotamos os pontos e exibimos suas respectivas classes (Classe A = Azul; Classe B = Rosa). Agora vamos plotar os três pontos que se deseja prever as classes:



Veja que – em laranja – são exibidos os três pontos (1,1), (0,0) e (-1,2). Para descobrir à qual classe esses pontos se referem, nós vamos buscar os três pontos mais próximos e analisar suas classes:

- Pontos mais próximos de (1,1):
(1,0) = A; (0,1) = A; (1,2) = A → logo, temos 3 ocorrências de A;

- Pontos mais próximos de (0,0):
(1,0) = A; (0,1) = A; (1,-1) = B ou (-1,1) = B → logo, temos 2 ocorrências de A e 1 ocorrência de B;

- Pontos mais próximos de (-1,2):
(0,2) = B; (-1,1) = B; (0,1) = A ⟶ logo, temos 2 ocorrências de B e 1 ocorrência de A;

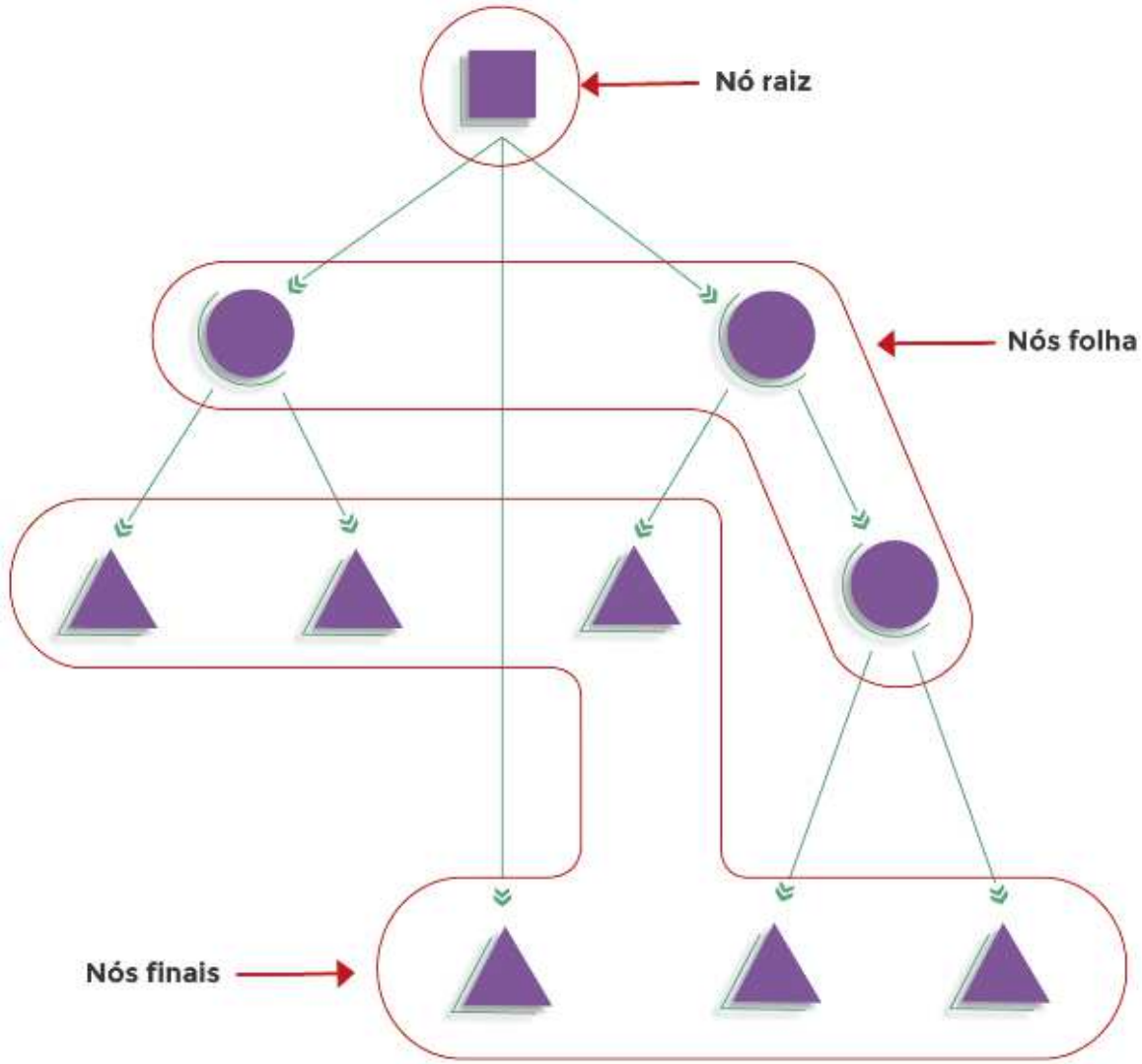
Dessa forma, podemos pegar agora apenas as classes com maior frequência para cada ponto. Assim, teremos respectivamente A, A, B (Letra D)

Árvores de Decisão

INCIDÊNCIA EM PROVA: BAIXÍSSIMA

Uma das principais ferramentas de classificação é a árvore de decisão. *O que é isso, Diego?* **É basicamente uma representação gráfica de regras de classificação[11]!** Elas demonstram visualmente as condições para categorizar dados por meio de uma estrutura que contém nó raiz, nós folha e nós finais. Na imagem seguinte, eles elementos estão representados respectivamente por um quadrado, um círculo e um triângulo.

É possível atravessar a árvore de decisão partindo do nó raiz até cada folha por meio de diversas regras de decisão – lembrando que o destino final (nó final) contém sempre uma das classes pré-definidas. É importante destacar também que uma árvore de decisão pode ser utilizada tanto para classificação quanto para regressão, mas o nosso foco aqui será na árvore de classificação porque nossas classes são **categóricas** e **finitas** e, não, **contínuas** e **infinitas**.



[11] Regras de Classificação são regras de fácil interpretação no estilo se (antecedente), então (consequente) em que o antecedente é um conjunto de testes e o consequente é a classe ou distribuição de probabilidades sobre as classes.

Cada nó interno denota um teste de um atributo, cada ramificação denota o resultado de um teste e cada nó folha apresenta um rótulo de uma classe definida de antemão por um supervisor. O objetivo dessa técnica é criar uma árvore que verifica cada um dos testes até chegar a uma folha, que representa a categoria, classe ou rótulo do item avaliado. *O que isso tem a ver com aprendizado de máquina, Diego?* A árvore de decisão, por si só, não tem a ver com aprendizado de máquina...

No entanto, seu processo de construção automático e recursivo a partir de um conjunto de dados pode ser considerado um algoritmo de aprendizado de máquina. O processo de construção do modelo de uma árvore de decisão se chama indução e busca fazer diversas divisões ou particionamentos dos dados em subconjuntos de forma automática, de modo que os subconjuntos sejam cada vez mais homogêneos. *Como assim, Diego?*

Imagine que eu tenha uma tabela com diversas linhas e colunas, em que as linhas representam pessoas que desejam obter um cartão de crédito e as linhas representam atributos ou variáveis dessas pessoas. Para que uma operadora de cartão decida se vai disponibilizar um cartão de crédito para uma pessoa ou não, ela deve avaliar qual é o risco de tomar um calote dessa pessoa. Logo, a nossa árvore de decisão buscará analisar variáveis para classificar o risco de calote de uma pessoa.

Nós já sabemos que as árvores de decisão são uma das ferramentas do algoritmo de classificação e também sabemos que algoritmos de classificação são supervisionados. Dessa forma, podemos concluir que a árvore de decisão necessita de um supervisor externo para treinar o algoritmo e indicar, de antemão, quais serão as categorias que ele deve classificar uma pessoa. Para o nosso exemplo, vamos assumir que as categorias/classes sejam: Risco Baixo, Risco Médio ou Risco Alto.

Legal! Então, o algoritmo de aprendizado de máquina vai analisar um conjunto de variáveis ou atributos de diversas pessoas e categorizá-las em uma dessas três classes possíveis. *E como o algoritmo vai descobrir quais são as variáveis mais importantes?* De fato, não é uma tarefa simples – ainda mais se existirem muitas variáveis! *A idade é mais relevante para definir o risco de calote de uma pessoa do que seu salário anual?*

O estado civil é mais relevante para definir o risco de calote de uma pessoa do que seu saldo em conta? Em que ordem devemos avaliar cada variável? Para humanos, essas perguntas são extremamente difíceis de responder, mas é aí que entre em cena o aprendizado de máquina! Existe uma lógica interna de construção da árvore de decisão que automaticamente pondera a contribuição de cada uma das variáveis com base em dados históricos. *Como assim, Diego?*

Você pode fazer a máquina aprender oferecendo para ela uma lista que contenha os valores dessas variáveis referentes a diversos clientes antigos e uma coluna extra que indique se esses clientes deram calote na operadora ou não. O algoritmo da árvore de decisão analisará cada uma dessas variáveis (e seus possíveis pontos de corte) a fim de descobrir quais são as melhores para realizar o particionamento dos dados de modo que se formem dois subgrupos mais homogêneos possíveis.

O que você quer dizer com grupos mais homogêneos possíveis, professor? Galera, esse é o momento em que o algoritmo fará diversos testes: ele poderá começar analisando, por exemplo, a variável de estado civil. Ele separa as pessoas em dois subgrupos (casados e não-casados) e verifica qual é a porcentagem de casados caloteiros e não-casados caloteiros. Pronto, ele anota esse valor para poder fazer diversas comparações posteriormente.

Depois ele pode analisar a idade: separa em dois subgrupos (< 25 anos e >= 25 anos) e verifica qual é a porcentagem de < 25 anos caloteiros e >= 25 anos caloteiros. Pronto, ele anota esse valor também. *Professor, por que 25 anos?* Eu dei apenas um chute, mas o algoritmo pode fazer diversos testes e verificar o melhor corte. *Vocês entenderam a lógica?* Pois é, o algoritmo fará isso para todas as variáveis!

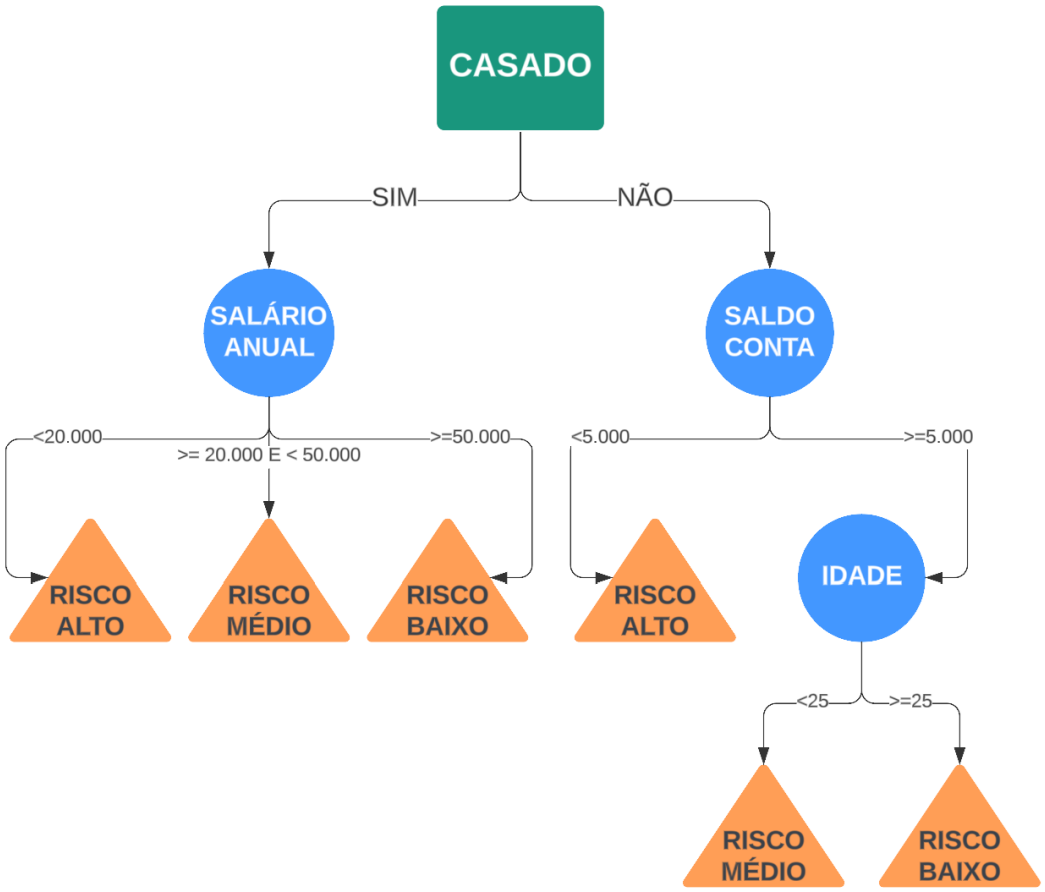
Agora vem o pulo do gato: eu disse para vocês que os subgrupos formados devem ser os mais homogêneos possíveis. Quando ele dividiu a primeira variável em casados caloteiros e não-casados caloteiros, ele pode ter descoberto que havia muito mais não-casados caloteiros do que casados caloteiros. Logo, o fato de uma pessoa ser casada tende a reduzir seu risco de calote. Pessoal, é claro que isso aqui é uma simplificação...

O algoritmo vai fazer milhares de testes com cada uma das variáveis, vai testar diversos pontos de corte diferentes e diversas sequências de análise de variáveis diferentes. O que importa aqui é que nós vamos sair de um grupo muito misturado (menos homogêneo) para dois subgrupos menos misturados (mais homogêneos). *Sabe qual é o nome disso no contexto de ciência de dados?* Ganho de Informação ou Redução de Entropia.

Quem sabe o que é entropia? Entropia é a uma medida que nos diz o quanto um conjunto de dados está desorganizado ou misturado. Ora, toda vez que nós particionamos os dados em subgrupos, nós obtemos dados mais homogêneos e organizados, logo nós reduzimos a entropia. O que a nossa árvore de decisão busca fazer é pegar um conjunto de dados e encontrar um conjunto de regras sobre variáveis ou pontos de corte que permite separar esses dados em grupos mais homogêneos.

Note que sempre que dividimos um grupo em outros subgrupos, esses subgrupos serão mais puros à medida que seus dados forem mais homogêneos. Em outras palavras, quanto mais homogêneos forem os dados de um subconjunto, mais puros serão seus dados. Para calcular a pureza de um subgrupo, isto é, quão homogêneos são seus dados, existem diversas métricas (Ex: Índice de Gini, Redução de Variância, etc).

Ao final do treinamento da máquina, chegamos a um possível modelo de árvore de decisão efetivamente construído e disponível para ser utilizado com novos clientes. Vejamos:



O algoritmo inicialmente faz uma avaliação se a pessoa é solteira ou casada. Caso ela casada: se seu salário anual for abaixo de 20.000, é classificada como risco alto; se seu salário anual for entre 20.000 e 50.000, é classificada como risco médio; se seu salário anual for maior que 50.000, é classificada como risco baixo. Caso não seja casada: se seu saldo em conta for abaixo de 5.000, é classificada como risco alto; se seu saldo em conta for acima de 5.000, analisa-se sua idade.

Se ela não é casada, possui mais de 5.000 em conta, mas tem menos de 25 anos, é classificada como risco médio; se ela não é casada, possui mais de 5.000 em conta, mas tem mais de 25 anos, é classificada como risco baixo. *Gostaram da nossa árvore?* Acho que ela faz sentido! Parece ser um bom modelo para avaliar risco de calote e decidir se deve receber um cartão de crédito ou não. *Ela é perfeita?* Não, nenhum modelo será perfeito!

Eu disse páginas atrás que o processo de construção do modelo de uma árvore de decisão busca fazer diversas divisões ou particionamentos dos dados em subconjuntos de forma automática, de modo que os subconjuntos sejam cada vez mais homogêneos. Ora, imaginem um contexto em que tenhamos dezenas de variáveis! Se o objetivo do algoritmo é obter automaticamente subconjuntos cada vez mais homogêneos, então temos um problema grave...

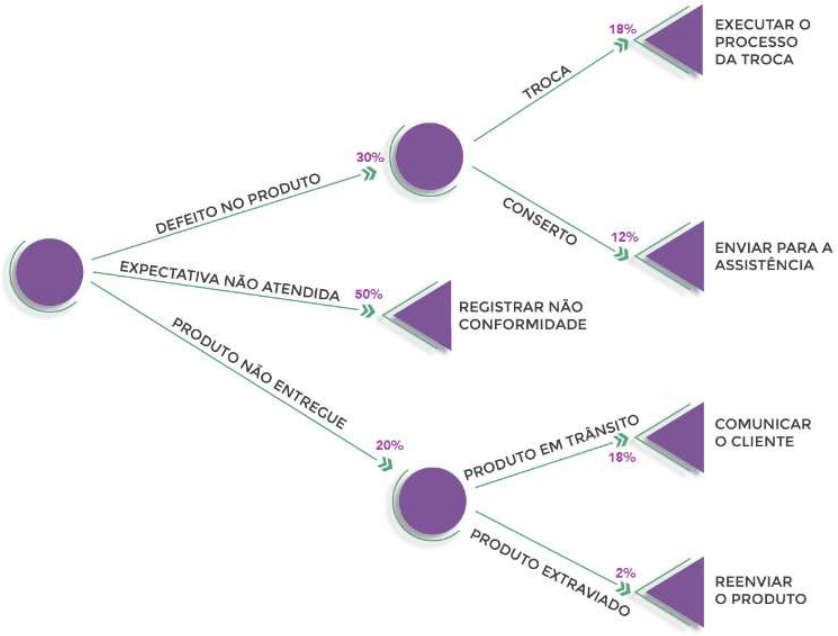
Quando o algoritmo vai parar de fazer as divisões? Em tese, ele pode ir dividindo, dividindo, dividindo indefinidamente até que – no pior caso – tenhamos um único dado para cada nó folha. O nome desse fenômeno é *overfitting*! Ele deve ser evitado porque pode tornar o modelo de árvore de

decisão completamente inútil. *E o que devemos fazer, professor?* É necessário estabelecer um limite para as divisões...

Há diversas maneiras: nós podemos definir uma altura/profundidade máxima da árvore – quando esse limite for atingido, interrompe as subdivisões; nós também podemos realizar a poda da árvore – deixamos a árvore crescer quanto quiser e depois vamos reduzindo as divisões que sejam pouco significativas (é como se realmente fizéssemos uma poda de uma árvore). Por fim, vamos ver algumas vantagens e desvantagens:

VANTAGENS DE ÁRVORES DE DECISÃO
As árvores de decisão podem gerar regras compreensíveis e executam a classificação sem exigir muitos cálculos, sendo capazes de lidar com variáveis contínuas e categóricas.
As árvores de decisão fornecem uma indicação clara de quais campos são mais importantes para predição ou classificação.
Por meio da técnica de estratificação (estrato = camadas), é capaz designar regras para cada caso a uma dentre várias categorias diferentes.

DESVANTAGENS DE ÁRVORES DE DECISÃO
As árvores de decisão são bastante propensas ao overfitting dos dados de treinamento.
Uma única árvore de decisão normalmente não faz grandes previsões, portanto várias árvores são frequentemente combinadas em forma de florestas chamadas (Random Forests).
Somente se as informações forem precisas e exatas, a árvore de decisão fornecerá resultados promissores. Mesmo se houver uma pequena alteração nos dados de entrada, isso pode causar grandes alterações na árvore.
Se o conjunto de dados é enorme, com muitas colunas e linhas, é uma tarefa muito complexa projetar uma árvore de decisão com muitos ramos.
Se uma das regras do modelo estiver incorreta, isso gerará divisões equivocadas da árvore, fazendo com que o erro se propague por todo o resto da árvore.
As árvores de decisão são menos apropriadas para tarefas de estimativa em que o objetivo é prever o valor de um atributo contínuo.
As árvores de decisão estão sujeitas a erros em problemas de classificação com muitas classes e um número relativamente pequeno de exemplos de treinamento.



É importante mencionar que esse algoritmo é capaz de classificar dados dentre de um conjunto finito de classes com base em valores de entrada por meio de uma abordagem chamada **estratificação**, que permite determinar as regras para que se possa designar ou direcionar cada caso a uma categoria pré-existente, separando-os em níveis diferentes (Ex: executar o processo da troca, enviar para a assistência, comunicar o cliente, reenviar o produto). *Bacana?*

(MDA – 2014) Nos processos de Data Mining, a partir de uma massa de dados, uma técnica estatística cria e organiza regras de classificação em formato de diagramas, que vão ordenar suas observações ou predizer resultados futuros. Uma das abordagens empregadas nessa técnica é a estratificação, que determina regras para que se possa designar cada caso a uma dentre várias categorias existentes, como, por exemplo, classificar um cliente tomador de crédito em um grupo de elevado, médio ou baixo risco. Essa técnica estatística é denominada:

a) diagrama de regressão.
b) gráfico de estrutura.
c) árvore de decisão.
d) rede neural.
e) histograma.

_____

Comentários: a técnica que cria e organiza regras de classificação e estratificação é a árvore de decisão (Letra C).

Os algoritmos de árvore de decisão mais comuns são: ID3, c4.5 e CART. O algoritmo ID3 é utilizado para gerar árvores a partir de um conjunto de dados. É usado em problemas de aprendizado supervisionado, onde os dados de entrada são divididos em categorias com base em certas condições. O algoritmo funciona selecionando o melhor atributo de um determinado conjunto de atributos para dividir os dados em dois (ou mais) subconjuntos.

Isso é feito de forma iterativa, onde – a cada passo do algoritmo – o melhor atributo é escolhido e os dados são divididos em dois (ou mais) subconjuntos. Este processo é repetido até que todos os dados sejam divididos em categorias. Depois que a árvore é gerada, ela pode ser usada para classificar novos pontos de dados seguindo o caminho da árvore até que o ponto de dados seja associado a uma categoria.

Ela trabalha apenas com atributos categóricos/qualitativos (Ex: Masculino ou Feminino), construindo a árvore a partir do nó raiz de cima para baixo recursivamente por meio do método dividir-para-conquistar. Já o algoritmo c4.5 é uma evolução do ID3 capaz de trabalhar tanto com atributos categóricos quanto numéricos/quantitativos (Ex: 95kg). Ele funciona dividindo iterativamente os dados em subconjuntos com base no atributo que melhor separa os dados.

A cada divisão, o C4.5 calcula a impureza de cada um dos subconjuntos e seleciona o atributo que produz os subconjuntos mais puros, isto é, aqueles que contêm apenas dados de apenas uma classe ou rótulo. O algoritmo então repete o processo até que todos os dados tenham sido divididos em subconjuntos puros ou até que um critério de parada seja atingido. Além disso, trata-se de um algoritmo bem mais rápido que o anterior.

Por fim, o algoritmo CART (*Classification and Regression Trees*) é um tipo de algoritmo de árvore de decisão utilizado para problemas de classificação e regressão. Ele funciona construindo uma árvore de decisão durante a fase de treinamento, que é usada para fazer previsões sobre dados não vistos. O objetivo do algoritmo é criar um modelo que preveja com precisão o valor alvo, além de ter o menor número possível de divisões.

Ele funciona dividindo recursivamente os dados de treinamento em subconjuntos menores com base na variável independente mais significativa. A árvore para de crescer quando atinge uma profundidade máxima especificada ou quando todas as amostras restantes pertencem à mesma classe. O algoritmo CART é capaz de trabalhar tanto com atributos categóricos/qualitativos quanto numéricos/quantitativos.

(MPE/AL – 2012 – II) O algoritmo ID3 é considerado uma evolução do algoritmo C4.5 e muito utilizado em mineração de dados.

_____

Comentários: na verdade, o algoritmo C4.5 é considerado uma evolução do algoritmo ID3 (Errado).

(TJ/RO – 2012 – A) A técnica de clustering em data mining utiliza os algoritmos ID3 e o C4.5. Esses algoritmos produzem árvores de decisão, o que permite gerar clusters de elementos que, por sua vez, geram mapeamento dos elementos em grupos predefinidos.

_____

Comentários: ID3 e C4.5 são algoritmos que realmente produzem árvores de decisão, mas são considerados uma técnica de classificação e, não, *clustering* (Errado).

(MPE/AL – 2012) Os algoritmos C4.5 e K-Means, muito utilizados para descoberta de conhecimento através de mineração de dados, são algoritmos de respectivamente:

a) agrupamento e agrupamento (clustering).
b) classificação e regras de associação.
c) regras de associação e classificação.
d) classificação e agrupamento (clustering).
e) agrupamento (clustering) e classificação.

_____

Comentários: C4.5 é um algoritmo de classificação e K-Means (veremos adiante) é um algoritmo de agrupamento (Letra D).

Florestas Aleatórias

RELEVÂNCIA EM PROVA: BAIXÍSSIMA

Árvores de decisão são ferramentas interessantes, mas não fazem previsões com acurácia. Em outras palavras, elas funcionam bem com dados de treinamento, mas não são tão flexíveis quando utilizadas para classificar novas amostras. Já florestas aleatórias (*random forests*) permitem combinar a simplicidade de árvores de decisão com flexibilidade para melhorar significativamente a acurácia das previsões, sendo utilizadas tanto para classificação quanto para regressão.

DOR NO PEITO	BOA CIRCULAÇÃO	ARTÉRIAS BLOQUEADAS	PESO	DOENÇA CARDÍACA
Não	Não	Não	125	Não
Sim	Sim	Sim	180	Sim
Sim	Sim	Não	210	Não
Sim	Não	Sim	167	Sim

Vamos ver como isso funciona? Imagine que nós tenhamos um conjunto de dados (*dataset*) original com diversas variáveis mostrado acima. Porém, em nosso primeiro passo, precisaremos criar um *bootstrapped dataset*. *O que diabos é isso, Diego?* É um conjunto de dados retirados do conjunto de dados original (inclusive com o mesmo tamanho do conjunto de dados original), mas com amostras aleatórias dos dados. Aqui já temos um ponto de atenção...

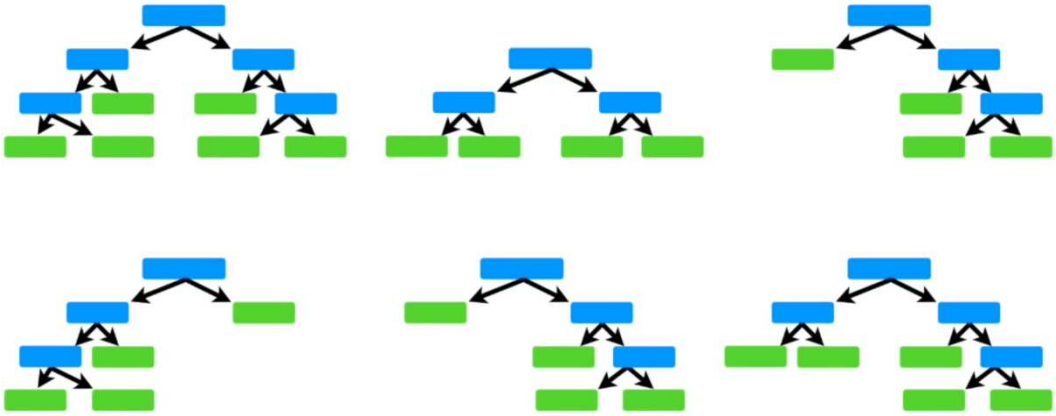
Você deve estar imaginando: se o *bootsrapped dataset* é um conjunto de dados **retirado** do conjunto de dados original e tem o **mesmo tamanho**, então deve ser uma cópia dos dados originais. Não, porque as amostras de dados retirada do conjunto de dados original pode conter amostras repetidas. Vejam um exemplo na tabela a seguir: a primeira linha é igual a segunda linha do *dataset* original; a segunda linha é igual a primeira do *dataset* original.

DATASET ORIGINAL				
DOR NO PEITO	BOA CIRCULAÇÃO	ARTÉRIAS BLOQUEADAS	PESO	DOENÇA CARDÍACA
Não	Não	Não	125	Não
Sim	Sim	Sim	180	Sim
Sim	Sim	Não	210	Não
Sim	Não	Sim	167	Sim

BOOTSTRAPPED DATASET				
DOR NO PEITO	BOA CIRCULAÇÃO	ARTÉRIAS BLOQUEADAS	PESO	DOENÇA CARDÍACA
Sim	Sim	Sim	180	Sim
Não	Não	Não	125	Não
Sim	Não	Sim	167	Sim
Sim	Não	Sim	167	Sim

Note que a quarta linha do conjunto de dados original foi repetida duas vezes e a terceira sequer foi selecionado. É por essa razão que o *bootstrapped dataset* é considerado um *dataset* retirado do *dataset* original e possui o mesmo tamanho. O segundo passo é criar uma árvore de decisão baseado no *bootstrap dataset*, mas com uma particularidade: nós vamos utilizar um subconjunto aleatório de colunas para cada nível da árvore.

Por exemplo: para a primeira árvore, nós vamos escolher duas variáveis quaisquer e verificar qual delas divide melhor as amostras. Vamos supor que tenhamos escolhido "Boa Circulação" e "Artérias Bloqueadas" e que, dentre essa duas, "Boa Circulação" divide melhor as amostras. Agora vamos para o próximo nível e fazemos os mesmos passos até utilizar todas as variáveis disponíveis. Essa iteração permitirá construir diversas árvores diferentes:



É a variedade que torna as florestas aleatórias mais efetivas que árvores de decisão utilizadas de forma individual. Bem, agora que nós temos uma enorme variedade de árvores de decisões (uma enorme variedade de árvores é uma... floresta), nós podemos tentar prever de uma determinada pessoa tem ou não doenças cardíacas. Considere abaixo os dados de uma nova pessoa que desejar prever doenças cardíacas:

NOVO PACIENTE				
DOR NO PEITO	BOA CIRCULAÇÃO	ARTÉRIAS BLOQUEADAS	PESO	DOENÇA CARDÍACA
Sim	Não	Não	168	?

Nós pegamos esses dados das variáveis e passamos por todas as árvores de decisão aleatórias criadas nos passos anteriores. Vamos supor que, de 100 árvores aleatórias, 75 indicaram que esse paciente possui doença cardíaca e 25 indicaram que ele não possui. Dessa forma, podemos prever que – sim – ele tem doença cardíaca. *Agora, como nós sabemos que a floresta aleatória fez um bom trabalho de previsão de dados?* Essa é a parte legal...

DATASET ORIGINAL				
DOR NO PEITO	BOA CIRCULAÇÃO	ARTÉRIAS BLOQUEADAS	PESO	DOENÇA CARDÍACA
Não	Não	Não	125	Não
Sim	Sim	Sim	180	Sim
Sim	Sim	Não	210	Não
Sim	Não	Sim	167	Sim

Vocês se lembram que no início da explicação eu disse que o bootstrapped dataset tinha o mesmo tamanho do dataset original e permitia dados duplicados? A consequência disso é que algumas amostras do *dataset* original não foram utilizadas (em geral, 1/3 das amostras de *datasets* originais não são utilizadas no *bootstrapped dataset*). Em nosso exemplo, a 3ª linha não foi utilizada, logo eu posso usar os dados reais das variáveis, passar pela floresta aleatória e ver se os resultados batem.

Assim, para saber a acurácia da previsão de dados, basta rodar os dados das variáveis da terceira linha e verificar se a floreta também retorna: **Não**. *Fechado?* Em relação às árvores de decisão individuais, podemos afirmar que esse algoritmo é capaz de fazer previsões mais acuradas, dado que ele combina previsões de diversas árvores de decisão treinadas em diferentes subconjuntos do conjunto de dados de treinamento original.

Ademais, ela é menos propensa a sofrer com *overfitting*, dado que árvores individuais geralmente se ajustam excessivamente aos dados de treinamento. Por fim, alguns conceitos importantes:

CARACTERÍSTICAS DAS FLORESTAS DE DADOS
A técnica que combina múltiplos *bootstrapping* de dados com uma agregação final é chamada de *bagging* (***b**ootstrapping **agg**regation*).
Florestas aleatórias são capazes de resolver problemas de classificação ou regressão e funcionam bem com variáveis categóricas ou contínuas.
A utilização de florestas aleatórias permite reduzir o *overfitting* (veremos mais à frente) ao reduzir a variância e, portanto, melhorar a acurácia;
Trata-se de um algoritmo bastante robusto a ruídos e dados anômalos, além de conseguir lidar bem com grandes conjuntos de dados.

(CRA/PR – 2022) Entre as técnicas de *machine learning*, a *random forest* é capaz de solucionar problemas de classificação e de regressão, por meio da construção e dos treinamentos de árvores de decisão.

_____

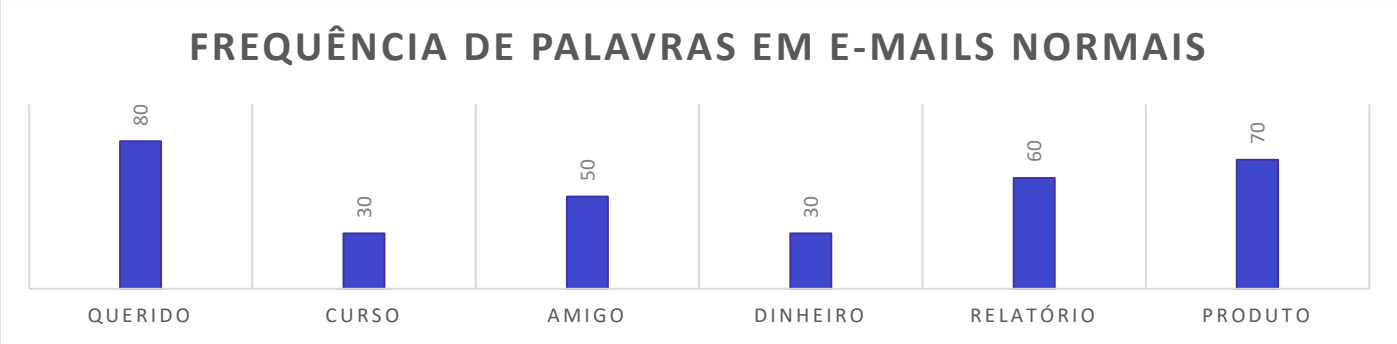
Comentários: esse método realmente é capaz de solucionar problemas de classificação/regressão por meio da construção e treinamento de árvores de decisão conforme acabamos de estudar (Correto).

Classificador Naive Bayes

RELEVÂNCIA EM PROVA: BAIXÍSSIMA

Para explicar o que é um Classificador Naive Bayes, eu preciso fazer uma contextualização! Deem uma olhada na caixa de entrada de seus e-mails agora. Em geral, há e-mails normais e e-mails de spam. Os e-mails normais são aquelas de amigos, colegas de trabalho e familiares; já os e-mails indesejados são aqueles de spam com tentativas de golpes ou anúncios não solicitados. *O que deveríamos fazer se quiséssemos filtrar as mensagens de spam?*

Nós podemos pegar o texto de todos os e-mails normais (aquelas de amigos, colegas e familiares) e criar um gráfico com a frequência de todas as palavras em suas mensagens. Exemplo...



Em seguida, nós podemos utilizar esse gráfico para calcular as probabilidades de encontrar cada palavra dado que elas estão presentes em um e-mail normal. Por exemplo: a probabilidade de vermos a palavra "Querido" dado que estamos vendo apenas e-mails normais é 80/320. *Por que?* Porque essa probabilidade é calculada pela frequência da palavra "querido" sobre a quantidade total de palavras contidas em todos os e-mails normais[12] da seguinte forma:

$$P(Normais) = \frac{Frequência(Querido)}{Total\ de\ Palavras} = \frac{80}{(80+30+50+30+60+70)} = \frac{80}{320} = 0{,}25$$

Se fizermos esse mesmo procedimento para as outras palavras contidas nos e-mails normais, vamos obter os seguintes resultados:

$$P(Normais) = \frac{Frequência(Curso)}{Total\ de\ Palavras} = \frac{30}{(80+30+50+30+60+70)} = \frac{30}{320} = 0{,}09$$

$$P(Normais) = \frac{Frequência(Amigo)}{Total\ de\ Palavras} = \frac{50}{(80+30+50+30+60+70)} = \frac{50}{320} = 0{,}16$$

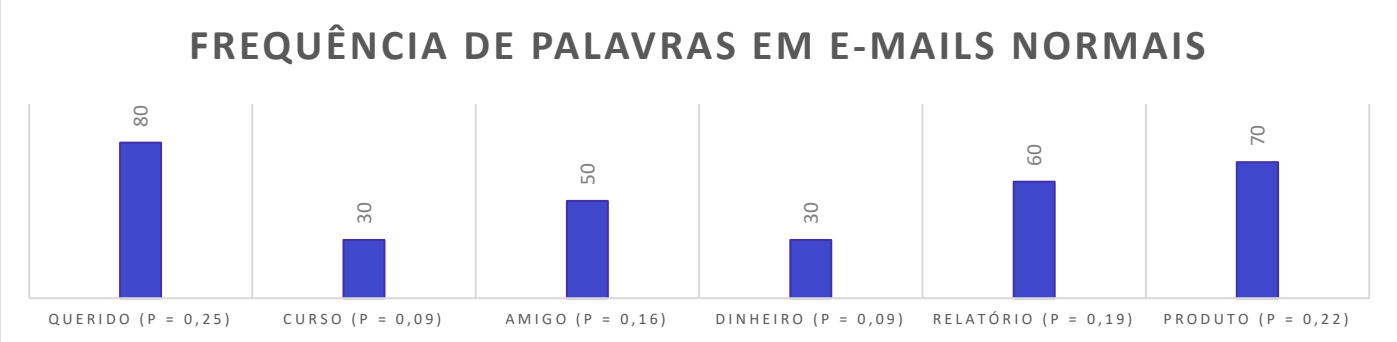
$$P(Normais) = \frac{Frequência(Dinheiro)}{Total\ de\ Palavras} = \frac{30}{(80+30+50+30+60+70)} = \frac{30}{320} = 0{,}09$$

[12] No exemplo, colocamos apenas algumas palavras no gráfico para facilitar a explicação didática, mas é óbvio que e-mails possuem outras palavras.

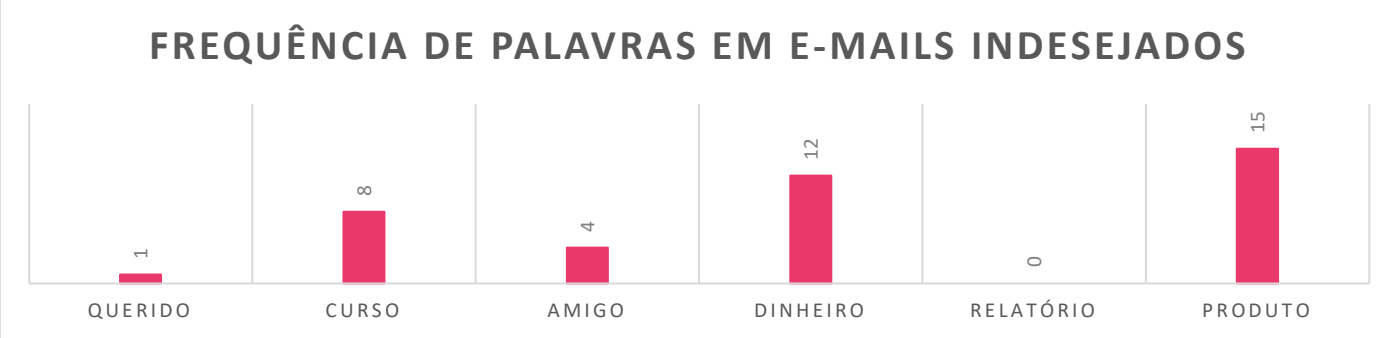
$$P(Normais) = \frac{Frequência(Relatório)}{Total\ de\ Palavras} = \frac{60}{(80+30+50+30+60+70)} = \frac{60}{320} = 0{,}19$$

$$P(Normais) = \frac{Frequência(Produto)}{Total\ de\ Palavras} = \frac{70}{(80+30+50+30+60+70)} = \frac{70}{320} = 0{,}22$$

Pronto! Agora para que nós não esqueçamos qual é a probabilidade de cada palavra, vamos colocá-las no gráfico de frequência de palavras em e-mails normais:



Agora vamos fazer a mesma coisa, mas para e-mails de spam! Para tal, pegamos o texto de todos os e-mails de spam e criamos um gráfico com a frequência de todas as palavras em suas mensagens:



Da mesma forma, nós podemos calcular a probabilidade de cada uma das palavras, porém agora dentro do contexto de e-mails indesejados[13]:

$$P(Indesejadas) = \frac{Frequência(Querido)}{Total\ de\ Palavras} = \frac{1}{(1+8+4\ +12+0+15)} = \frac{1}{40} = 0{,}03$$

Se fizermos esse mesmo procedimento para as outras palavras contidas nos e-mails indesejados, vamos obter os seguintes resultados:

$$P(Indesejadas) = \frac{Frequência(Curso)}{Total\ de\ Palavras} = \frac{8}{(1+8+4\ +12+0+15)} = \frac{8}{40} = 0{,}20$$

[13] Quando não há ocorrências de uma palavra nos dados de treinamento (Ex: Relatório = 0), é necessário utilizar técnicas de suavização para evitar que o resultado total seja 0. Uma das técnicas mais utilizadas é a Estimação de Laplace, que soma uma unidade para todas as frequências.

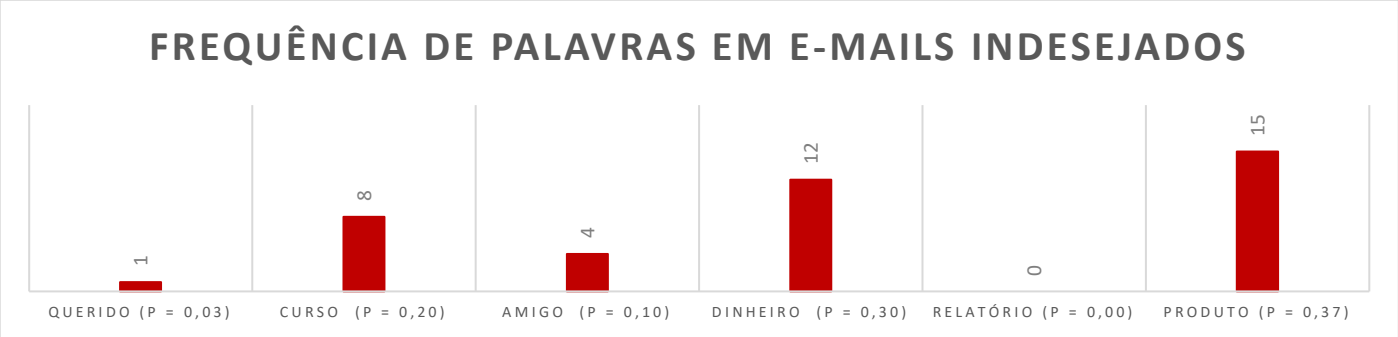
$$P(Indesejadas) = \frac{Frequência(Amigo)}{Total\ de\ Palavras} = \frac{4}{(1+8+4\ +12+0+15)} = \frac{4}{40} = 0{,}10$$

$$P(Indesejadas) = \frac{Frequência(Dinheiro)}{Total\ de\ Palavras} = \frac{12}{(1+8+4\ +12+0+15)} = \frac{12}{40} = 0{,}30$$

$$P(Indesejadas) = \frac{Frequência(Relatório)}{Total\ de\ Palavras} = \frac{0}{(1+8+4\ +12+0+15)} = \frac{0}{40} = 0{,}00$$

$$P(Indesejadas) = \frac{Frequência(Produto)}{Total\ de\ Palavras} = \frac{15}{(1+8+4\ +12+0+15)} = \frac{15}{40} = 0{,}37$$

Pronto! Agora para que nós não esqueçamos qual é a probabilidade de cada palavra, vamos colocá-las no gráfico de frequência de palavras em e-mails indesejados:



PROBABILIDADES DE E-MAILS NORMAIS		PROBABILIDADES DE E-MAILS INDESEJADOS	
P(QUERIDO | NORMAL)	0,25	P(QUERIDO | INDESEJADO)	0,03
P(CURSO | NORMAL)	0,09	P(CURSO | INDESEJADO)	0,20
P(AMIGO | NORMAL)	0,16	P(AMIGO | INDESEJADO)	0,10
P(DINHEIRO | NORMAL)	0,09	P(DINHEIRO | INDESEJADO)	0,30
P(RELATÓRIO | NORMAL)	0,19	P(RELATÓRIO | INDESEJADO)	0,00
P(PRODUTO | NORMAL)	0,22	P(PRODUTO | INDESEJADO)	0,37

Bacana! Agora imagine que eu acabo de receber uma mensagem que começa com:

"Querido amigo,

(...)"

Como saber se esse e-mail é normal ou indesejado? Vamos começar dando um palpite inicial sobre a probabilidade de um e-mail qualquer (independentemente do que diz em seu texto) seja um e-mail

normal. Como é um palpite, podemos sugerir qualquer probabilidade que quisermos (Ex: 50%), mas um palpite mais direcionado pode ser estimado a partir do nosso conjunto de dados de treinamento (Ex: 89%). *Como assim, Diego?*

Vamos lá! Estamos querendo descobrir a probabilidade de um determinado e-mail ser um e-mail normal. Logo, podemos chutar – por exemplo – que existe uma probabilidade de 50% desse e-mail ser um e-mail normal. Por outro lado, nós podemos dar um chute mais direcionado. Ora, se nosso conjunto de treinamento continha 320 e-mails normais e 40 e-mails indesejados, logo um palpite inicial mais razoável seria:

$$P(Normal) = \frac{Quantidade\ (Normal)}{Quantidade\ (Total)} = \frac{320}{360} = 0{,}89$$

Como queremos calcular a probabilidade de um e-mail que contenha as palavras "Querido amigo", devemos multiplicar nosso palpite inicial pela probabilidade de a palavra "*Querido*" ocorrer em um e-mail normal e multiplicar pela probabilidade de a palavra "*amigo*" ocorrer em um e-mail normal. Ora, nós temos essas probabilidades todas calculadas da etapa anterior em nossa tabelinha apresentada na página passada. Vamos fazer os cálculos...

$$P(Normal)\ x\ P(Normal)\ x\ P(Amigo|Normal) = 0{,}89\ x\ 0{,}25\ x\ 0{,}16 = 0{,}0356 = 3{,}56\%$$

Agora vamos repetir o procedimento para e-mails indesejados. Comecemos pelo cálculo do nosso palpite inicial:

$$P(Indesejado) = \frac{Quantidade\ (Indesejado)}{Quantidade\ (Total)} = \frac{40}{360} = 0{,}11$$

Como queremos calcular a probabilidade de um e-mail que contenha as palavras "Querido amigo", devemos multiplicar nosso palpite inicial pela probabilidade de a palavra "*Querido*" ocorrer em um e-mail indesejado e multiplicar pela probabilidade de a palavra "*amigo*" ocorrer em um e-mail indesejado. Ora, nós temos essas probabilidades todas calculadas da etapa anterior em nossa tabelinha apresentada na página passada. Vamos fazer os cálculos...

$$P(Indesejado)\ x\ P(Indesejado)\ x\ P(Amigo|Indesejado) = 0{,}11\ x\ 0{,}03\ x\ 0{,}1 = 0{,}00033 = 0{,}03\%$$

Ora, como a probabilidade que calculamos para que essa mensagem fosse um e-mail normal é maior que a probabilidade que calculamos para que essa mensagem fosse um e-mail indesejado (**3,56%** > **0,03%**), logo esse e-mail será considerado normal e, não, indesejado. Pronto, vocês acabaram de entender o funcionamento prático desse classificador. Agora vamos ver algumas formalizações teóricas...

O Classificador Naive Bayes é um classificador probabilístico baseado no Teorema de Bayes com hipótese forte de independência entre seus atributos/variáveis. *O que é o Teorema de Bayes?* Trata-se de um teorema que descreve a probabilidade condicional de um evento, baseado em um

conhecimento anterior que pode estar relacionado ao evento. A fórmula desse teorema é extremamente simples, sendo basicamente uma aplicação direta de sua definição:

$$P(A|B) = \frac{P(A)\ x\ P(A)}{P(B)}$$

Esse teorema pode ser lido como: probabilidade de ocorrência de um evento A dado que um evento B ocorreu é igual à probabilidade de ocorrência de um evento B dado que um evento A ocorreu multiplicado pela probabilidade de ocorrência de um evento A sobre a probabilidade de ocorrência de um evento B. De forma mais simples, a probabilidade de A sabendo B é igual à probabilidade de B sabendo A, multiplicado pela probabilidade de A e dividido pela probabilidade de B.

Parece complicado, mas não é! Esse teorema permite para calcular uma probabilidade condicional, isto é, probabilidade de ocorrência de um evento dado que outro evento ocorreu. Por exemplo: *qual é a probabilidade de uma pessoa qualquer ser uma mulher dado que essa pessoa tem 1,90m de altura?* Pelo Teorema de Bayes, é a probabilidade de uma pessoa ter 1,90m de altura dado que ela é uma mulher, multiplicado pela probabilidade de ser mulher dividido pela probabilidade de ter 1,90m.

O Teorema de Bayes é a base para estudar o Classificador Naive Bayes. A palavra de origem inglesa *naive* significa ingênuo, ou seja, esse tópico trata do classificador ingênuo de Bayes. *Por que ele é chamado de ingênuo?* Porque pressupõe-se que as variáveis ou atributos contribuem para a probabilidade de forma independente uma da outra. Em outras palavras, trata-se de uma aplicação ingênua do Teorema de Bayes por considerar que as variáveis são independentes.

Na prática, isso significa que esse classificador retorna a mesma probabilidade independente da ordem dos eventos. Logo, a probabilidade para "Querido Amigo" é igual à de "Amigo Querido":

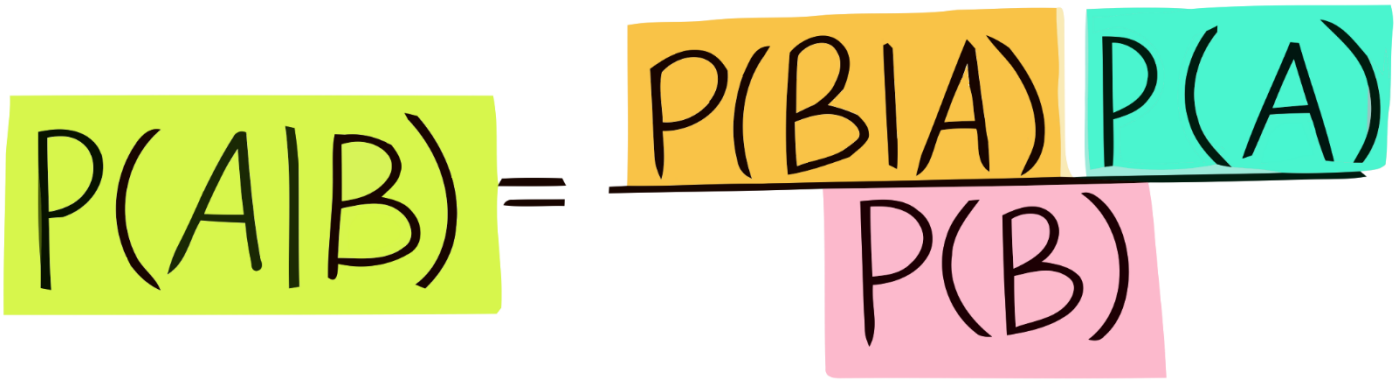
$$P(Normal)\ x\ P(Normal)\ x\ P(Amigo|Normal) = 0{,}89\ x\ 0{,}25\ x\ 0{,}16 = 0{,}0356 = 3{,}56\%$$

$$=$$

$$P(Normal)\ x\ P(Normal)\ x\ P(Querido|Normal) = 0{,}89\ x\ 0{,}16\ x\ 0{,}25 = 0{,}0356 = 3{,}56\%$$

Nós temos regras gramaticais que regulam a forma em que falamos ou escrevemos, mas esse classificador ignora completamente essa característica. Ele considera que a língua escrita é simplesmente um saco cheio de palavras e cada mensagem de e-mail é só um conjunto aleatório dessas palavras, mas – de forma curiosa – esse classificador funciona bem com aprendizado de máquina. Logo, diz-se que ele tem alto viés e baixa variância (veremos em outro tópico).

Note que existe um paralelismo interessante entre esse classificador e o algoritmo de árvores de decisão, visto que ambos respondem perguntas para descobrir a probabilidade de algo pertencer a uma determinada classe (Ex: um e-mail com as palavras "querido" e "amigo" ser um spam). Só que, para o algoritmo de árvore de decisão, a ordem das perguntas importa; no classificador ingênuo, a ordem das perguntas não importa porque as variáveis são consideradas não correlacionáveis.



Por fim, um dos grandes benefícios desse classificador é que ele é muito simples, rápido e escalável, dado que as variáveis são independentes (e isso facilita muito os cálculos!). Em alguns casos, a velocidade é preferível à maior precisão. Além disso, ele funciona bem com a classificação de texto, processamento de linguagem natural, detecção de spam, entre outros. Ele também é capaz de realizar, com precisão, o treinamento de um modelo com uma quantidade reduzida de amostras.

Em outras palavras, ele requer apenas um pequeno número de dados de treinamento para estimar os parâmetros necessários para a classificação. Ele precisa apenas de dados suficientes para entender a relação probabilística entre cada variável independente isoladamente em relação à variável alvo. Se os dados não fossem independentes, precisaríamos de uma amostra maior para entender a relação probabilística entre as combinações de variáveis.

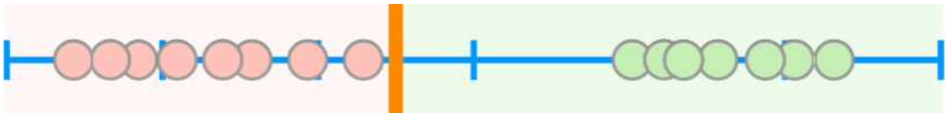
Como desvantagens, podemos destacar que ele assume que as variáveis são independentes, o que raramente ocorre na vida real e há o problema de variáveis com nenhuma ocorrência de frequência. Em suma: o Classificador Naive Bayes se baseia na probabilidade condicional do Teorema de Bayes para encontrar a mais provável das possíveis classificações considerando que as variáveis/atributos não são correlacionadas, isto é, são independentes entre si (ou ingênuos).

Support Vector Machine (SVM)

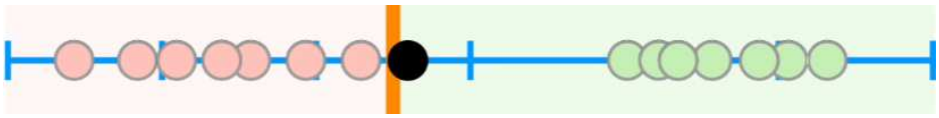
RELEVÂNCIA EM PROVA: BAIXÍSSIMA

SVM (*Support Vector Machine*) é método de mineração de dados de aprendizado supervisionado não probabilístico utilizado tanto para classificação quanto para regressão. Para entender seu funcionamento, vamos partir de um exemplo de peso de ratos de laboratório. Suponha que coletemos o peso de cada rato e coloquemos seus valores em uma linha de tal forma que as bolinhas vermelhas representem ratos não obesos e as bolinhas verdes representem ratos obesos.

Dessa maneira, nós podemos escolher um valor limite de tal forma que, caso uma nova observação possua um peso menor que esse valor limite, ela será classificada como não obesa; e caso possua um peso maior, será classificada como obesa. Esse valor limite é representado na imagem seguinte pela barra laranja (apesar de ser um ponto). Note que qualquer nova observação à esquerda da barra alaranjada será considerada não obesa e qualquer observação à direita será considerada obesa.

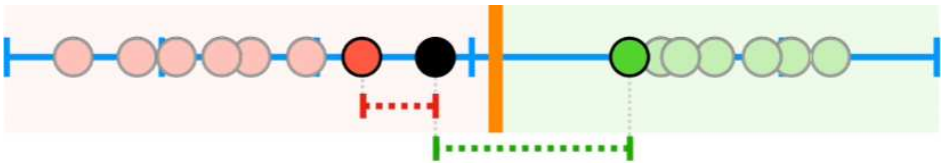


Infelizmente as coisas não são tão simples assim! *E se tivermos uma nova observação que esteja bem próximo do valor limite como a bolinha preta abaixo?*



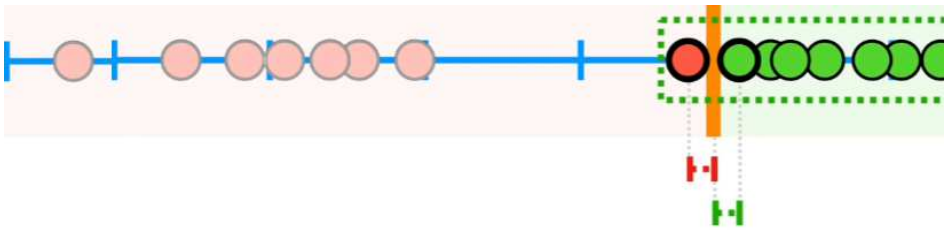
Ora, está à direita do valor limite, logo deveria ser classificada com obesa. No entanto, isso não faz muito sentido porque ela está mais próxima dos não obesos do que dos obesos. Podemos concluir que esse valor limite que escolhemos não foi muito legal! *Vamos escolher outro?* Uma maneira interessante seria focar nas observações que estão localizadas na borda interna de cada classe de dados conforme apresenta a imagem e utilizar o ponto central entre as duas como limite:

Note que agora, se uma nova observação estiver localizada à esquerda do valor limite, ela estará sempre mais próxima das observações classificadas como não obesas do que das observações classificadas como obesas conforme apresenta a imagem seguinte, logo faz todo sentido classificá-la realmente como não obesa. E é claro que isso ocorre também de forma análoga para as observações classificadas como obesas:



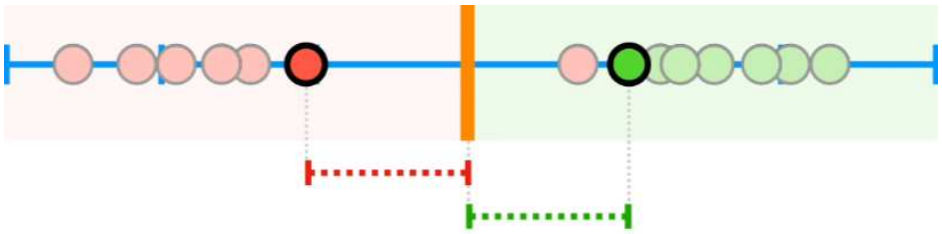
A menor distância entre observações e o valor limite é chamada de margem – isso será importante daqui a pouco. Quando o valor limite está justamente no ponto central entre as duas observações localizadas na borda interna de cada classe de dados, as distâncias são iguais e a margem é a maior possível. Caso movamos o valor limite para um lado ou para outro, como a margem é a menor distância entre as observações e o valor limite, ela sempre será menor.

Vamos complicar um pouquinho mais os nossos dados de treinamento. Imagine agora que tenhamos um valor anômalo (*outlier*) dentro do meu conjunto de dados:

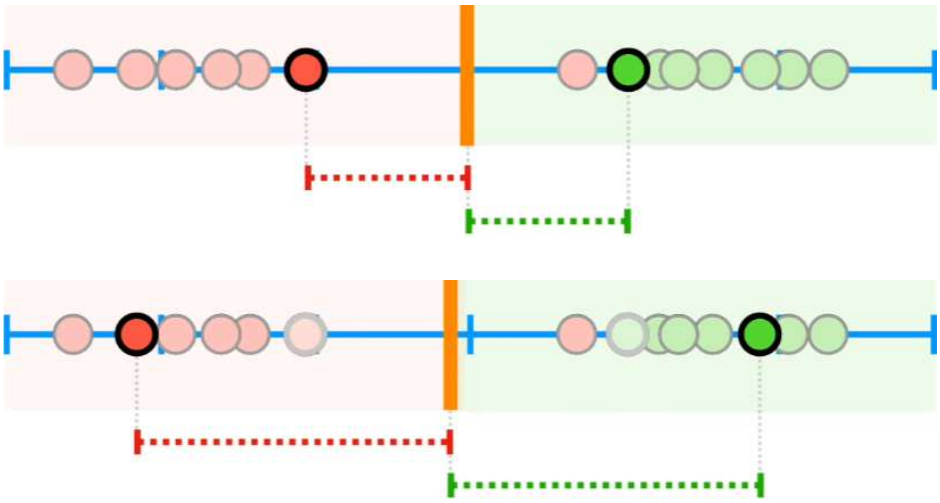


Note que o dado anômalo foi classificado como não obeso, mas está bem mais próximo da classe de obesos. O valor limite ficaria no ponto central entre as duas observações localizadas na borda interna de cada classe de dados conforme apresenta a imagem acima. Se tivermos uma nova observação como a representada pela bolinha preta na imagem seguinte, ela seria classificada como não obesa, muito embora esteja mais próxima da classe de obesos.

Podemos concluir que é complexa a atividade de escolher um limite quando o conjunto de dados de treinamento possui dados anômalos. A solução para esse tipo de problema é permitir classificações incorretas. No exemplo abaixo: se ignorarmos a bolinha vermelha da direita e utilizarmos um valor limite central entre as observações localizadas na borda interna de cada classe, nós vamos errar a classificação dos dados anômalos, mas classificaremos corretamente o restante.

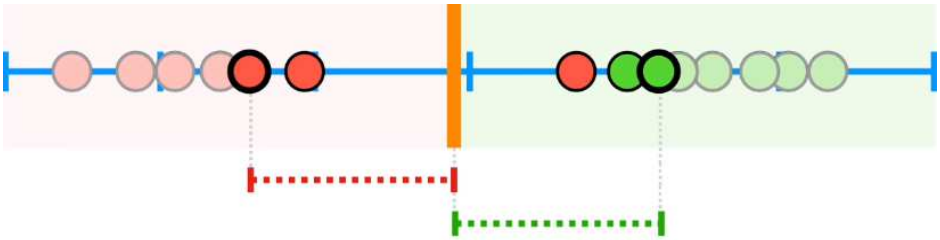


Quando nós permitimos classificações incorretas, a distância entre as observações e o valor limite é chamada de margem flexível (*soft margin*). Agora vejam as duas imagens a seguir:



Pergunto: *qual possui a margem flexível melhor?* Para responder a essa pergunta, utilizamos uma técnica chamada *cross-validation* (veremos adiante). Ela permite determinar quantas classificações incorretas e quantas observações devem ser permitidas dentro da margem flexível para conseguir a melhor classificação. Quando usamos uma margem flexível para determinar a localização ótima de um valor limite, estamos usando um Support Vector Classifier (SVC)[14].

Em tradução livre, seria chamado de Classificador de Vetores de Suporte. *E você sabe o que são os vetores de suporte?* São as observações dentro da margem flexível!



Todos esses exemplos são unidimensionais, mas poderíamos ter duas dimensões (Ex: Peso e Altura), três dimensões (Ex: Peso, Altura e Idade) ou mais dimensões. Quando temos apenas uma dimensão, o classificador de vetores de suporte é um ponto; quando temos duas dimensões, o classificador de vetores de suporte é uma linha; quando temos três dimensões, o classificador de vetores de suporte é um plano[15]; e quando temos mais dimensões, o classificador é um hiperplano.

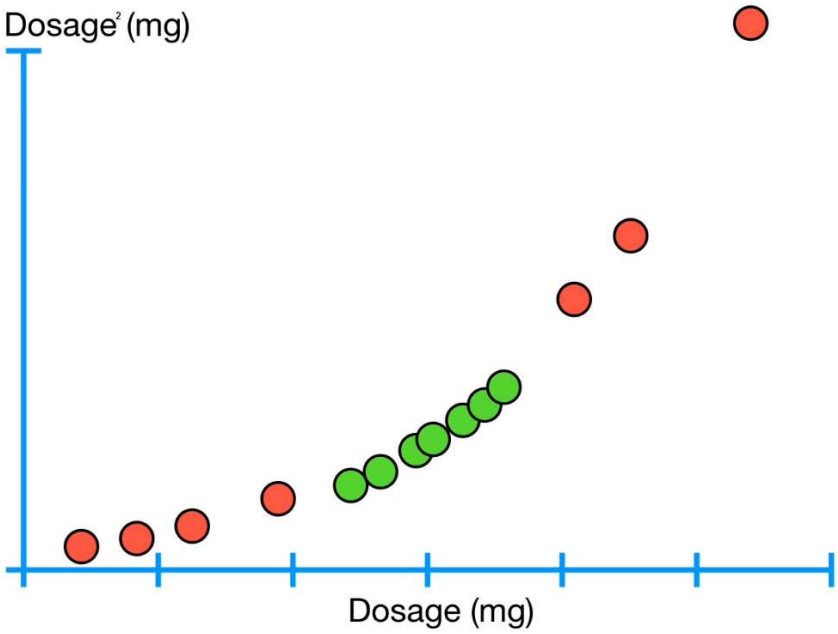
Agora vejam o conjunto de dados apresentado a seguir que representa a dosagem ideal de um remédio em miligramas. Há um ditado popular que diz: "*A diferença entre o remédio e o veneno é a dose*". Isso faz sentido porque uma dose muito baixa de um remédio não faz o efeito esperado e uma dose muito alta pode até causar danos. Os pontos apresentados mostram mais ou menos isso: em vermelho, pacientes que não foram curados e em verde pacientes que foram curados.

[14] Caso estivéssemos usando regressão, seria um Support Vector Regression (SVR).
[15] Pode-se dizer também hiperplano unidimensional, hiperplano bidimensional e hiperplano tridimensional respectivamente.

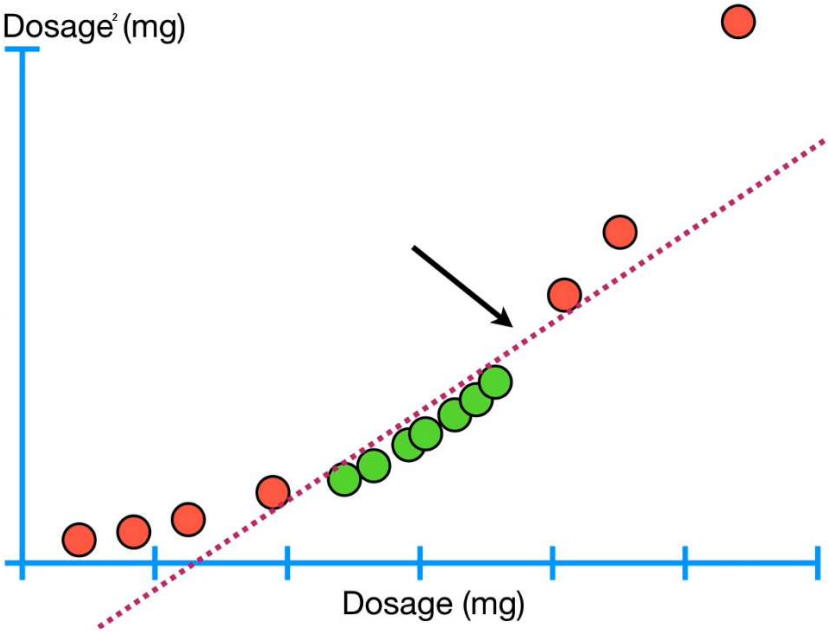
Podemos concluir que a dose não funciona caso seja pequena demais ou grande demais. E agora temos um problema: qualquer que seja a posição do valor limite, nós sempre teremos muitas classificações incorretas. Vocês podem tentar à vontade escolher um ponto na imagem acima que faça uma boa classificação dos dados que vocês não vão conseguir. Pode ser no início, no meio ou no fim, sempre haverá uma grande quantidade de classificações incorretas.

Pois é, os classificadores de vetores de suporte não possuem um bom desempenho para esse tipo de conjunto de dados. *E o que fazer agora, Diego?* Chegou o momento de ele brilhar*: Support Vector Machines* (SVM). Ele resolve o problema de selecionar um valor limite que permita classificar com acurácia um conjunto de dados por meio da adição de dimensões ao problema, isto é, o que era um problema unidimensional agora será um problema bidimensional.

Vejamos um exemplo: vamos construir um plano (bidimensional) em que o eixo X continuará sendo a dosagem do remédio e o eixo Y será a dosagem ao quadrado:



Por que o gráfico ficou desse jeito? Imaginem que a primeira bolinha vermelha representava 0,5mg no eixo X. Como o eixo Y é dosagem², seu valor será 0,5 x 0,5 = 0,25mg. Dessa forma, a bolinha vermelha que ocupava a posição 0,5 em um hiperplano unidimensional, agora ocupará a posição (0,5, 0,25) em um hiperplano bidimensional; e assim por diante para os demais pontos do gráfico. *E por que fazer tudo isso?* Porque agora é possível traçar uma linha que separa os dados:



Viram a mágica? Agora se um novo dado surgir, será possível classificá-lo de forma mais satisfatória. Em síntese, a ideia por trás desse algoritmo é: iniciar com dados em uma dimensão relativamente baixa (Ex: uma dimensão), depois mover os dados para uma dimensão maior (Ex: duas dimensões), e por fim encontrar um classificador de vetores de suporte que consiga separar os dados de dimensão maior em dois grupos.

E por que nós usamos dosagem ao quadrado? Não poderia ser outra dimensão? O SVM utiliza um conceito chamado Funções de Kernel (ou simplesmente Kernel). Essas funções permitem transformar dados de entrada no formato necessário ao encontrar classificadores de vetores de suporte em dimensões maiores, tais como funções lineares, polinomiais ou radiais. *E quais são as vantagens e desvantagens desse algoritmo?*

Como vantagens, podemos afirmar que ele é bastante efetivo quando as classes dos pontos de dados estão bem separadas; é efetivo até em espaços de alta dimensionalidade; é bastante eficiente nos casos em que o número de dimensões (variáveis) é maior que o número de amostras (linhas de uma tabela); é relativamente eficiente em relação ao usuário de memória computacional; e trabalha bem com imagens.

Como desvantagens, podemos afirmar que ele não é ideal para grandes conjuntos de dados; não tem um bom desempenho quando o conjunto de dados possui muito ruído ou quando as classes de dados não são bem separadas; é desafiador escolher uma função de kernel ótima; em grandes bases de dados, ele precisa de mais treinamento; como não se trata de um modelo probabilístico, não é possível explicar a classificação por esses termos; e é mais difícil de interpretar que outros métodos.

Por fim, é importante saber que o SVM é considerado um método não paramétrico[16], dado que não existe uma função de mapeamento entre dados de entrada e dados de saída definida por um conjunto finito de números (chamados de parâmetros) que plenamente caracterizam todos os possíveis resultados da função de mapeamento e cujos valores específicos são determinados durante o treinamento. *E as funções de kernel?*

As funções de kernel são escolhidas pelo usuário e não otimizadas pelo processo de treinamento, logo elas são consideradas hiperparâmetros (não é necessário entender isso agora). Além disso, podemos afirmar que o SVM é sensível à escala dimensional dos conjuntos de dados. *Como assim, Diego?* Isso significa que mudanças na escala do conjunto de dados afetam as predições de um modelo de aprendizado de máquina.

Veja que estamos falando de mudanças na escala do conjunto de dados e, não, de mudanças nas dimensões do conjunto de dados. Como dados podem vir em diferentes tamanho e formatos, é importante realizar o pré-processamento (transformação, padronização ou normalização) para que eles fiquem com dimensões razoáveis de forma que o algoritmo possa definir bem a fronteira dos dados por meio dos vetores de suporte (Ex: escalar cada atributo entre [0,1] ou [-1,1]).

(SEFAZ/PI – 2021) Considerando os métodos de Mineração de Dados, analise a seguinte descrição: "constrói os denominados classificadores lineares, que separam o conjunto de dados por meio de um hiperplano, sendo considerado um dos mais efetivos para a tarefa de classificação." Trata-se de:

a) Wang-Mendel.
b) Backpropagation.
c) SVM (Support Vector Machines).
d) Classificador Bayesiano Ingênuo.

_____

Comentários: o método que permite construir classificadores que separam conjuntos de dados por meio de um hiperplano sendo útil para a tarefa de classificação é o SVM (*Support Vector Machine*) (Letra C).

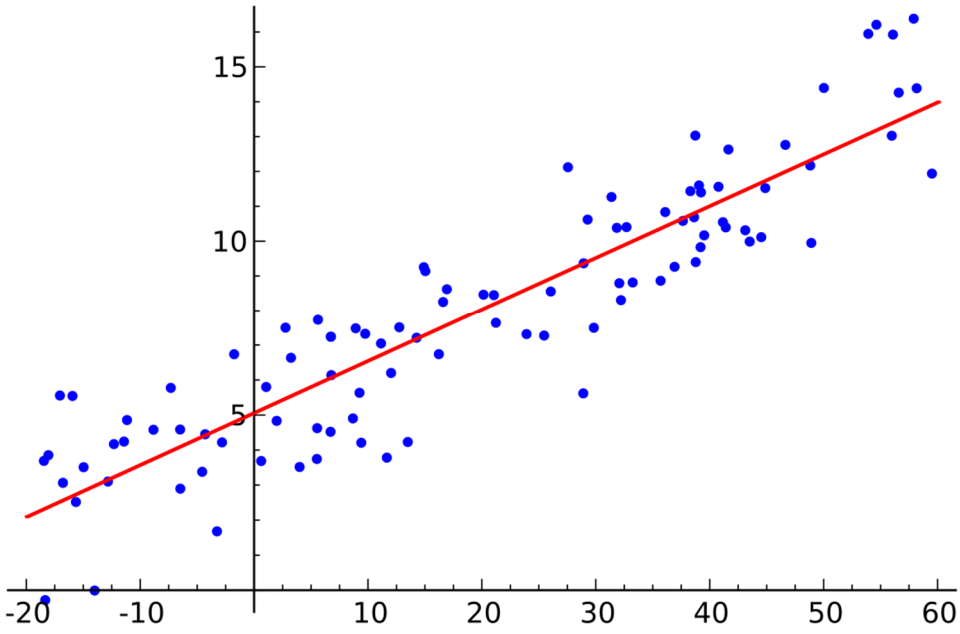
[16] Com uma exceção: quando estamos tratando de SVM Linear. Nesse caso, temos uma função de mapeamento linear.

Regressão

INCIDÊNCIA EM PROVA: BAIXA

Galera, um parente muito próximo da classificação é a regressão. **Na regressão, em vez de prever uma categoria, o objetivo é prever um número**. Vamos pegar o exemplo da Target novamente! Eles queriam saber não apenas se cada cliente estava grávida, mas quando enviar cada cupom de desconto. Então eles conseguiram estimar as datas de nascimento também dos bebês. Essa é uma questão de regressão, isto é, quantas semanas até a cliente dar à luz.

A Regressão depende muitas vezes de dezenas ou mesmo milhares de variáveis ou características que descrevam cada exemplo e encontra uma equação ou curva para ajustar os pontos de dados. Como na classificação, muitas técnicas de regressão dão a cada característica um peso, então combinam contribuições positivas e negativas dos recursos ponderados para obter uma estimativa. Vejam no exemplo a seguir:



Notem que baseado em diversos pontos de dados, foi traçada uma linha capaz de estimar dados – poderia ter sido não-linear também. Galera, essa linha do gráfico é dada por uma equação baseada em variáveis. Logo, se eu tenho a equação, eu posso estimar qualquer outro valor – basta jogar na equação e esperar o resultado. Vamos supor que o exemplo acima seja dado pela equação $y = 2{,}2923x - 46{,}244$ – se eu quiser saber qualquer valor de y, é só mudar o valor de x.

E, assim como a classificação, a regressão também é usada em vários lugares. Um dos exemplos mais conhecidos é o Google Trend da Gripe. Em 2008, o Google começou a publicar estimativas em tempo real de quantas pessoas teriam gripe com base em pesquisas por palavras como "febre" e "tosse". **Em alguns casos, ele foi capaz de prever surtos regionais de gripe até 10 dias antes de serem notificados pelo CDC (Centros de Controle e Prevenção de Doenças).**

Em 2010, o CDC identificou um pico de casos de gripe na região do Atlântico dos Estados Unidos. No entanto, os dados das consultas de pesquisa do Google sobre os sintomas da gripe conseguiram mostrar esse mesmo pico duas semanas antes do relatório do CDC! Inicialmente, o Google tinha uma precisão de 97% em relação ao CDC, porém em anos subsequentes ele reduziu sua precisão e o Google decidiu retirar do ar enquanto não houvesse uma precisão melhor.

HTTPS://WWW.GOOGLE.ORG/FLUTRENDS

Em uma definição formal, Navathe afirma que a regressão é uma aplicação especial da regra de classificação. Se uma regra de classificação é considerada uma função sobre variáveis que as mapeia em uma classe destino, a regra é chamada regressão. **Isto ocorre quando, ao invés de mapear um registro de dados para uma classe específica, o valor da variável é previsto (calculado) baseado em outros atributos do próprio registro**. *Bacana?*

(CRF/SP – 2018) "A etapa de Mineração de Dados compreende a busca efetiva por conhecimentos úteis no contexto da aplicação de KDD (Knowledge Discovery in Database), ou Descoberta do Conhecimento em Bases de Dados. É a principal etapa do processo de KDD." Acerca de algumas das tarefas do KDD, analise a assertiva a seguir: "compreende a busca por uma função que mapeie os registros de um banco de dados em um intervalo de valores reais". Assinale a alternativa que apresenta esta tarefa.

a) Regressão.
b) Classificação.
c) Sumarização.
d) Agrupamento.

_____

Comentários: a tarefa que busca uma função que mapeie registros de um banco de dados é a Regressão (Letra A).

(TCE/SP – 2009) Uma das abordagens de mining define que, se uma regra de classificação é considerada uma função sobre variáveis que as mapeia em uma classe destino, a regra é chamada:

a) categorização.
b) Apriori.
c) algoritmo genético.
d) regressão.
e) minimização.

_____

Comentários: regressão é uma aplicação especial da regra de classificação. Se uma regra de classificação é considerada uma função sobre variáveis que as mapeia em uma classe destino, a regra é chamada regressão. Uma aplicação de regressão ocorre quando, em vez de mapear uma tupla de dados de uma relação para uma classe específica, o valor da variável é previsto baseado naquela tupla (Letra D).

Regressão Linear

INCIDÊNCIA EM PROVA: BAIXÍSSIMA

A regressão linear é um tipo de algoritmo de aprendizado de máquina supervisionado utilizado na mineração de dados. Ela é usada para prever uma variável de destino contínua ajustando uma equação linear aos pontos de dados. Baseia-se na relação entre as variáveis independentes (preditoras) e a variável dependente (alvo). O algoritmo de regressão linear encontra a melhor linha de ajuste que minimiza a soma dos erros quadrados.

Essa linha de melhor ajuste é então usada para prever valores para a variável dependente. Dito de outra forma, a ideia é modelar o relacionamento entre uma variável dependente (Y) e uma ou mais variáveis independentes (Y). Esse modelo é utilizado para fazer previsões sobre a variável dependente ajustando uma equação linear aos dados observados. A equação de regressão linear mostra a relação linear entre a(s) variável(is) independente(s) e a variável dependente.

Sendo assim, o modelo de regressão para uma única variável preditora x, ou regressão linear simples, pode ser definido pela seguinte equação da reta:

$$y = a + bx$$

em que a e b são coeficientes de regressão (pesos) e especificam o intercepto do eixo y e a inclinação da reta, respectivamente.

Deve-se, então, encontrar valores para os coeficientes de regressão, de forma que a reta (ou o plano/hiperplano, se for uma regressão linear multivariada) se ajuste aos valores assumidos pelas variáveis nos exemplares de um conjunto de dados. O melhor ajuste da reta pode ser encontrado, por exemplo, pelo método dos mínimos quadrados, o qual minimiza o erro entre os valores das variáveis nos exemplares do conjunto de dados e os valores estimados pelo regressor.

Regressão Logística

INCIDÊNCIA EM PROVA: BAIXÍSSIMA

Na mineração de dados, a regressão logística é uma técnica de modelagem preditiva utilizada para problemas de classificação. Trata-se de um algoritmo de aprendizado supervisionado que usa um conjunto de dados de treinamento rotulados para construir um modelo que pode prever com precisão os resultados de dados não vistos. A regressão logística é usada para identificar padrões em grandes conjuntos de dados e para estimar a probabilidade de um determinado evento ocorrer.

Análise de Agrupamentos

INCIDÊNCIA EM PROVA: MÉDIA

Análise de agrupamentos (também chamados de clusters, grupos, aglomerados, segmentos, partições ou agregações) é uma técnica que visa fazer agrupamentos automáticos de dados segundo o seu grau de semelhança, permitindo a descoberta por faixa de valores e pelo exame de atributos das entidades envolvidas. **Como o nome sugere, o objetivo é descobrir diferentes clusters em uma massa de dados e agrupá-los de uma forma que ajude com sua análise**.

Um agrupamento é uma coleção de registros similares entre si, porém diferentes dos outros registros nos demais agrupamentos. Esta tarefa difere da classificação, uma vez não necessita que os registros sejam previamente categorizados – trata-se de um aprendizado não-supervisionado. Além disso, ela não tem a pretensão de classificar, estimar ou predizer o valor de uma variável, ela apenas identifica os grupos de dados similares.

Vamos imaginar um site: www.mercadolivre.com.br! Lá existem milhões de produtos de absolutamente tudo que você imaginar – mesmo dentro de uma única categoria, existem uma infinidade de produtos diferentes. **Por essa razão, o Mercado Livre organiza as coisas em subcategorias, mas é inviável ter um funcionário que fique organizando todos os anúncios em categorias, subcategorias, entre outros.**

Em vez disso, a empresa pode usar técnicas de agrupamento para agrupar automaticamente os produtos. Mais uma vez: cada produto deve primeiro ser dividido em características numéricas, como quantas vezes a palavra "impressora" aparece na descrição ou quem é a fabricante. **O método de cluster mais simples é adivinhar quantas subcategorias distintas devem existir**. Sim, você dá um "chute" de categorias...

Em seguida, você agrupa itens aleatoriamente em várias categorias diferentes e depois continua mudando itens entre categorias para tornar cada o grupo mais preciso. No final, acreditem ou não, produtos similares acabam se agrupando, mas não precisamos parar por aí! Imaginem dois anúncios de um mesmo modelo de câmera, mas com cores diferentes! Eles não precisam ficar em categorias separadas, porque são apenas variantes do mesmo produto.

Dessa forma, além das subcategorias, seria interessante mesclar alguns grupos. Sites como o Mercado Livre fazem isso por meio de uma técnica chamada clustering hierárquico. **Em vez de um único conjunto de categorias, o agrupamento hierárquico produz uma espécie de árvore taxonômica, aglomerando ou dividindo elementos**. *Adivinhem quem utilizou essa técnica?* Cambridge Analytica. Ela agrupou eleitores que respondiam ao mesmo tipo de publicidade!

Em síntese, essa técnica realiza agrupamentos comuns chamados *clusters*, de modo que o grau de associação seja forte entre os membros do mesmo grupo e fraco entre membros de grupos diferentes. *Bacana?* Antes de falarmos sobre os principais algoritmos de agrupamento nos tópicos seguintes, vamos primeiro ver algumas classificações. Iniciemos pela classificação em determinístico e estocástico:

CLASSIFICAÇÃO	DESCRIÇÃO
DETERMINÍSTICO	Métodos determinísticos apresentam sempre o mesmo agrupamento, independe de parâmetros do algoritmo e/ou da condição inicial.
ESTOCÁSTICO	Métodos estocásticos podem apresentar diferentes soluções dependendo dos parâmetros e/ou da condição inicial.

Os algoritmos de agrupamento também podem ser divididos em algoritmos hierárquicos ou algoritmos particionais:

CLASSIFICAÇÃO	DESCRIÇÃO
HIERÁRQUICO	Métodos hierárquicos criam uma decomposição hierárquica dos dados, podendo ser divididos em aglomerativos ou divisivos, baseados em como o processo de composição/decomposição é efetuado. Métodos aglomerativos começam com cada objeto pertencendo a um grupo e unem sucessivamente objetos. Métodos divisivos começam com todos os objetos fazendo parte do mesmo grupo e particionam sucessivamente os grupos em grupos menores.
PARTICIONAL	Dado um conjunto com n objetos, um método particional constrói k partições dos dados, sendo que cada partição representa um cluster (k <= n). Dado o número k de partições, um método particional cria uma partição inicial e emprega um algoritmo de realocação iterativa que tem por objetivo melhorar o particionamento movendo objetos entre grupos (Ex: K-Means e K-Medoids).

Por fim, os algoritmos particionais podem ser divididos em algoritmos rígidos (*hard*), algoritmos suaves (*soft*) ou algoritmos difusos (*fuzzy*).

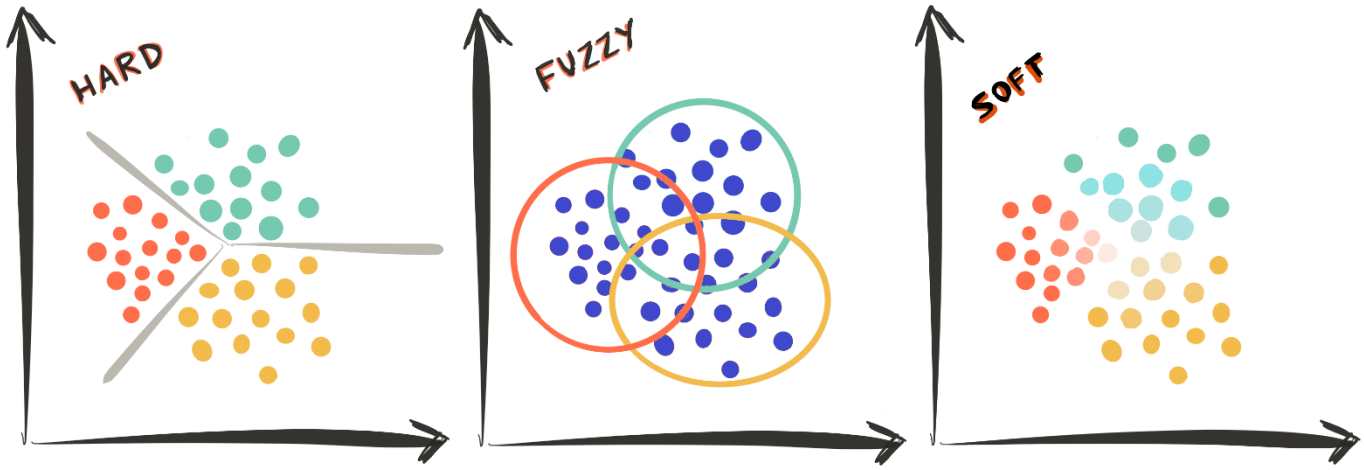
CLASSIFICAÇÃO	DESCRIÇÃO
RÍGIDOS (HARD)	Em agrupamentos rígidos, cada objeto pertence a um único grupo.
SUAVES (SOFT)	Em agrupamentos suaves, cada objeto pertence completamente a mais de um grupo.
DIFUSOS (FUZZY)	Em agrupamentos difusos, cada objeto pertence parcialmente a mais de um grupo.

Vamos supor que em nosso *dataset*, queremos agrupar um conjunto de mangas. Ao utilizar um método de agrupamento rígido, um possível agrupamento poderia ser: manga verde, manga

amarela e manga vermelha. Ao utilizar um método de agrupamento suave, poderíamos ter esses mesmos rótulos, mas baseado no grau de pertencimento (calculado por meio de probabilidade) aos grupos (Ex: Manga 1: 45% Verde, 40% Vermelha, 15% Amarela, logo será considerada Verde).

E ao utilizar um método de agrupamento difuso, poderíamos ter esses mesmos rótulos, mas uma manga pode ser considerada simultaneamente (Verde e Vermelha) ou (Verde e Amarela) ou (Amarela e Vermelha) ou (Verde, Vermelha e Amarela) de acordo com seus graus de pertencimento. *Então qual é a diferença para o anterior, professor?* No suave, apesar dos graus de pertencimento, cada objeto tem um único rótulo; no difuso, o objeto pode ter mais de um rótulo.

Nas três imagens seguintes, eu tentei representar esses três cenários em um plano cartesiano (eu fiz esse desenho no iPad, então não ficou muito bom, mas é possível de entender).



(MPU – 2013) Em se tratando de mineração de dados, a técnica de agrupamento (clustering) permite a descoberta de dados por faixa de valores, por meio do exame de alguns atributos das entidades envolvidas.

_____

Comentários: a técnica de agrupamento realmente permite a descoberta de dados por faixa de valores, por meio do exame de alguns atributos das entidades envolvidas (Correto).

(SERPRO – 2013) Em algoritmos de clusterização hierárquica, os clusters são formados gradativamente por meio de aglomerações ou divisões de elementos, gerando uma hierarquia de clusters.

_____

Comentários: o agrupamento hierárquico ou *clustering* hierárquico tem como característica um processo de junções (aglomerações) ou separações de grupos ou elementos. Ele visa detectar a existência de diferentes grupos dentro de um determinado conjunto de dados e, em caso de existência, determina dividi-los ou aglomerá-los. O Clustering tenta identificar um conjunto finito de categorias ou clusters para os quais cada registro ou elemento possa ser mapeado (Correto).

Ainda no contexto de análise de agrupamentos, é importante falar sobre o **Coeficiente de Silhouette**. Trata-se de uma utilizada para calcular a qualidade dos agrupamentos gerados por um

algoritmo de clusterização. Este coeficiente oferece uma forma de avaliar o quão bem cada objeto foi agrupado, se está no cluster correto (ou agrupamento) e o quão separados os clusters estão entre si.

O valor do coeficiente de silhouette varia de -1 a 1, onde: (1) um valor próximo a +1 indica que o objeto está bem agrupado dentro de seu cluster e longe dos outros clusters; (2) um valor de 0 indica que o objeto está na fronteira ou muito próximo entre dois clusters; (3) um valor próximo a -1 indica que o objeto foi colocado no cluster errado. Para calcular o coeficiente de silhouette de um objeto, você precisa de:

- **Coerência (a):** trata-se da média da distância entre o objeto e todos os outros objetos no mesmo cluster;
- **Separação (b):** trata-se da menor média da distância entre o objeto e todos os objetos em outro cluster (o mais próximo);

O coeficiente de silhouette é particularmente útil para determinar o número ótimo de clusters em um conjunto de dados. Ao calcular o coeficiente médio de silhouette para todos os objetos em diferentes números de clusters, o número de clusters que produz o maior coeficiente médio de silhouette é considerado o número ótimo de clusters para aqueles dados, indicando uma boa separação e coerência dos clusters formados.

(FGV / Câmara dos Deputados – 2023) O Coeficiente Silhouette é utilizado na análise de agrupamentos, principalmente para examinar:

a) a separação e a coesão dos agrupamentos.
b) a preservação de pequenos agrupamentos.
c) a completude e a interseção dos agrupamentos.
d) a heterogeneidade dos agrupamentos.
e) a forma convexa dos agrupamentos.

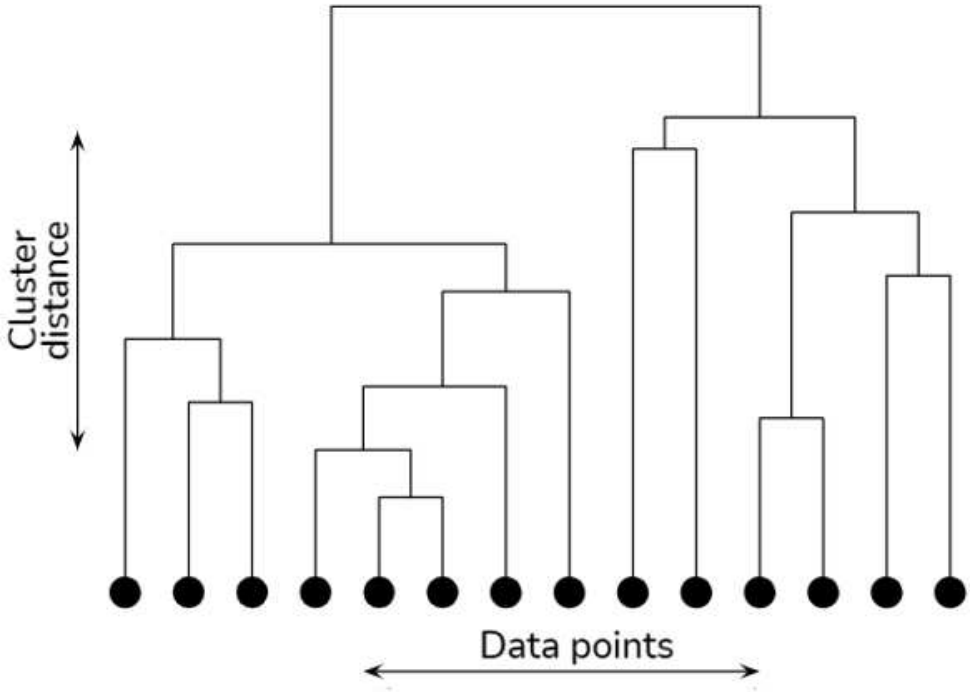
_____

Comentários: o coeficiente silhouette é uma métrica usada para calcular e interpretar a qualidade dos agrupamentos gerados por algoritmos de clusterização. Ele mede quão bem um ponto foi agrupado, considerando **tanto a coesão dentro dos clusters** (quão próximos os pontos dentro de um mesmo cluster estão entre si) **quanto a separação entre diferentes clusters** (quão distantes os clusters estão uns dos outros). Valores mais altos do coeficiente indicam um agrupamento mais adequado, onde os pontos dentro de um cluster estão próximos entre si, e os clusters estão bem separados uns dos outros (Letra A).

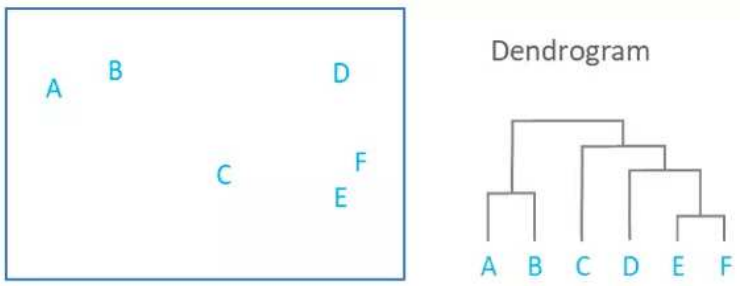
Agrupamento Hierárquico

INCIDÊNCIA EM PROVA: BAIXÍSSIMA

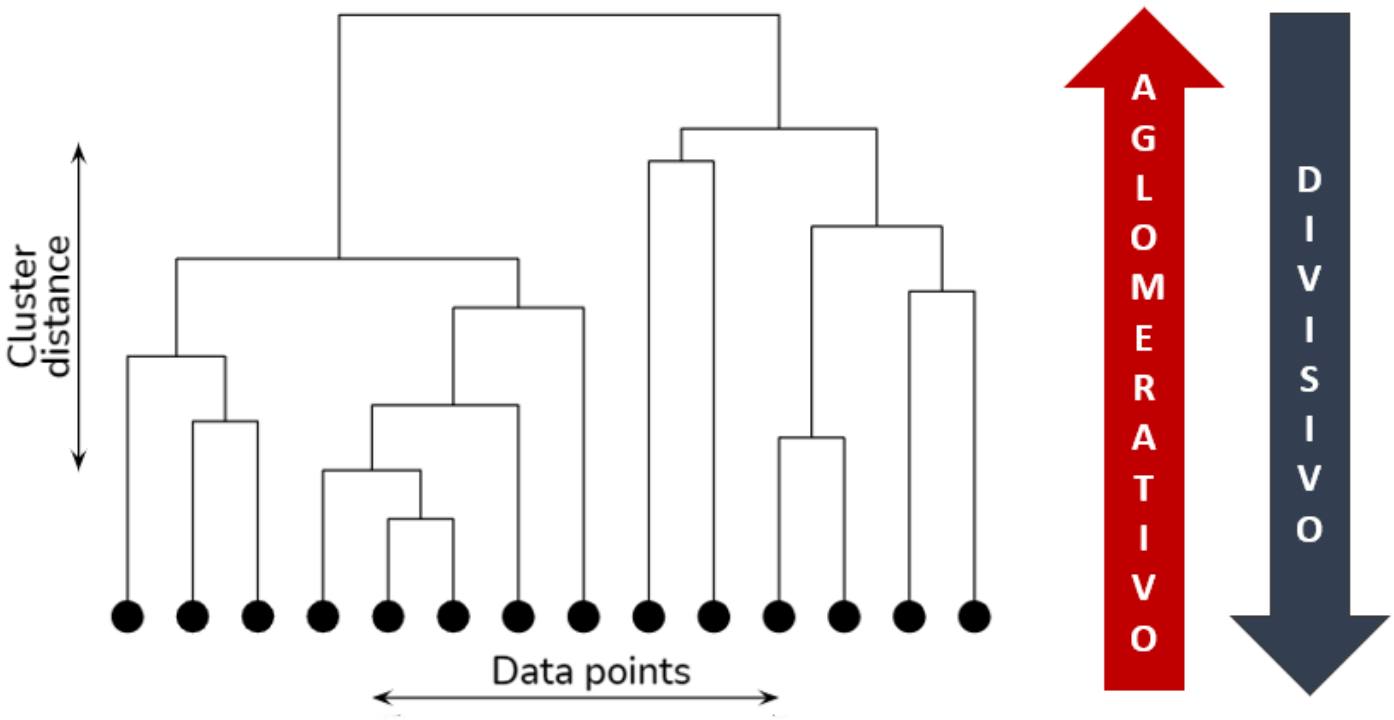
O conceito de agrupamento hierárquico se baseia na construção e análise de um dendrograma. *O que seria isso, professor?* **É basicamente um diagrama que exibe a relação hierárquica entre objetos.** No dendrograma apresentado a seguir, a parte de cima representa as gerações mais antigas (avós) e a parte de baixo representa as gerações mais novas (filhos), contendo no meio gerações intermediárias.



Para interpretar um dendrograma, é preciso entender que no eixo das abscissas (x) temos os pontos de dados e no eixo das ordenadas (y) temos a distância entre os grupos. No exemplo a seguir, isso fica mais claro: os pontos mais próximos são E e F, logo notem que eles possuem uma altura menor; em seguida, os pontos mais próximos são A e B, logo notem que ele tem a altura um pouco maior; depois o ponto mais próximo é entre o grupo (E,F) e D; depois (E,F,D) e C; depois (E,F,D,C) e (A,B).



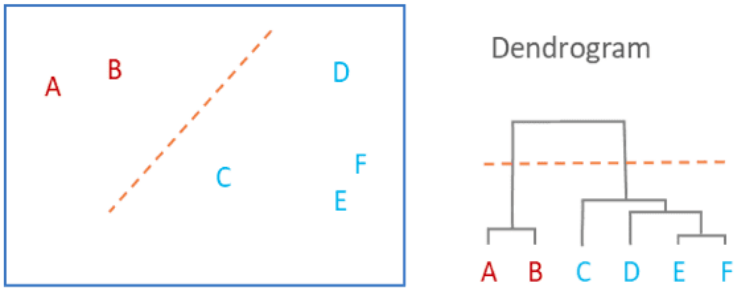
Vocês notaram que nós construímos o dendrograma de baixo para cima (bottom-up)? Pois é, utilizamos um método aglomerativo! Se fosse de cima para baixo (*top-down*), seria divisivo.



O **Método Aglomerativo** é também chamado de Método AGNES (*Agglomerative Nesting*). Conforme vimos, inicialmente cada objeto é considerado um grupo de um único elemento (folha) e, a cada passo do algoritmo, dois grupos similares são combinados para formar um grupo maior (nós). Esse procedimento ocorre iterativamente de baixo para cima até que todos os pontos sejam membros de um único grupo (raiz).

O **Método Divisivo** é também chamado de Método DIANA (*Divisive Analysis*). Conforme vimos, trata-se do inverso do método anterior: ele começa inicialmente com a raiz, na qual todos os objetos são incluídos em um único grupo. A cada passo da iteração, os grupos mais heterogêneos são divididos em dois. O processo permanece iterando até que todos os objetos sejam seu próprio cluster de um único objeto.

Após terminar a construção do dendrograma, podemos escolher um ponto da árvore para realizar a partição dos grupos conforme é exibido na imagem seguinte. De acordo com Matt Harrisson [Machine Learning – Guia de Referência Rápida: Trabalhando com dados estruturados em Python]: "*Após terminar a construção, tem-se um dendrograma, isto é, uma árvore que controla quando os clusters foram criados e qual é a métrica das distâncias*".



Essa definição já foi cobrada em prova em sua literalidade, mas está errada! Dendrograma é apenas uma representação visual do resultado do processo de agrupamento e não controla coisa alguma.

Caso de Uso de Dendrogramas

Uma empresa coleta dados sobre seus clientes, incluindo informações demográficas (idade, gênero, localização), comportamento de compra (frequência de compra, categorias de produtos adquiridos) e preferências (pesquisas de satisfação, feedback sobre produtos). O objetivo é agrupar clientes com características semelhantes para identificar segmentos de mercado distintos.

- Utiliza-se um algoritmo de agrupamento hierárquico para analisar as semelhanças entre os clientes com base em suas características. O algoritmo começa tratando cada cliente como seu próprio cluster e, em seguida, progressivamente mescla clusters com base em sua semelhança.
- O processo de fusão dos clusters é representado visualmente por meio de um dendrograma. Cada ramificação do dendrograma representa a fusão de dois ou mais clusters, e a altura das ramificações indica a distância ou diferença entre os clusters mesclados.
- Analisando o dendrograma, a empresa pode determinar o número ótimo de segmentos de mercado pela observação dos pontos naturais de corte no dendrograma, onde a fusão de clusters resulta em um aumento significativo na distância.

Com base no dendrograma, a empresa identifica, por exemplo, quatro segmentos principais de clientes. Cada segmento reflete um grupo de clientes com necessidades, comportamentos e preferências similares. Isso permite à empresa adaptar suas estratégias de marketing, desenvolver produtos específicos para cada segmento e melhorar a comunicação com diferentes grupos de clientes, maximizando a satisfação do cliente e a eficiência das vendas.

(SEFA/PA – 2022) A estratégia de agrupamento hierárquico em que a construção da árvore é iniciada pelo nó raiz, onde todos os exemplares são alocados, inicialmente, a um único grupo e, interativamente, os grupos são divididos de acordo com algum critério de dissimilaridade, aplicado aos exemplares que os constituem e, além disso, enquanto houver grupos formados por mais de um exemplar, dois grupos distintos são criados a cada divisão, dando origem aos demais nós internos da árvore, é conhecida como:

a) método AGNES.
b) método DIANA.
c) método de k-médias.
d) método DBSCAN.
e) mapa auto organizáveis.

_____

Comentários: agrupamento hierárquico em que a construção da árvore é iniciada pelo nó raiz é o Método DIANA (Letra B).

(SERPRO – 2021) Nos agrupamentos hierárquicos, um dendrograma é uma árvore que controla quando os clusters são criados e que determina qual é a métrica das distâncias.

_____

Comentários: a questão ter sido retirada em sua literalidade de um livro não a torna correta, então eu vou continuar discordando do gabarito definitivo da questão (Correto).

(FUNPAR / UNILA – 2014) No uso do método de agrupamento hierárquico, a formação dos agrupamentos é feita usando ligações e os resultados obtidos são dispostos a partir de um gráfico chamado:

a) Esquema de Mahalanobis.
b) Ligação completa.
c) Dendrograma.
d) Esquema de similaridades.
e) Função discriminante de Fisher.

_____

Comentários:

(a) Errado. O Esquema de Mahalanobis não é um gráfico usado para mostrar a formação de agrupamentos, mas sim uma medida de distância que leva em conta a correlação entre as variáveis;

(b) Errado. Ligação Completa não é um tipo de gráfico, mas sim uma abordagem específica para medir distâncias durante o processo de agrupamento;

(c) Correto. O dendrograma é o gráfico utilizado para ilustrar a formação de agrupamentos no método de agrupamento hierárquico. Ele mostra as relações de proximidade entre os objetos ou pontos de dados e como os grupos são formados passo a passo, desde os menores agrupamentos (compostos por objetos individuais) até o agrupamento total que engloba todos os objetos analisados;

(d) Errado. A análise de similaridade é um conceito mais amplo que pode ser visualizado de várias maneiras, mas não é o nome de um gráfico específico como o dendrograma;

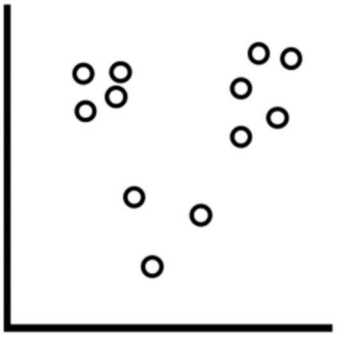
(e) Errado. O objetivo da Função Discriminante de Fisher é maximizar a separação entre as categorias, o que é diferente do propósito de agrupamento hierárquico, que visa agrupar dados com base em sua similaridade (Letra C).

K-Médias

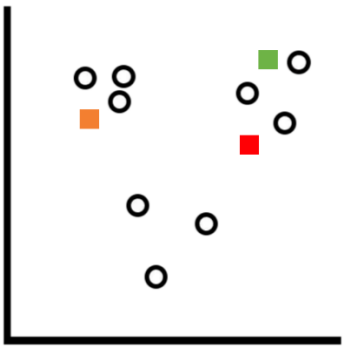
INCIDÊNCIA EM PROVA: BAIXÍSSIMA

Também chamado de K-Means, trata-se de algoritmo de agrupamento que basicamente agrupa dados em k grupos, em que k é um valor arbitrário definido pelo usuário. Esse algoritmo busca minimizar a soma de todos os quadrados das distâncias entre os pontos de dados e um ponto chamado centroide (também conhecido como semente). Galera, é um pouco complicado explicar de forma satisfatória esse algoritmo em texto, mas eu fiz uns desenhos no Paint para ajudar 😊

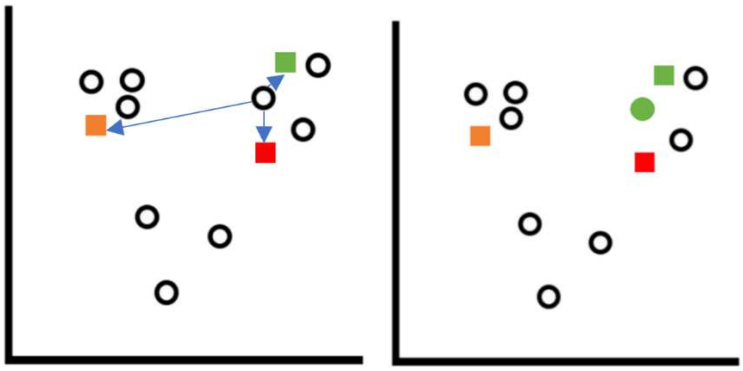
O primeiro passo desse algoritmo é reunir os dados e determinar em quantos grupos nós queremos dividi-los. Em nosso exemplo, eu vou querer dividir em três grupos (k = 3).



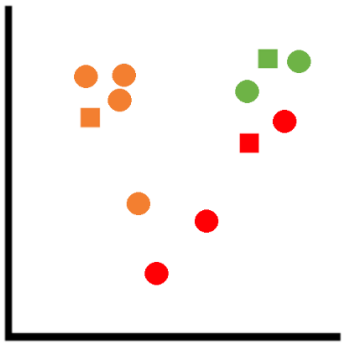
No segundo passo, nós selecionamos k pontos aleatórios. Como nós decidimos no passo anterior que k = 3, então vamos selecionar três pontos no *dataset* (esses pontos são os centroides).



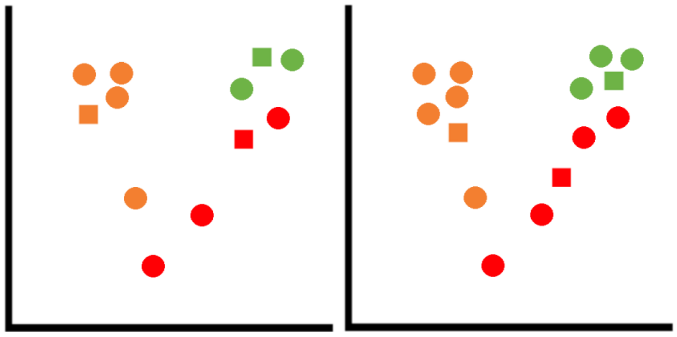
No terceiro passo, nós vamos calcular a distância de cada um dos pontos para cada um dos centroides. Por exemplo: começando pelo ponto lá no canto superior direito conforme mostra a imagem seguinte. Note que eu inseri três setas azuis saindo desse ponto em direção aos três centroides. Ao calcular a distância, é possível ver que o ponto verde é o mais próximo, logo esse primeiro ponto analisado será marcado como verde.



Agora faltam 8 pontos! Seguimos para o próximo e repetimos esse procedimento para os pontos restantes até chegar ao resultado apresentado a seguir:



No quarto passo, nós temos que identificar novos centroides. Para tal, nós vamos migrá-los de suas posições iniciais para a posição mais próxima dos centros dos pontos da sua respectiva cor. *Diego, segura a onda aí que foi rápido demais!* Vamos lá: para cada ponto de um determinado grupo (laranja, vermelho e verde), nós vamos somar o valor de suas coordenadas e dividir pela quantidade de pontos (por isso é uma média). Essa coordenada resultante será a nova posição do centroide 😊



O quinto passo trata basicamente de repetir o processo até o momento em que os centroides não alteram mais suas posições (ou a alteração é minúscula) e as observações não mudam mais de grupo, alcançando a posição ótima[17]. Lembrando que esse algoritmo gera agrupamentos em função da distância euclidiana (soma do quadrado das diferenças das coordenadas) de um ponto central. É importante dizer que o algoritmo k-means não é determinístico. *Como assim, Diego?*

[17] Apesar disso, não há garantia de que um ótimo global será atingido, ou seja, de que o melhor particionamento possível dos dados será encontrado.

Isso significa que a escolha aleatória inicial dos centroides influencia no agrupamento gerado, logo – a cada execução do algoritmo com diferentes escolhas iniciais de centroide – podemos obter resultados distintos. O k-Means é um algoritmo rápido, eficiente, fácil de entender e implementar, no entanto a escolha do valor ótimo de k é desafiadora, a escolha do centroide inicial influencia no resultado final e ele é bastante sensível a anomalias e ruídos.

(TCE/PE – 2017) O método de clustering k-means objetiva particionar 'n' observações entre 'k' grupos; cada observação pertence ao grupo mais próximo da média.

Comentários: o método k-means realmente particiona n observações em k grupos, sendo que cada observação pertence ao grupo mais próximo da média. Em outras palavras, ele segrega n dados em torno de k centroides, em que cada dado está associado ao grupo mais próximo da média (Correto).

K-Medoides

INCIDÊNCIA EM PROVA: BAIXÍSSIMA

Em primeiro lugar: *o que é um medoide?* Medoide é o objeto com menor dissimilaridade média em relação a todos os outros objetos, isto é, aquele mais centralmente localizado de um grupo. Nós acabamos de ver que os centroides não são pontos reais, mas simplesmente a média dos pontos presentes em um agrupamento. A ideia do k-Medoides é fazer com que os centroides finais sejam pontos de dados reais.

Em outras palavras, o k-Medoides escolhe objetos da própria base como protótipo[18] em lugar de um centroide, ao passo que o k-Médias calcula o centro do grupo a partir dos objetos contidos neles. Trata-se de um algoritmo mais resistente a ruídos e valores anômalos que o k-Médias, dado que o centro de um agrupamento será um objeto da própria base. Dessa forma, ruídos e valores anômalos influencia menos a definição do centro.

Fuzzy K-Médias

INCIDÊNCIA EM PROVA: BAIXÍSSIMA

O Fuzzy K-Médias é basicamente uma extensão do k-Médias em que cada objeto pode pertencer a mais de um grupo. Ele se baseia na ideia de que cada objeto possui um grau de pertencimento em relação a cada um dos agrupamentos. O grau de pertencimento é um valor normalizado entre 0 e 1, mas não representam probabilidades, logo a soma dos valores pode ser maior que 1. De resto, esse algoritmo é bastante similar ao K-Médias.

Ele tem alguns problemas: está sujeito a uma localização ótima do centroide, o resultado também depende da condição inicial e o valor de k deve ser definido a priori.

[18] No caso do k-médias, designa-se centroide do grupo; no caso do k-medoides, designa-se medoide do grupo.

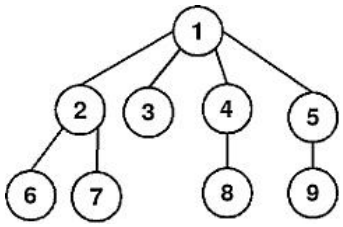
Árvore Geradora Mínima

INCIDÊNCIA EM PROVA: BAIXÍSSIMA

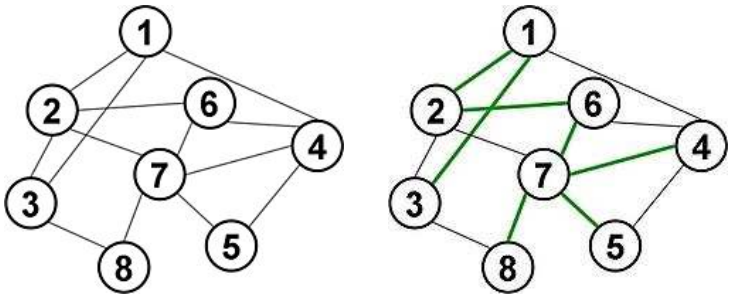
A Árvore Geradora Mínima (Minimal Spanning Tree – MST) é um algoritmo utilizado em diversas áreas de tecnologia da informação. Em nosso contexto, trata-se de um método baseado em teoria dos grafos[19] utilizado para segmentar um conjunto de dados em diferentes grupos. *Qual é o diferencial dela?* Bem, nos algoritmos anteriores, partia-se sempre de um conjunto inicial de protótipos e havia um processo iterativo de alocação de objetos aos protótipos.

Lembrando que os protótipos são aqueles "chutes" iniciais (centroide no k-Médias e medoide no k-Medoide). Esse processo ocorria repetidamente, calculando novos protótipos a fim de particionar um conjunto de dados em k grupos e minimizar uma função de custo proporcional à distância intragrupo. Aqui é diferente: não há definição de protótipos para segmentar a base de dados em diferentes grupos.

Podemos afirmar que uma árvore é dita geradora se ela interliga (direta ou indiretamente) todos os nós de um grafo, ou seja, se todos os nós do grafo fazem parte da árvore - não ficou nenhum nó "isolado". Perceba que, por exemplo, os nós 1 e 8 da árvore seguinte não estão diretamente ligados (não existe aresta entre eles), mas mesmo assim ambos fazem parte da árvore. Se não existisse a aresta que liga 4 e 8, o nó 8 ficaria "sozinho", isolado do resto da estrutura.

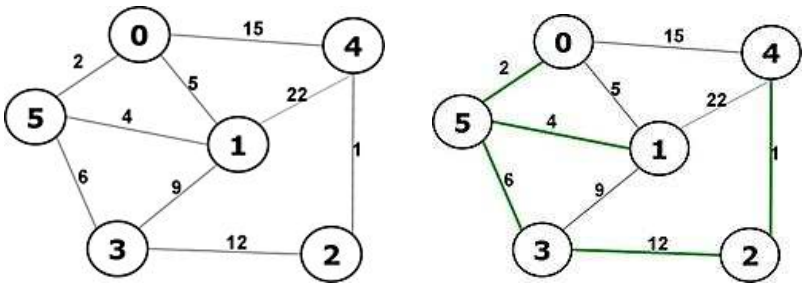


Os nós {1,2,3,4,5,6,7,9} continuariam a formar uma árvore, mas essa árvore não seria mais geradora, pois não inclui todos os nós do grafo - o nó 8 ficou de fora. A imagem da esquerda logo abaixo apresenta um grafo qualquer com 8 nós e 13 arestas. Vamos escolher somente algumas dessas arestas (o menor número possível), de forma que o grafo fique conectado. Na imagem da direita, as arestas escolhidas estão destacadas.



[19] Grafo é uma estrutura de dados que consiste em vértices ou nós conectados por arestas ou linhas, sendo utilizado para armazenar relações ou conexões entre objetos (Ex: redes sociais, redes de computadores, rotas de entrega, etc).

Não precisamos escolher nenhuma aresta a mais, já que todos os nós já estão conectados. Também não podemos escolher nenhuma aresta a menos, pois estaríamos desconectando os nós. Perceba como escolhemos justamente uma árvore geradora - um grafo sem ciclos (árvore) que conecta todos os nós (geradora). Lembrando que esse é apenas um exemplo, mas existiam outras escolhas de arestas para chegar ao mesmo objetivo.



Dito isso, agora nós queremos chegar à árvore geradora mínima! Para tal, vamos levar em consideração o peso das arestas - queremos escolher, entre as árvores geradoras possíveis, aquela que possua o menor peso total, onde o peso total é dado pela soma dos pesos das arestas da árvore. Vejamos um exemplo: na esquerda, temos um grafo qualquer com 6 nós e 9 arestas (com seus respectivos pesos).

Na direita, temos as arestas escolhidas que formam uma árvore geradora de peso 1+2+4+5+12 = 25. Como essa é a árvore geradora que gera o menor peso, ela é chamada de árvore geradora mínima. Existem diversos algoritmos que permitem calcular qual é a árvore geradora mínima, mas não vamos entrar nesses detalhes. É importante também saber que o peso da aresta pode representar qualquer valor em um problema real, como custo, fluxo, confiabilidade, tempo, entre outros.

Em nosso contexto, a Árvore Geradora Mínima permite identificar os limites de valores de um conjunto de dados que permitem dividir essa base de dados em grupos (*clusters*).

(TRE/BA – 2017) O agrupamento de dados no processo de data mining procura, em uma massa de dados que caracterizam uma população de indivíduos, grupos semelhantes e diferentes. O algoritmo baseado na teoria dos grafos e que dispensa a definição de protótipos utilizado para segmentar a base de dados em diferentes grupos é denominado:

a) K média.
b) K medoides.
c) Apriori.
d) DBSCAN.
e) Árvore geradora mínima.

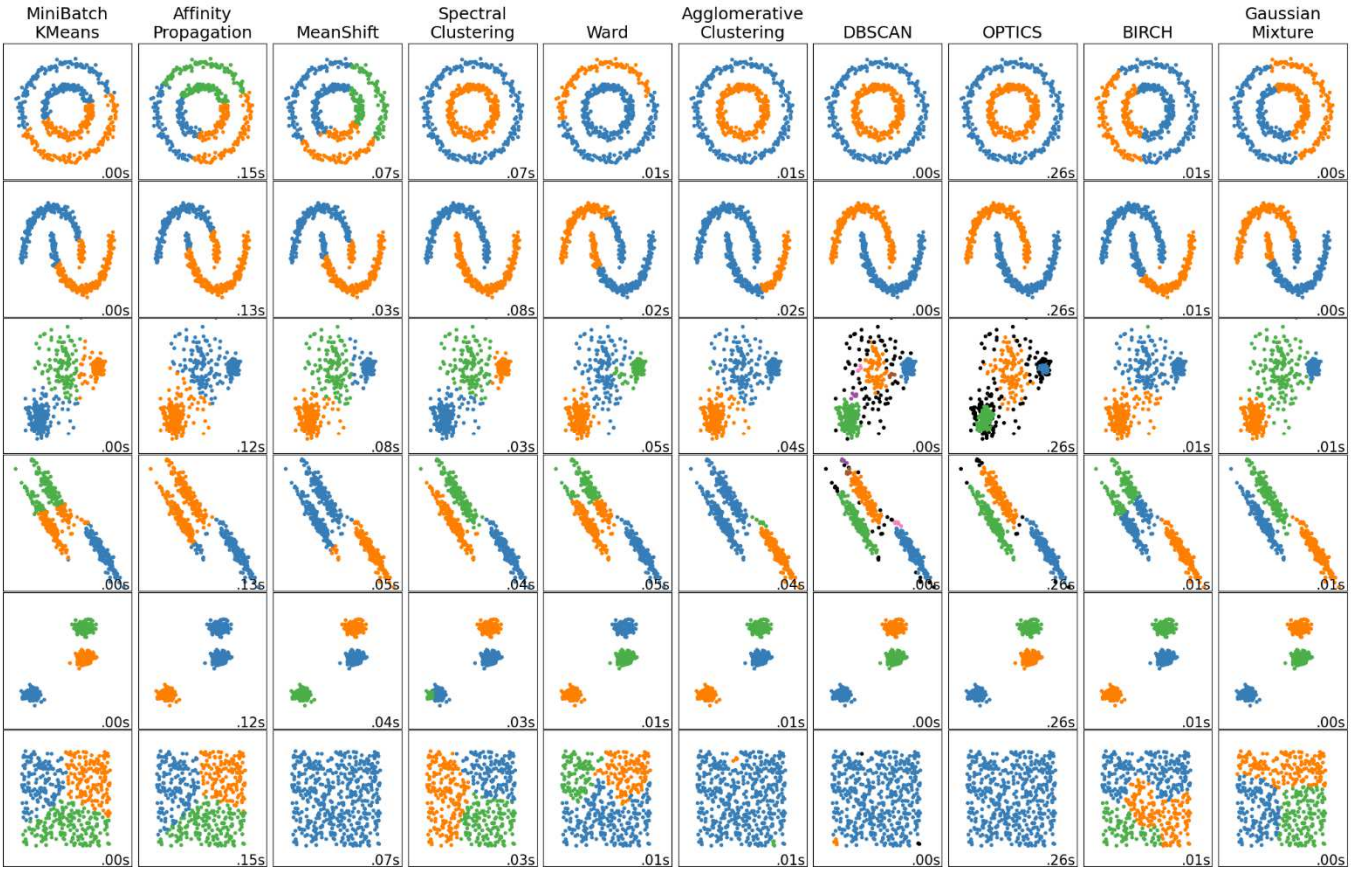
_____

Comentários: dentre as opções apresentadas, a única que é baseada em teoria dos grafos e dispensa a definição de protótipos para segmentar uma base de dados em diferentes grupos é a Árvore Geradora Mínima (Letra E).

DBSCAN

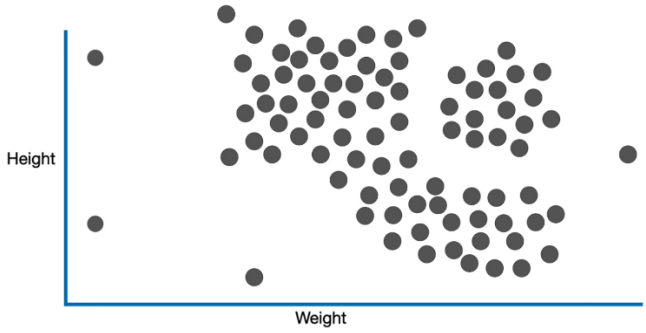
INCIDÊNCIA EM PROVA: BAIXÍSSIMA

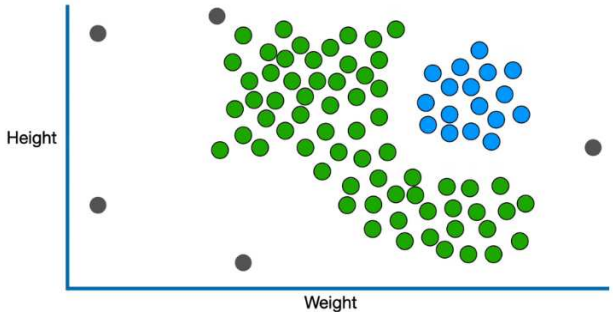
DBSCAN (*Density Based Spatial Clustering of Applications With Noise*) é um algoritmo utilizado para encontrar agrupamentos de diferentes formatos e ruído nas bases de dados, baseado na densidade de objetos no espaço. Galera, nós vimos vários exemplos de pontos plotados em um plano cartesiano e como os algoritmos de agrupamento funcionam sobre eles. Ocorre que alguns métodos funcionam melhor ou pior dependendo do formato dos dados. *Como assim, Diego?*



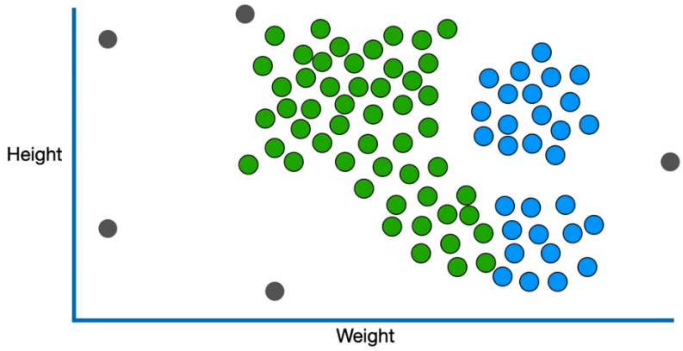
Na imagem acima, temos um mesmo conjunto de dados agrupados por meio de diversos algoritmos de agrupamento diferentes. Vejam que os resultados variam bastante dependendo da disposição dos pontos de dados. Vocês devem se lembrar também que eu falei que algoritmos como o k-Means são bastante sensíveis a ruídos e anomalias. Além disso, nem sempre temos apenas duas dimensões como na imagem anterior – por vezes, temos dezenas de dimensões.

Vejam o conjunto de dados da imagem à esquerda! Nós, humanos, conseguimos bater o olho e ver com certa facilidade dois grupos e alguns dados anômalos (conforme a imagem à direita).





No entanto, por conta de ruídos, anomalias e do formato dos pontos verdes contornando os pontos azuis, o algoritmo k-Médias teria dificuldade de agrupar os dados podendo gerar algo assim:



Então vamos listar os problemas dos algoritmos vistos até agora: eles têm dificuldades com alguns formatos de pontos de dados; são bastante sensíveis a ruídos e anomalias de dados; e não são ideais para problemas com mais de duas dimensões. É esse o contexto em que o DBSCAN chega para bilhar! Ele é especialista justamente em realizar agrupamentos de dados espaciais (> 2 dimensões) em aplicações com ruído e dados anômalos.

Para tal, ele se baseia na densidade dos pontos de dados, isto é, a quantidade de objetos dentro de um raio de vizinhança. *Como assim, Diego?* Na primeira imagem acima (dos pontos cinzas), como nós – humanos conseguimos identificar dois grupos por meio da densidade dos pontos. Tem um bolinho de pontos juntos de um lado e um bolinha de pontos juntos do outro. O DBSCAN faz algo muito parecido e, por isso, diz-se que ele é baseado em densidade.

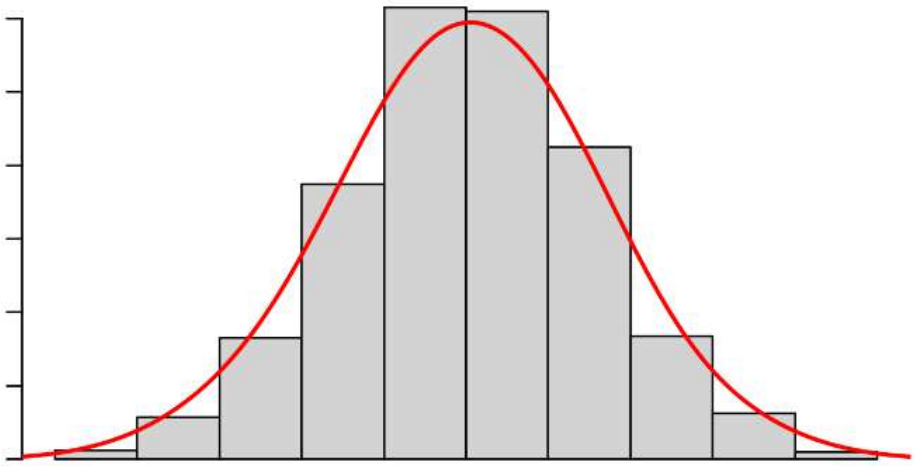
Justamente por conta de ser baseado em densidade é que ele lida bem com outliers (dados anômalos). Como a densidade de pontos é baixa próximo de dados anômalos, ele consegue identificá-los com maior facilidade. Note, por meio da imagem anterior, que métodos baseados em densidade são adequados para descobrir agrupamentos com forma arbitrária, tais como elíptica, cilíndrica ou espiralada.

Misturas Gaussianas

INCIDÊNCIA EM PROVA: BAIXÍSSIMA

Vamos passo a passo para entender esse algoritmo: em estatística, **uma distribuição é uma função matemática que descreve a probabilidade de ocorrência de um determinado evento ou valor em um conjunto de dados.** Em outras palavras, a distribuição descreve a maneira como os dados são espalhados ou distribuídos em torno de um valor central. Existe uma distribuição muito famosa chamada distribuição normal. *Vocês já ouviram falar?*

Essa distribuição é caracterizada por sua forma de sino simétrica e é frequentemente utilizada para modelar fenômenos naturais que seguem uma distribuição simétrica em forma de sino, como a altura de uma população, o peso, o tempo que leva para realizar uma tarefa, entre outros. **Ela é definida por dois parâmetros: sua média e seu desvio padrão.** A média representa o centro da distribuição, enquanto o desvio padrão representa a dispersão dos dados em torno da média.



A distribuição normal é simétrica em torno da média, o que significa que a mesma quantidade de dados está localizada em ambos os lados da média. Quem descobriu a distribuição normal foi um matemático alemão chamado Carl Friedrich Gauss e, por essa razão, ela ficou conhecida também como Distribuição Gaussiana. Pulamos dois séculos para os dias atuais e chegamos ao Modelo de Misturas Gaussianas (*Gaussian Mixture Model* – GMM). *O que seria isso?*

Matematicamente falando, uma mistura gaussiana é uma função composta por várias gaussianas, cada uma identificada por $k \in \{1, ..., k\}$, onde k é o número de clusters do conjunto de dados. Cada gaussiano k na mistura é composto pelos seguintes parâmetros: uma média μ que define seu centro; uma covariância $\sum$ que define sua largura; e uma probabilidade de mistura π que define quão grande ou pequena será a função gaussiana.

De outra forma, podemos dizer que uma mistura gaussiana é um modelo estatístico e uma técnica de aprendizado de máquina não supervisionado que assume que os dados são gerados a partir de uma mistura de várias distribuições gaussianas. Logo, a distribuição gaussiana é a distribuição básica utilizada no GMM para modelar cada componente da mistura. Esse modelo tem o objetivo de modelar distribuições de dados complexas.

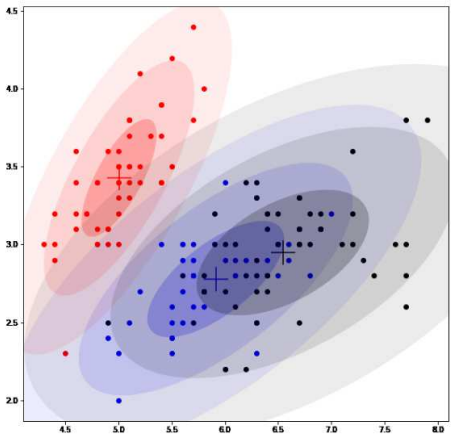
Muitas vezes, os dados reais têm uma distribuição que não é facilmente descrita por uma única distribuição gaussiana. Logo, para modelar esse conjunto de dados complexos, podemos utilizar uma mistura de diversas distribuições gaussianas, cada uma com seus próprios parâmetros (média e desvio padrão). **Cada componente da mistura é ponderado por um coeficiente que representa a proporção de dados que pertencem a essa componente.**

O modelo de misturas gaussianas é frequentemente utilizado em tarefas de agrupamento, análise de misturas de populações, detecção de anomalias e modelagens de dados complexos em geral. Em análise de agrupamento, o objetivo é agrupar os dados em clusters de modo que as observações dentro de um mesmo cluster sejam mais semelhantes entre si do que com as observações em outros clusters.

O GMM pode ser usado para modelar a distribuição dos dados em cada cluster, permitindo que cada cluster seja representado por uma distribuição gaussiana diferente. Tudo começa com uma estimativa inicial do número de clusters e dos parâmetros das distribuições gaussianas, incluindo as médias, desvios padrão e coeficientes de mistura. Em seguida, o algoritmo itera entre duas etapas: a etapa de expectativa (E) e a etapa de maximização (M).

Na etapa E, é calculada a probabilidade de cada observação pertencer a cada cluster, com base nos parâmetros atuais do modelo. **Essa probabilidade é calculada usando a função de densidade de probabilidade das distribuições gaussianas.** Já na etapa M, os parâmetros do modelo são atualizados para maximizar a verossimilhança dos dados, considerando as probabilidades calculadas na etapa E.

Essa etapa envolve a atualização dos parâmetros das distribuições gaussianas, bem como o cálculo dos novos coeficientes de mistura. As etapas E e M são repetidas até que a verossimilhança dos dados não melhore significativamente ou até que um critério de convergência seja atingido. Ao final do algoritmo, cada observação é atribuída ao cluster com maior probabilidade. *Bacana?* **Para fechar, podemos dizer que GMM é uma generalização do K-Médias.**



Em outras palavras, nós podemos afirmar que o K-Médias é um caso especial do modelo de misturas gaussianas, em que as distribuições gaussianas são consideradas iguais e esféricas, com o mesmo desvio padrão em todas as direções! **Por outro lado, o GMM é mais flexível do que o K-Médias, uma vez que pode modelar conjuntos de dados com distribuições complexas e não necessariamente esféricas – isto é, com um formato mais elipsóide.** No entanto, o GMM pode ser mais difícil de ajustar, pois envolve a estimação de muitos parâmetros. Logo, o K-Means pode ser visto como uma aproximação simplificada do GMM, com menos parâmetros a serem ajustados. *Fechado?* Fim!

Single/Complete-Linkage

INCIDÊNCIA EM PROVA: BAIXÍSSIMA

Para finalizar os algoritmos de agrupamento, vamos falar bem rapidamente sobre o Single-Linkage e o Complete-Linkage. O Single-Linkage é um método aglomerativo de agrupamento hierárquico no qual novos grupos são criados unindo os grupos mais semelhantes. O agrupamento inicial é formado apenas por singletons (grupos formados por apenas um objeto), e a cada iteração um novo grupo é formado por meio da união dos dois grupos mais similares da iteração anterior.

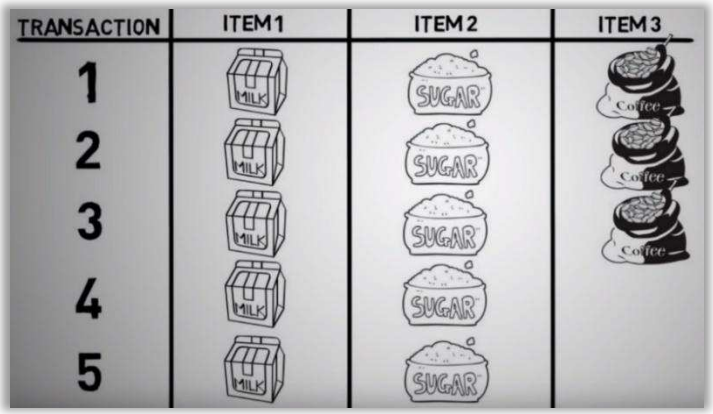
Neste método, a distância (proximidade) entre o novo grupo e os demais é determinada como a menor distância entre os elementos do novo grupo e os grupos remanescentes. Já o complete-linkage é o inverso: a distância do novo grupo aos demais é calculada como a distância máxima entre os elementos do novo grupo aos grupos restantes. Eu ainda não vi caindo em provas de tecnologia da informação – apenas em provas de estatístico cujo foco é diferente do nosso.

Regras de Associação

INCIDÊNCIA EM PROVA: ALTA

Uma das principais tecnologias de mineração de dados envolve a descoberta de regras de associação. O banco de dados é considerado uma coleção de transações, cada uma envolvendo um conjunto de itens. Um exemplo comum é o de dados de um carrinho de supermercado. Nesse contexto, o carrinho de supermercado corresponde aos conjuntos de itens que um consumidor compra em um supermercado durante uma visita.

Para esse exemplo, podemos definir as seguintes regras de associação: se leite for comprado, açúcar provavelmente também será; se açúcar for comprado, leite provavelmente também será; se leite e açúcar forem comprados, café provavelmente também será em 60% das transações. **Percebam que geralmente as regras de associação são escritas em um formato como: se [algo acontecer], então [algo acontecerá] ou se [evento], então [ações].**



É possível pensar em dezenas de outros exemplos! **Mulheres que compram um sapato em uma loja tendem a comprar uma bolsa também na mesma loja**. Músicos que compram uma nova guitarra tendem a comprar também novas palhetas. *Vocês se lembram do exemplo das cervejas e das fraldas?* Pois é, também são outro excelente exemplo de como utilizar regras de associação para minerar dados. Vamos resumir...

Na mineração de dados, uma regra de associação é um evento que relaciona a presença de um conjunto de itens com outra faixa de valores de um outro conjunto de variáveis. **Uma regra de associação pode ser vista como uma expressão da forma X → Y, onde há a relação dos valores de X e Y em um certo conjunto de valores (Ex: {fralda} → {cerveja}).** Isso significa que quem compra leite e fralda também tende a comprar cerveja na mesma transação[20].

Existem duas variações comuns de regras de associação: padrões sequenciais e os padrões temporais. Nos padrões sequenciais, uma sequência de ações é buscada. Exemplo: se um paciente passou por uma cirurgia de ponte de safena para artérias bloqueadas e um aneurisma e, depois, desenvolveu ureia sanguínea alta dentro de um ano de cirurgia, ele provavelmente sofrerá de insuficiência renal nos próximos dezoito meses.

Já nos padrões temporais (ou padrões dentro de série temporal), as similaridades podem ser detectadas **dentro de posições de uma série temporal de dados**, que é uma sequência de dados tomados em intervalos regulares, como vendas diárias ou preços de ações de fechamento diário.

[20] Essa é uma técnica de coocorrência conhecida como Análise de Cesta de Compra/Mercado, cujo objetivo é identificar combinações de itens que ocorrem com frequência significativa em bancos de dados e podem caracterizar, por exemplo, hábitos de consumo de clientes em um supermercado.

Exemplo: uma ação de uma companhia de energia e outra de uma companhia financeira tiveram o mesmo padrão durante um período de três anos em relação a preço de fechamento de ações.

Vamos diferenciar tudo isso? **A técnica de regras de associação visa simplesmente descobrir o relacionamento ou correlação entre variáveis de um banco de dados**. Já a técnica de Padrões Sequenciais busca descobrir padrões sequenciais de eventos de forma equivalente a certos relacionamentos temporais. Por fim, a técnica de Padrões Temporais é bastante semelhante à técnica de Padrões Sequenciais, mas sempre envolve um fator temporal que permite diferenciá-los.

(MAPA – 2010) No tocante ao data mining, o tipo de informação que é extraída desta ferramenta, em que se pode descobrir, por exemplo, que, quando se compra uma casa, compra-se também uma geladeira, considerando-se que essas compras são realizadas num período de duas semanas, é:

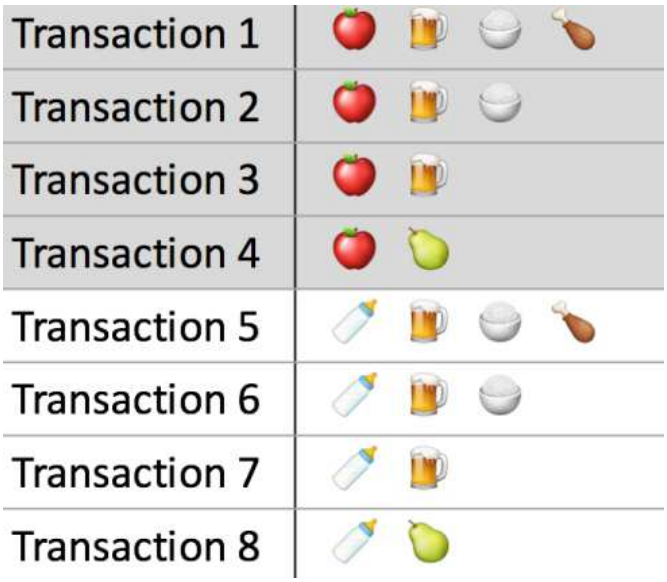
a) classificação b) aglomeração c) prognósticos d) associação e) sequência

Comentários: quando fiz essa questão fui direto na associação, mas observe que ele menciona que essas compras foram realizadas num período de duas semanas. Logo, temos uma alternativa mais específica: padrões temporais. A detecção de padrões sequenciais é equivalente à detecção de associações entre eventos com certos relacionamentos temporais (Letra E).

Existem duas medidas capazes de indicar a qualidade ou grau de certeza de uma regra de associação. São elas: suporte e confiança. Vejamos...

MEDIDAS DE INTERESSE	DESCRIÇÃO
SUPORTE/ PREVALÊNCIA	Trata-se da frequência com que um conjunto de itens específicos ocorrem no banco de dados, isto é, o percentual de transações que contém todos os itens em um conjunto. Em termos matemáticos, a medida de suporte para uma regra X → Y é a frequência em que o conjunto de itens aparece nas transações do banco de dados. Um suporte alto nos leva a crer que os itens do conjunto X e Y costumam ser comprados juntos, pois ocorrem com alta frequência no banco (Ex: 70% das compras realizadas em um mercado contém arroz e refrigerante).
CONFIANÇA/ FORÇA	Trata-se da probabilidade de que exista uma relação entre itens. Em termos matemáticos, a medida de confiança para uma regra X → Y é a força com que essa regra funciona. Ela é calculada pela frequência dos itens Y serem comprados dado que os itens X foram comprados. Uma confiança alta nos leva a crer que exista uma alta probabilidade de que se X for comprado, Y também será (Ex: existe uma probabilidade de 70% de que clientes que compram fraldas também comprem cerveja).

Vamos ver um exemplo de cálculo de suporte e confiança porque esse conteúdo será importante para entender o funcionamento dos algoritmos de regras de associação. Vejamos a tabela a seguir:



REGRA DE ASSOCIAÇÃO: {MAÇÃ → CERVEJA}

Para realizar a medida de suporte, basta calcular a razão da sua frequência pela quantidade de transações. Como maçã aparece em 4 das 8 transações, possui um suporte de 50%:

$$\text{Support } \{🍎\} = \frac{4}{8}$$

Para realizar a medida de confiança, basta calcular em quantas dessas quatro transações ocorreu cerveja como consequente. Como cerveja aparece em 3 das 4 transações, temos confiança de 75%.

$$\text{Confidence } \{🍎 \rightarrow 🍺\} = \frac{\text{Support } \{🍎, 🍺\}}{\text{Support } \{🍎\}} = \frac{3}{4}$$

Bacana! Agora estamos prontos para entender os principais algoritmos de regras de associação. Vamos iniciar pelo algoritmo *apriori*.

Apriori

INCIDÊNCIA EM PROVA: BAIXA

O algoritmo *apriori* é um método de mineração de dados não supervisionado utilizado para minerar conjuntos de dados frequentes e regras de associação relevantes. Para entender seu funcionamento, vamos ver um exemplo com dados hipotéticos e vamos passar com cada um dos passos quem permitem chegar a essas regras. A tabela seguinte apresenta um conjunto de transações com os respectivos produtos comprados em um supermercado hipotético:

TRANSAÇÃO	PRODUTOS
1	Cerveja, Vinho, Queijo
2	Cerveja, Batata Chips
3	Ovos, Farinha de Trigo, Manteiga, Queijo
4	Ovos, Farinha de Trigo, Manteiga, Cerveja, Batata Chips
5	Vinho, Queijo
6	Batata Chips
7	Ovos, Farinha de Trigo, Manteiga, Vinho, Queijo
8	Ovos, Farinha de Trigo, Manteiga, Cerveja, Batata Chips
9	Vinho, Cerveja
10	Cerveja, Batata Chips
11	Manteiga, Ovos
12	Cerveja, Batata Chips
13	Farinha de Trigo, Ovos
14	Cerveja, Batata Chips
15	Ovos, Farinha de Trigo, Manteiga, Vinho, Queijo
16	Cerveja, Vinho, Batata Chips, Queijo
17	Vinho, Queijo
18	Cerveja, Batata Chips
19	Vinho, Queijo
20	Cerveja, Batata Chips

O primeiro passo do algoritmo é calcular a medida de suporte para cada item individual. *Por quê?* Porque precisamos reduzir o custo computacional! A quantidade de regras de associação possíveis de serem geradas em bases transacionais cresce exponencialmente com o número de itens da base. Um conjunto de dados simples e pequeno já é capaz de produzir centenas de regras de associação, onerando o processamento dos dados.

Qualquer algoritmo de força bruta, ou seja, que gera todas as combinações possíveis de itens para depois buscar as regras relevantes, não será computacionalmente factível para diversas bases transacionais. Dito isso, uma forma de reduzir o custo computacional dos algoritmos de mineração

de regras de associação é desacoplar os requisitos de suporte e os requisitos de confiança mínima das regras.

Como o suporte da regra depende apenas do conjunto de itens, conjuntos de itens pouco frequentes podem ser eliminados no início do processo sem que seja necessário calcular sua confiança. Eu sei que deve ter embaralhado a cabeça de vocês, então eu vou explicar um pouco melhor! *Qual é o objetivo do algoritmo apriori?* Minerar conjuntos de itens frequentes a partir de regras de associação relevantes: com alto suporte a alta confiança.

Uma maneira de fazer isso seria ir testando todas as combinações possíveis de itens até encontrar regras de associação significativas, mas isso tem um custo computacional tão alto que torna essa estratégia inviável. Dito isso, antes de me preocupar em calcular a confiança, eu posso inicialmente calcular o suporte. *Por quê?* Porque se o suporte for baixo, eu já descarto esse item! *Isso faz sentido para vocês?* Se um item tem um suporte baixo, significa que ele tem uma frequência baixa.

No contexto de um supermercado, um item com suporte baixo pode ser jiló. Quase ninguém gosta de jiló, logo ele deve aparecer em pouquíssimas transações, portanto possui um baixo suporte. *Entendido?* Então vamos voltar para o nosso exemplo! Eu preciso calcular o suporte para cada item individual. Como o suporte é apenas a frequência (número de ocorrências) de um determinado produto, basta calcular as quantidades:

PRODUTO	FREQUÊNCIA
VINHO	8
QUEIJO	8
CERVEJA	11
BATATA CHIPS	10
OVOS	7
FARINHA DE TRIGO	6
MANTEIGA	6

O segundo passo do nosso algoritmo é decidir qual será o suporte mínimo! Produtos cujo suporte seja menor que esse limite mínimo são produtos com baixa frequência que podem ser descartados. *Como esse patamar é escolhido?* Ele é escolhido basicamente de forma arbitrária! Em nosso caso, vamos escolher um suporte mínimo de 7. Isso significa que produtos com suporte (frequência de ocorrência) abaixo desse valor serão descartados: Farinha de Trigo e Manteiga.

O terceiro passo consiste em realizar o mesmo procedimento anterior, mas agora utilizando pares de produtos em vez de produtos individuais. Como já descartamos alguns produtos no passo anterior, teremos um pouco menos de combinações para testar. Nós vamos ignorar todas as combinações que contenham Farinha de Trigo e Manteiga. O resultado dessa análise é apresentado na tabela seguinte:

CONJUNTO DE ITENS	FREQUÊNCIA
VINHO, QUEIJO	7
VINHO, CERVEJA	2
VINHO, BATATA CHIPS	1
VINHO, OVOS	2
QUEIJO, CERVEJA	2
QUEIJO, BATATA CHIPS	1
QUEIJO, OVOS	3
CERVEJA, BATATA CHIPS	9
CERVEJA, OVOS	2
BATATA CHIPS, OVOS	2

Da mesma forma, é possível notar que apenas duas combinações possuem medida de suporte maior que 7: {Vinho, Queijo} e {Cerveja, Batata Chips}. O próximo passo é continuar com o procedimento anterior, mas para conjuntos maiores. Já testamos produtos individuais, testamos conjuntos de dois itens e agora é o momento de testar conjuntos de três itens. Vamos ver todas as possíveis combinações:

CONJUNTO DE ITENS	FREQUÊNCIA
VINHO, QUEIJO, CERVEJA	1
VINHO, QUEIJO, BATATA CHIPS	1
CERVEJA, BATATA CHIPS, VINHO	1
QUEIJO, CERVEJA, BATATA CHIPS	1

Note que nenhum conjunto de três itens contém um suporte maior que o mínimo pré-definido. Na verdade, nem era necessário calcular – bastava lembrar que {Queijo, Cerveja}, {Queijo, Batata Chips}, {Vinho, Batata Chips} e {Queijo, Cerveja} já haviam sido descartados no passo anterior. Agora que já temos o maior conjunto de itens frequentes, basta gerar as regras de associação e realizar o cálculo da confiança. *Fechado?*

Bem, como todos os conjuntos de três itens foram descartados, então ficamos com os conjuntos de dois itens: {Vinho, Queijo} e {Cerveja, Batata Chips}. Agora vamos pegar todos os subconjuntos desses conjuntos, criar as regras e verificar se elas possuem um patamar mínimo de confiança. *E qual seria esse valor, Diego?* Da mesma forma do suporte, trata-se de um valor arbitrário – nós vamos escolher o valor mínimo de 85%. Para cada subconjunto S de I, teremos a regra:

$$S \mapsto (I - S)$$

Logo, se temos um conjunto de dados I = {A, B, C}, temos os subconjuntos {A}, {B}, {C}, {A,B}, {A, C} e {B,C}. Dessa forma, uma regra poderia ser: {A} ↦ {A, B, C} – {A}, logo {A} ↦ {B, C}. Vamos lá...

- **Conjunto de Itens 1:** {Vinho, Queijo}
- **Subconjuntos:** {Vinho}, {Queijo}, {Vinho, Queijo}

Regra de Associação 1:
{Vinho} ↦ {Vinho, Queijo} – {Vinho} = **{Vinho} ↦ {Queijo}**

Regra de Associação 2:
{Queijo} ↦ {Vinho, Queijo} – {Queijo} = **{Queijo} ↦ {Vinho}**

- **Conjunto de Itens 2:** {Cerveja, Batata Chips}
- **Subconjuntos:** {Cerveja}, {Batata Chips}, {Cerveja, Batata Chips}

Regra de Associação 3:
{Cerveja} ↦ {Cerveja, Batata Chips} – {Cerveja} = **{Cerveja} ↦ {Batata Chips}**

Regra de Associação 4:
{Batata Chips} ↦ {Cerveja, Batata Chips} – {Batata Chips} = **{Batata Chips} ↦ {Cerveja}**

Lembrando que desconsideramos regras cujo antecedente ou consequente sejam nulos. Agora é o momento de calcular a Confiança = Suporte(I)/Suporte(S). Nós já calculamos esses valores antes:

PRODUTO	FREQUÊNCIA
VINHO	8
QUEIJO	8
CERVEJA	11
BATATA CHIPS	10

CONJUNTO DE ITENS	FREQUÊNCIA
VINHO, QUEIJO	7
CERVEJA, BATATA CHIPS	9

O cálculo da confiança é bastante intuitivo: nós vamos ver a proporção – dentre as transações que contenham o conjunto de itens – daquelas transações em que a regra de transação é válida. Por exemplo: dada uma regra de associação X ↦ Y, se temos 10 transações em que ocorre o item {X}, vamos calcular em quantas dessas transações X ocorreu também Y. *Entendido?* Então, faremos isso para as quatro regras de transação descobertas:

Regra de Associação 1: {Vinho} ↦ {Queijo}
Confiança: Suporte(Vinho, Queijo)/Suporte(Vinho) = 7/8 = 87,5% > 85% (Regra aceita)

Regra de Associação 2: {Queijo} ↦ {Vinho}

Confiança: Suporte(Queijo,Vinho)/Suporte(Queijo) = 7/8 = 87,5% > 85% (Regra aceita)

Regra de Associação 3: {Cerveja} ↦ {Batata Chips}
Confiança: Suporte(Cerveja, Batata Chips)/Suporte(Cerveja) = 9/11 = 81% < 85% (Regra rejeitada)

Regra de Associação 4: {Batata Chips} ↦ {Cerveja}
Confiança: Suporte(Batata Chips,Cerveja)/Suporte(Batata Chips) = 9/10 = 90% > 85% (Regra aceita)

Em síntese: o algoritmo *apriori* é um método baseado no conceito de itens frequentes que constrói um conjunto de regras das mais simples (único item) às mais complexas (múltiplos itens), sendo que – para cada nível de regra – calcula-se o número de ocorrências nos dados (suporte) e eliminam-se as regras com suporte inferior a um determinado patamar mínimo de tal forma que regras que subsistirem sejam expandidas para mais um produto e assim por diante.

(Prefeitura de Manaus/AM – 2022) A mineração de dados (Data Mining) envolve um conjunto de algoritmos e ferramentas que são utilizados para a exploração de dados. Assinale o algoritmo/método usado na extração de regras de associação.

a) Apriori.
b) C4.5
c) K-mean.
d) Naive Bayes.
e) PageRank.

_____
Comentários: algoritmo utilizado para extração de regras de associação é o *apriori* (Letra A).

FP-Growth

INCIDÊNCIA EM PROVA: BAIXÍSSIMA

O FP-Growth (*Frequent Pattern – Growth*) é uma técnica utilizada para descobrir padrões de associação entre itens em um conjunto de dados muito eficaz para trabalhar com grandes conjuntos de dados usando uma memória limitada. A técnica é baseada na construção de uma árvore de frequência de itens (em uma abordagem bottom-up), que é usada para descobrir padrões frequentes nos dados.

Inicia-se com os itens individuais e eles são agrupados em padrões mais complexos. O algoritmo é muito eficiente e pode ser usado para descobrir padrões em conjuntos de dados muito grandes. O método *apriori* tem dificuldade para tratar uma grande quantidade de conjuntos candidatos e execução de repetidas passagens pela base de dados. Já a árvore é capaz de armazenar de forma comprimida a informação sobre padrões frequentes. Esse método ainda não foi cobrado em prova.

CRISP-DM

INCIDÊNCIA EM PROVA: BAIXA

Galera, vocês já devem ter notado que nós – malucos da área de tecnologia da informação – adoramos processos! É impressionante, nem advogados e juízes gostam tanto de processos! Vejam só: nós temos um modelo de referência para gerenciamento de projetos, um para análise de negócio, um para gerenciamento de serviços, um para qualidade de software, um para governança de tecnologia da informação... **e todos eles são baseados em processos**!

Ora, se nós temos modelos de referência para esse bando de coisas, por que não temos um modelo de referência para a mineração de dados? Pois é, mas nós já temos!!! O **CRISP-DM** (**C**ross **I**ndustry **S**tandard **P**rocess for **D**ata **M**ining) é um modelo de referência[21] de mineração de dados que descreve um conjunto de processos para realizar projetos de mineração de dados em uma organização baseado nas melhores práticas utilizadas por profissionais e acadêmicos do ramo.

Galera, como nascem todos esses modelos de referência? **Em geral, reúnem-se os maiores especialistas da área e eles exibem como fazem para resolver problemas recorrentes.** As ideias são, então, organizadas em forma de processos e tarefas em um documento de modo que outras pessoas que desejem resolver problemas semelhantes (em nosso caso, projetos de mineração de dados) possam usá-lo como referência. *Entendido?* Vamos ver agora o que dizia o site oficial:

> *"O Projeto CRISP-DM desenvolveu um modelo de processos de mineração de dados com foco industrial e independente de ferramentas. Partindo dos processos embrionários de descoberta de conhecimento usados atualmente na indústria e respondendo diretamente aos requisitos do usuário, este projeto definiu e validou um processo de mineração de dados aplicável em diversos setores da indústria. Isso tornará grandes projetos de mineração de dados mais rápidos, mais baratos, mais confiáveis e mais gerenciáveis. Até casos de mineração de dados em pequena escala se beneficiarão do uso do CRISP-DM".*

Dessa forma, caso você queira fazer um projeto de mineração de dados em sua organização, você poderá utilizar esse modelo de processos como referência – **é importante destacar que se trata de uma metodologia não proprietária que pode ser aplicada livremente a qualquer projeto independentemente do tamanho ou tipo do negócio**. Bem, essa metodologia possui um ciclo de vida não-linear composto por seis fases ou etapas. Vejamos...

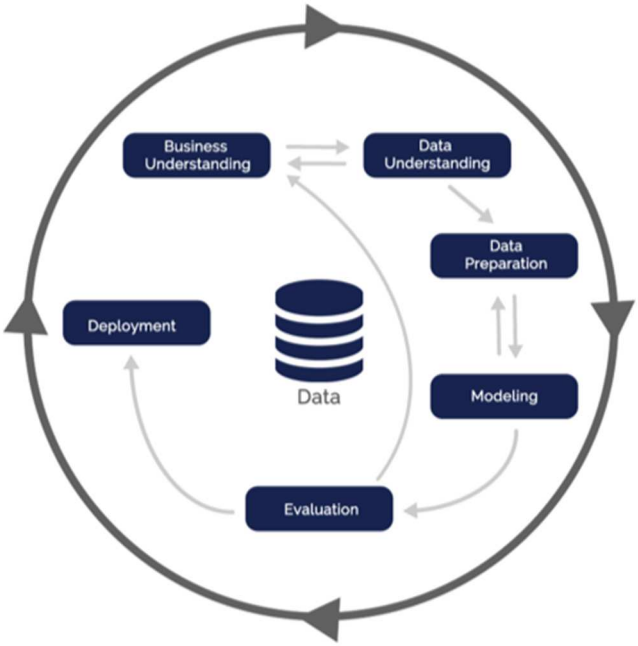
(TCE/PA – 2016) CRISP-DM é uma metodologia proprietária que identifica as fases Business Understanding e Data Understanding na implantação de um projeto de data mining.

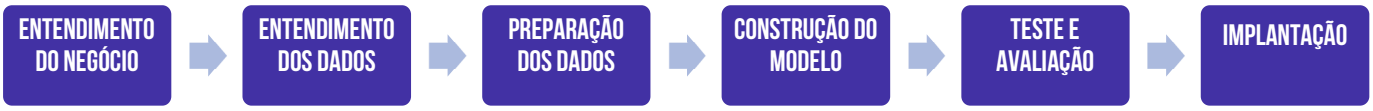
Comentários: na verdade, não se trata de uma metodologia proprietária (Errado).

[21] Também pode ser considerado um modelo de processos, um framework de processos, uma metodologia, entre outros.

Professor, essas fases devem ser executadas em uma sequência rigorosa? Não, conforme é possível ver na imagem a seguir, é sempre necessário ir e voltar entre diferentes fases – e isso depende do resultado de cada fase. **Observem também na imagem que as setas indicam as dependências mais importantes e frequentes entre as fases.** Além disso, o círculo externo simboliza a natureza cíclica da própria mineração de dados. *Como assim, Diego?*

Um processo de mineração de dados continua após a implantação de uma solução. As lições aprendidas durante o processo podem desencadear novas questões comerciais, geralmente mais focadas. Os processos subsequentes de mineração de dados se beneficiarão das experiências dos anteriores. *Beleza?* **As fases são: (1) Entendimento do Negócio; (2) Entendimento dos Dados; (3) Preparação dos Dados; (4) Modelagem; (5) Avaliação; e (6) Implantação.**





(TCE/RS – 2018) 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) Entendimento do Negócio, Entendimento dos Dados, Preparação dos Dados, Modelagem, Avaliação e Implantação.

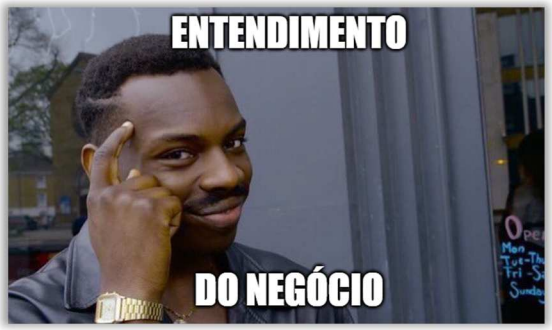
d) Preparação do Negócio, Replicação dos Dados, Indexação dos Dados, Diagramação do Negócio, Estimativa e Organização.

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

Comentários: as fases são: Entendimento do Negócio, Entendimento dos Dados, Preparação dos Dados, Modelagem, Avaliação e Implantação (Letra C).

Entendimento do Negócio

Essa fase inicial concentra-se no entendimento dos objetivos e requisitos do projeto de uma perspectiva de negócio e, em seguida, na conversão desse conhecimento em uma definição de problema de mineração de dados e em um plano preliminar desenvolvido para atingir os objetivos. **Em outras palavras, essa fase busca entender qual problema o negócio quer resolver!** *Professor, não entendi!* Calma, vamos lá...



Pessoal, é muito comum que uma área de tecnologia da informação faça um projeto de mineração de dados para uma área que ela não domina o assunto! Você pode aplicar a tecnologia de mineração à área de saúde, finanças, turismo, esportes, comércio, etc. **A galera da área de tecnologia entende de tecnologia, não entende de finanças por exemplo. Logo, antes de começar o projeto, é importantíssimo que ela entenda do negócio.**

Essa é a fase em que os analistas de tecnologia vão entender qual é o problema que o negócio quer resolver, seus objetivos, como está a situação atual, quais são os requisitos do projeto, quais são os principais pressupostos e limitações, em qual forma será a entrega dos resultados, quais são os critérios de sucesso, entre outros parâmetros – **tudo isso para desenvolver um planejamento de projeto**. *Bacana?*

(TCE/PE – 2017) Durante a fase de entendimento do negócio, busca-se descrever claramente o problema, fazer a identificação dos dados e verificar se as variáveis relevantes para o projeto não são interdependentes.

Comentários: fazer a identificação dos dados e verificar se as variáveis relevantes para o projeto não são interdependentes (ou seja, não se sobrepõem) são atividades da fase de entendimento dos dados e, não, do negócio (Errado).

Entendimento dos Dados

A segunda fase começa com uma coleta inicial dos dados e prossegue com atividades para explorá-los com o intuito de obter um maior conhecimento e familiaridade. **Em seguida, busca-se avaliar a qualidade dos dados, descobrir as primeiras ideias sobre os dados ou detectar subconjuntos interessantes para formar hipóteses de informação ocultas e descobrir insights.** Essa fase também é responsável por descrever os dados – por vezes, utilizando estatísticas.

Na descrição, pode-se obter uma espécie de fotografia dos dados contendo a localização, o formato, a fonte, o número de registros, o número de atributos, como será a extração dos dados, e outras características que interessem, assegurando que esses dados consigam representar o problema em análise. **Esta etapa também envolve o que geralmente é denominado análise exploratória de dados.** Enfim... o lance aqui é ser o mais íntimo possível dos dados em análise.

Preparação dos Dados

Também chamada de pré-processamento, nessa fase ocorre a preparação dos dados para a fase de modelagem. Essa etapa ocorre quando já entendemos o problema do negócio e já exploramos os dados disponíveis. Ela abrange todas as atividades para construir o conjunto de dados final a partir dos dados brutos iniciais, isto é, aqueles que serão alimentados na ferramenta de modelagem. *Professor, e que atividades são essas?* Bem, me acompanhem...

Essa lista não é exaustiva, mas inclui tarefas como seleção de tabelas, integração, transformação, limpeza e organização de dados – além da seleção e engenharia de recursos. Essas atividades visam a melhoria na qualidade dos dados originais e para realizá-las existem ferramentas de mineração de dados que dispõem de funcionalidades específicas que garantem agilidade nas operações, ou seja, você não precisa fazer isso "na mão".

Galera, é bem provável que as tarefas de preparação de dados sejam executadas várias vezes, de forma iterativa e sem nenhuma sequência predefinida. **Além disso, trata-se da fase mais demorada, ocupando mais de 70% do tempo/esforço total gasto em qualquer projeto de ciência de dados.** *Por que?* Porque ela é a responsável por carregar os dados identificados na etapa anterior e prepará-los para análise por meio de métodos de mineração de dados.

Vamos falar em uma linguagem mais clara agora: no mundo real, os dados não estão todos organizados bonitinhos prontos para serem analisados. **Tem dado faltando, dado incompleto, dado errado, dados discrepantes, dados inconsistentes, dados em formatos estranhos, entre outros.** *Como nós vamos analisar dados totalmente desestruturados?* Logo, essa etapa busca organizá-los da melhor maneira possível.

Dessa forma, é necessário integrar os dados! **Há momentos em que os dados estão disponíveis em várias fontes e, portanto, precisam ser combinados com base em determinadas chaves ou atributos para melhor uso.** É necessário também selecionar os dados desejados no modelo! Por exemplo: talvez você não queira usar dados *outliers* (fora da curva) ou talvez você não queira utilizar todas as colunas de uma tabela. Enfim, escolha tudo que será relevante para seu modelo!

Legal, mas o que fazemos em relação aos dados incorretos? Bem, nós devemos limpá-los! **Datas em formatos incorretos e números inteiros sendo interpretados como textos são exemplos de "sujeira" que podem ser encontradas no seu dado – esse é o momento de tratá-las.** Há também momentos em que precisamos derivar ou gerar recursos a partir dos existentes. *Como assim?* Ex: algumas vezes, você deve derivar a idade de uma pessoa a partir da data de nascimento.

Enfim... todas essas atividades são realizadas para tratar a qualidade dos dados da melhor forma possível de modo que possam ser utilizadas pelos algoritmos de mineração de dados. *Fechou?*

Construção do Modelo

Também chamada de Modelagem, nessa fase ocorre a seleção das técnicas, ferramentas e algoritmos a serem utilizados, como também a elaboração e execução da modelagem sobre o conjunto de dados preparado na fase anterior. Aliás, retornar à fase de preparação é bem frequente e necessário nessa etapa – lembrem-se que se trata de um processo iterativo e cheio de vai e volta! Você pode criar diferentes modelos e compará-los na próxima fase! *Bacana?*

Em outras palavras, essa etapa utiliza os dados limpos e formatados preparados na etapa anterior para fins de modelagem. **Ela inclui a criação, avaliação e ajuste fino de modelos e parâmetros para valores ideais, com base nas expectativas e critérios estabelecidos durante a fase de entendimento dos negócios.** Dependendo da necessidade do negócio, a tarefa de mineração de dados pode ser de uma classificação, uma regressão, uma associação, uma clusterização, etc.

Teste e Avaliação

Bem, chegou a hora de avaliar os resultados da modelagem. *Você se lembra que nós definimos critérios de sucesso lá na primeira fase?* **Agora é a hora de verificar se ela foi atingida. Se não foi, é necessário voltar a primeira fase e entender o que deu de errado, determinar um novo escopo e tentar novamente.** Caso tudo tenha dado certo, é hora de seguir em frente para a última fase. *E o que mais?*

Nessa fase do projeto, você construiu um modelo (ou vários modelos) que parece ter alta qualidade do ponto de vista da análise de dados. Antes de prosseguir para a implantação final do modelo, é importante avaliá-lo mais detalhadamente e revisar as etapas executadas para garantir que ele atinja adequadamente os objetivos de negócios. No final desta fase, uma decisão sobre o uso dos resultados da mineração de dados deve ser alcançada.

Implantação/Implementação

Também chamada de desenvolvimento, essa fase busca colocar o modelo para funcionar. Ele coloca fim ao seu projeto, mas é necessário se lembrar de monitorar os resultados e de adaptar o modelo sempre que necessário. Os modelos que foram desenvolvidos, ajustados, validados e testados durante várias iterações são salvos e preparados para o ambiente de produção (o nome é estranho, mas esse é o ambiente em que o software está de fato funcionando).

É necessário criar um plano de implantação adequado que inclui detalhes sobre os requisitos de hardware e software. **O estágio de implantação também inclui a verificação e o monitoramento de aspectos para avaliar o modelo em produção quanto a resultados, desempenho e outras métricas.** Dependendo dos requisitos, a fase de implantação pode ser tão simples quanto gerar um relatório ou tão complexa quanto implementar um processo de mineração de dados repetível.

Em muitos casos, será o cliente, não o analista de dados, quem executará as etapas de implantação. **No entanto, mesmo que o analista não realize o esforço de implantação, é importante que o cliente entenda antecipadamente quais ações precisarão ser executadas para realmente fazer uso dos modelos criados.** *Entendido, galera?* Agora vamos fazer um resumão de tudo para ver se vocês realmente entenderam...

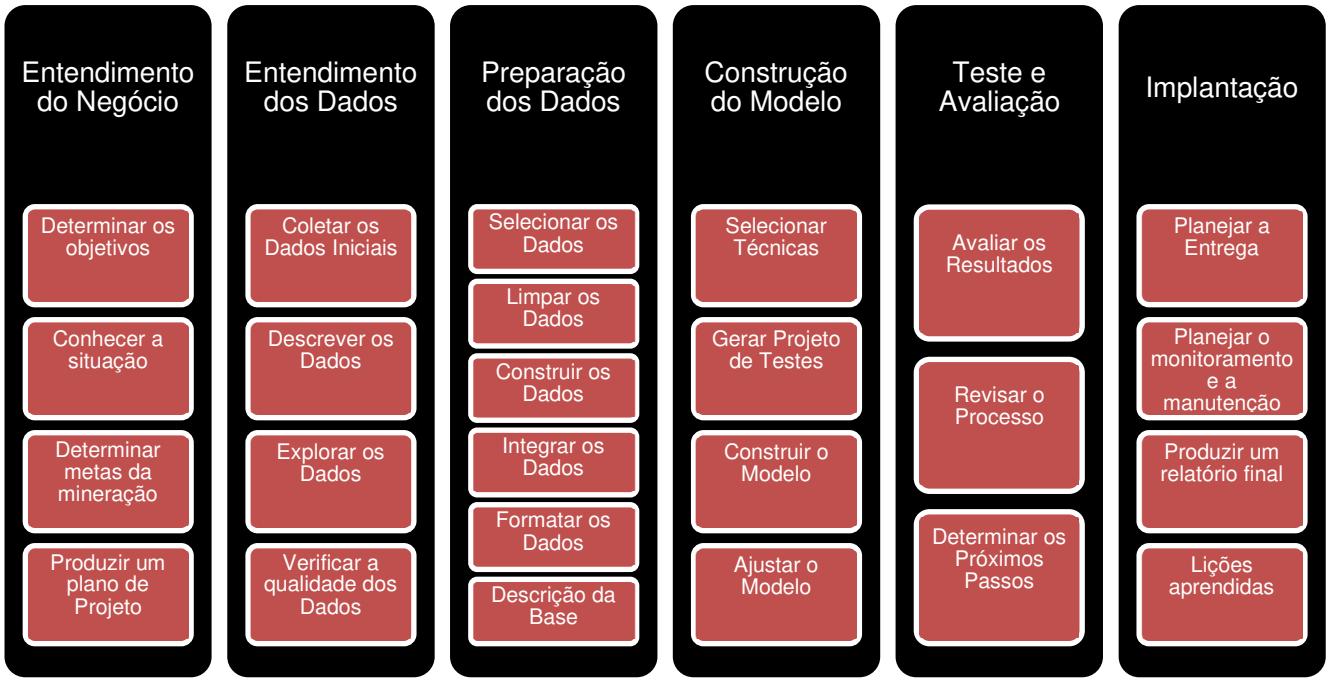
Acompanhem o Tio Diego aqui! Vamos supor que eu tenha sido chamado para fazer um projeto de mineração de dados da Bolsa de Valores! *Eu entendo algo desse negócio?* Não, então meu primeiro passo é entender o negócio. **Eu vou lá na Bolsa de Valores, entrevisto algumas pessoas, converso**

com outras até entender qual é o problema que se quer resolver, quais são os objetivos, os requisitos, entre outros. *Certinho até aqui?* Bacana...

Em seguida, eu vou lá ver como estão os dados que serão utilizados. Eu vejo que os dados vêm de dez sistemas diferentes, analiso como está a qualidade desses dados, qual é a quantidade, entre outras coisas. *Para quê?* **Para que eu me familiarize com os dados! Em seguida, eu vou pré-processar os dados.** Ora, tem dado corrompido, inconsistente, incompleto, faltante, etc – eu preciso fazer aquela limpeza básica para começar a trabalhar. Vamos para a modelagem...

Agora vou escolher ferramentas, técnicas e algoritmos que serão utilizados para modelar os meus dados. *Que técnicas eu posso utilizar?* Eu posso utilizar a classificação, a estimativa, a previsão, a análise de afinidades, a análise de agrupamentos, entre outras. *E que algoritmos?* Eu posso utilizar, por exemplo, árvores de decisão ou redes neurais. *E que ferramentas?* Eu posso utilizar *SAS Enterprise Miner* ou *IBM Intelligent Miner* ou *Oracle Darwin Data Mining Software*.

Em seguida, eu vou testar e avaliar os modelos desenvolvidos quanto à precisão e generalidade. *Foram atendidos os objetivos de negócio?* **Se sim, partimos para a fase de implantação, que é colocar o modelo para funcionar!** Segue abaixo uma imagem com as principais atividades de cada fase. Galera, esse não é um tema que cai muito em prova – na verdade, eu só encontrei cinco questões. Saibam disso para dosar bem os estudos de vocês. *Fechou?* ;)



(TCM/BA – 2015) Em um processo de mineração, durante a etapa de preparação dos dados, são analisados os requisitos de negócio para consolidar os dados.

_____

Comentários: há uma etapa específica responsável pela análise de requisitos de negócio chamada Entendimento do Negócio, que obtém conhecimentos sobre os objetivos do negócio e seus requisitos; a etapa de preparação de dados é responsável por limpar, transformar, integrar e formatar os dados selecionados (Errado).

(ME – 2020) 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.

_____

Comentários: a etapa de Modelagem ou Construção de Modelo CRISP-DM permite realmente a escolha e aplicação de diversas técnicas de mineração sobre os dados a serem analisados (Correto).

(INEP – 2012) Conforme o modelo CRISP-DM o ciclo de vida de um projeto de mineração de dados consiste de 6 (seis) fases que são:

a) Compreensão do Negócio, Compreensão dos Dados, Preparação dos Dados, Modelagem, Avaliação, e Desenvolvimento.

b) Preparação dos Dados, Modelagem, Avaliação, Requisitos, Escopo, Ambiente.

c) Requisitos, Escopo, Ambiente, Modelagem, Avaliação, e Desenvolvimento.

d) Compreensão do Negócio, Compreensão dos Dados, Preparação dos Dados, Requisitos, Escopo e Ambiente.

e) Requisitos, Escopo, Ambiente, Compreensão dos Dados, Preparação dos Dados e Modelagem.

_____

Comentários: as fases são: Compreensão do Negócio, Compreensão dos Dados, Preparação dos Dados, Modelagem, Avaliação e Desenvolvimento (Letra A).

RESUMO

DEFINIÇÕES DE DATA MINING

Data Mining é o processo de explorar grande quantidade de dados para extração não-trivial de informação implícita desconhecida.

Palavras-chave: exploração; informação implícita desconhecida.

Data Mining é uso de teorias, métodos, processos e tecnologias para organizar uma grande quantidade de dados brutos para identificar padrões de comportamentos em determinados públicos.

Palavras-chave: teorias; métodos; processos; tecnologias; organizar dados brutos; padrões de comportamentos.

Data Mining é a categoria de ferramentas de análise denominada open-end e que permite ao usuário avaliar tendências e padrões não conhecidos entre os dados.

Palavras-chave: ferramenta de análise; open-end; tendências e padrões.

Data Mining é o processo de descoberta de novas correlações, padrões e tendências entre as informações de uma empresa, por meio da análise de grandes quantidades de dados armazenados em bancos de dados usando técnicas de reconhecimento de padrões, estatísticas e matemáticas.

Palavras-chave: descoberta; correlações; padrões; tendências; reconhecimento de padrões; estatística; matemática.

Data Mining constitui em uma técnica para a exploração e análise de dados, visando descobrir padrões e regras, a princípio ocultos, importantes à aplicação.

Palavras-chave: exploração e análise de dados; padrões; regras; ocultos.

Data Mining é o conjunto de ferramentas que permitem ao usuário avaliar tendências e padrões não conhecidos entre os dados. Esses tipos de ferramentas podem utilizar técnicas avançadas de computação como redes neurais, algoritmos genéticos e lógica nebulosa (fuzzy), dentre outras.

Palavras-chave: tendências; padrões; redes neurais; algoritmos genéticos; lógica nebulosa.

Data Mining é o conjunto de ferramentas e técnicas de mineração de dados que têm por objetivo buscar a classificação e o agrupamento (clusterização) de dados, bem como identificar padrões.

Palavras-chave: classificação; agrupamento; clusterização; padrões.

Data Mining é o processo de explorar grandes quantidades de dados à procura de padrões consistentes com o intuito de detectar relacionamentos sistemáticos entre variáveis e novos subconjuntos de dados.

Palavras-chave: padrões; relacionamentos.

Data Mining consiste em explorar um conjunto de dados visando a extrair ou a ajudar a evidenciar padrões, como regras de associação ou sequências temporais, para detectar relacionamentos entre estes.

Palavras-chave: exploração; padrões; regras; associação; sequência temporal; detecção.

Data Mining são ferramentas que utilizam diversas técnicas de natureza estatística, como a análise de conglomerados (cluster analysis), que tem como objetivo agrupar, em diferentes conjuntos de dados, os elementos identificados como semelhantes entre si, com base nas características analisadas.

Palavras-chave: estatística; análise de conglomerados; agrupamento.

Data Mining é o conjunto de técnicas que, envolvendo métodos matemáticos e estatísticos, algoritmos e princípios de inteligência artificial, tem o objetivo de descobrir relacionamentos significativos entre dados armazenados em repositórios de grandes volumes e concluir sobre padrões de comportamento de clientes de uma organização.

Palavras-chave: métodos matemáticos e estatístico; inteligência artificial; relacionamentos; padrões; comportamentos.

Data Mining é o processo de explorar grandes quantidades de dados à procura de padrões consistentes, como regras de associação ou sequências temporais, para detectar relacionamentos sistemáticos entre variáveis, detectando assim novos subconjuntos de dados.

Palavras-chave: padrões; regras de associação; sequências temporais; relacionamentos.

Data Mining é o processo de identificar, em dados, padrões válidos, novos, potencialmente úteis e, ao final, compreensíveis.

Palavras-chave: padrões; utilidade.

Data Mining é um método computacional que permite extrair informações a partir de grande quantidade de dados.

Palavras-chave: extração.

Data Mining é o processo de explorar grandes quantidades de dados à procura de padrões consistentes, como regras de associação ou sequências temporais.

Palavras-chave: exploração; padrões consistentes; regras de associação; sequência temporal.

Data Mining é o processo de analisar de maneira semi-automática grandes bancos de dados para encontrar padrões úteis.

Palavras-chave: padrões.

CARACTERÍSTICAS DE MINERAÇÃO DE DADOS

Há diferentes tipos de mineração de dados: (1) diagnóstica, utilizada para entender os dados e/ou encontrar causas de problemas; (2) preditiva, utilizada para antecipar comportamentos futuros.

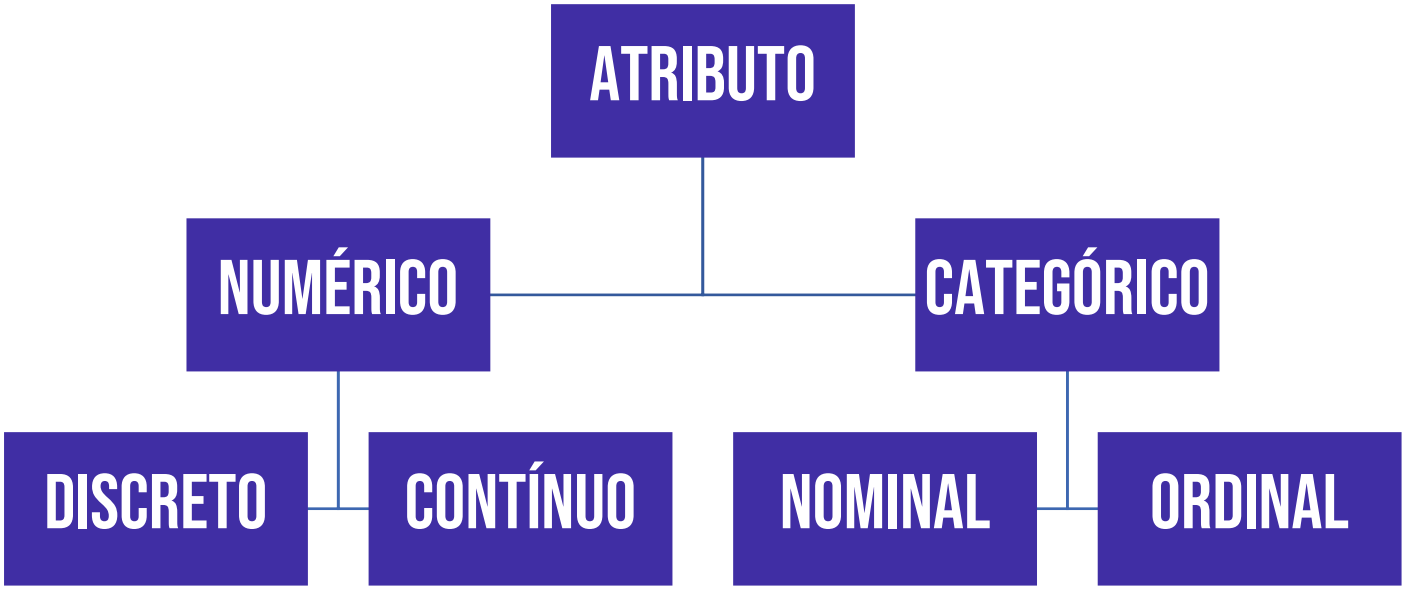
As provas vão insistir em afirmar que a mineração de dados só pode ocorrer em bancos de dados muito grades como Data Warehouses, mas isso é falso – apesar de comum, não é obrigatório.

Em geral, ferramentas de mineração de dados utilizam uma arquitetura web cliente/servidor, sendo possível realizar inclusive a mineração de dados de bases de dados não estruturadas.

Não é necessário ter conhecimentos de programação para realizar consultas, visto que existem ferramentas especializadas que auxiliam o usuário final de negócio.

TIPOS DE DADOS	DESCRIÇÃO
DADOS ESTRUTURADOS	São aqueles que residem em campos fixos de um arquivo (Ex: tabela, planilha ou banco de dados) e que dependem da criação de um modelo de dados, isto é, uma descrição dos objetos juntamente com as suas propriedades e relações.
DADOS SEMIESTRUTURADOS	São aqueles que não possuem uma estrutura completa de um modelo de dados, mas também não é totalmente desestruturado. Em geral, são utilizados marcadores (*tags*) para identificar certos elementos dos dados, mas a estrutura não é rígida.
DADOS NÃO ESTRUTURADOS	São aqueles que não possuem um modelo de dados, que não está organizado de uma maneira predefinida ou que não reside em locais definidos. Eles costumam ser de difícil indexação, acesso e análise (Ex: imagens, vídeos, sons, textos livres, etc).

ATRIBUTO DEPENDENTE	Representa um atributo de saída que desejamos manipular em um experimento de dados (também chamado de variável alvo ou variável target).
ATRIBUTO INDEPENDENTE	Representa um atributo de entrada que desejamos registrar ou medir em um experimento de dados.



TIPOS DE ATRIBUTOS	DESCRIÇÃO
ATRIBUTO NUMÉRICO	Também chamado de atributo quantitativo, é aquele que pode ser medido em uma escala quantitativa, ou seja, apresenta valores numéricos que fazem sentido.
DISCRETO	Os valores representam um conjunto finito ou enumerável de números, e que resultam de uma contagem (Ex: número de filhos, número de bactérias por amostra, número de logins em uma página web, entre outros).
CONTÍNUO	Os valores pertencem a um intervalo de números reais e representam uma mensuração (Ex: altura de uma pessoa, peso de uma marmita, salário de um servidor público, entre outros).

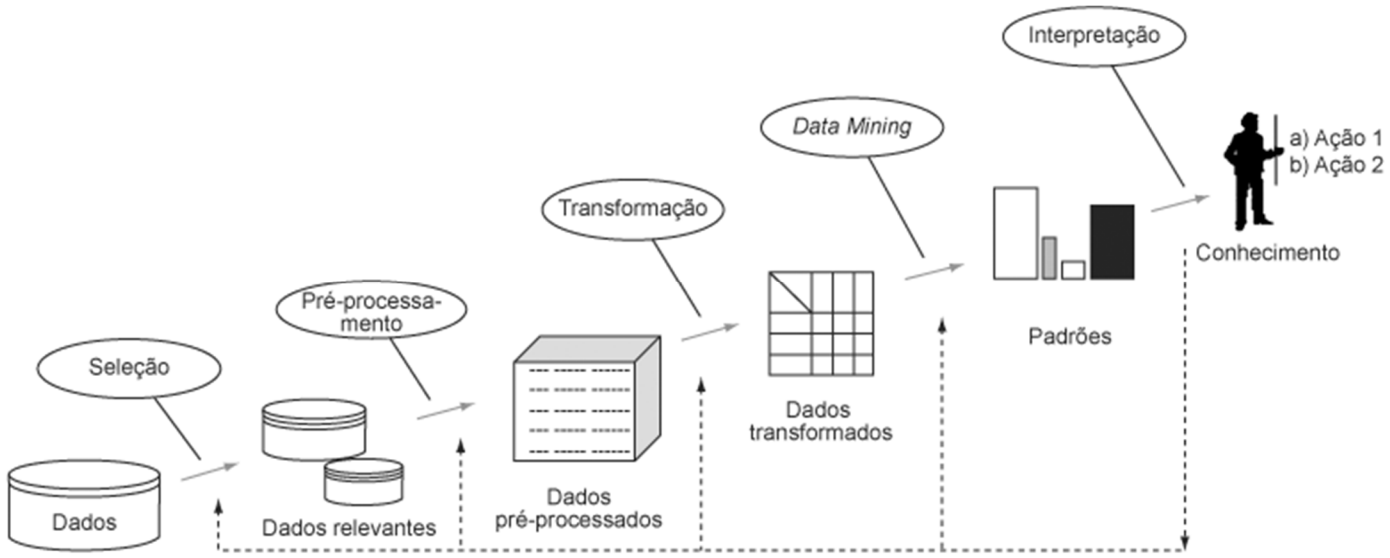
TIPOS DE ATRIBUTOS	DESCRIÇÃO
ATRIBUTO CATEGÓRICO	Também chamado de atributo qualitativo, é aquele que pode assumir valores categóricos, isto é, representam uma classificação.
NOMINAL	São aquelas em que não existe uma ordenação própria entre as categorias (Ex: sexo, cor dos olhos, fumante/não fumante, país de origem, profissão, religião, raça, time de futebol, entre outros).
ORDINAL	São aquelas em que existe uma ordenação própria entre as categorias (Ex: Escolaridade (1º, 2º, 3º Graus), Estágio de Doença (Inicial, Intermediário, Terminal), Classe Social (Classe Baixa, Classe Média, Classe Alta), entre outros)

OBJETIVOS DA MINERAÇÃO	DESCRIÇÃO

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: prever um terremoto com alta probabilidade).
IDENTIFICAÇÃO	Padrões de dados podem ser usados para identificar a existência de um item, um evento ou uma atividade (Ex: padrões de comportamento de hackers permitem identificar possíveis intrusos acessando sistema).
CLASSIFICAÇÃO	A mineração de dados permite particionar os dados de modo que diferentes classes ou categorias possam ser identificadas com base em combinações de parâmetros (Ex: clientes podem ser categorizados pelos seus perfis de compradores).
OTIMIZAÇÃO	A mineração de dados pode otimizar o uso de recursos limitados, como tempo, espaço, dinheiro ou materiais e maximizar variáveis de saída como vendas ou lucros sob determinadas restrições (Ex: tempo, escopo e custo de um projeto).

PROCESSO DE DESCOBERTA DE CONHECIMENTO

Data Mining faz parte de um processo muito maior de descoberta de conhecimento chamada KDD (Knowledge Discovery in Databases – Descoberta de Conhecimento em Bancos de Dados).

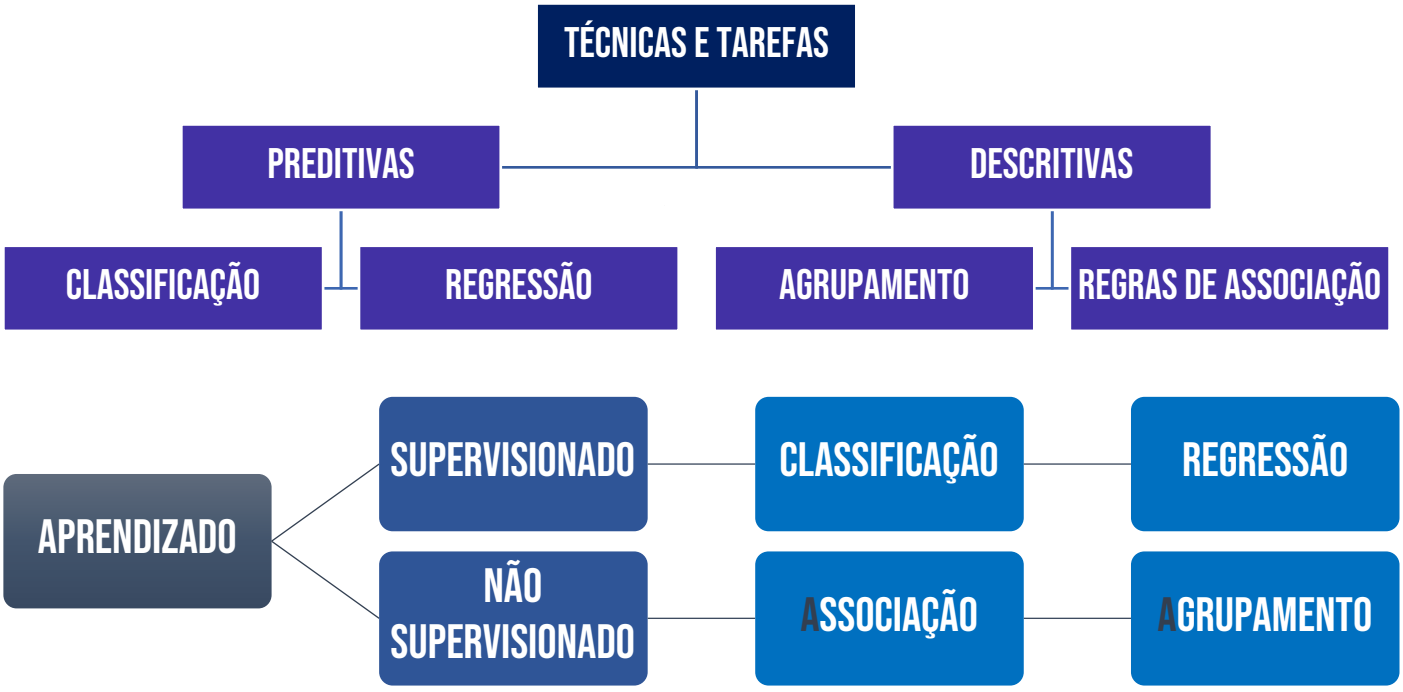


1	**SELEÇÃO DE DADOS**	Dados sobre itens ou categorias são selecionados.
2	**LIMPEZA DE DADOS**	Dados são corrigidos ou eliminados dados incorretos
3	**ENRIQUECIMENTO DE DADOS**	Dados são melhorados com fontes de informações adicionais
4	**TRANSFORMAÇÃO DE DADOS**	Dados são reduzidos por meio de sumarizações, agregações e discretizações[1].
5	**MINERAÇÃO DE DADOS**	Padrões úteis são descobertos.
6	**EXIBIÇÃO DE DADOS**	Informações descobertas são exibidas ou relatórios são construídos.

[1] Variáveis numéricas são convertidas em classes ou categorias.

PRÉ-PROCESSAMENTO DE DADOS

Trata-se do processo de preparação de dados para análise e modelagem adicionais. Envolve a transformação de dados brutos em um formato mais adequado para algoritmos de aprendizado de máquina por meio de tarefas como limpeza, normalização e organização dos dados para que possam ser analisados mais facilmente e usados para fazer previsões. É também uma etapa essencial em qualquer projeto de aprendizado de máquina, pois garante que os dados estejam em um formato adequado para o processo de modelagem.

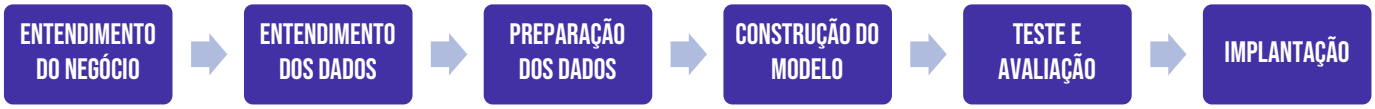


TÉCNICAS	DESCRIÇÃO
CLASSIFICAÇÃO	Hierarquia de classes com base em um conjunto existente de eventos ou transações.
REGRESSÃO	Aplicação especial da regra de classificação em busca de uma função que mapeie registros de um banco de dados.
REGRAS DE ASSOCIAÇÃO	Busca descobrir relacionamentos entre variáveis correlacionando a presença de um item com uma faixa de valores para outro conjunto de variáveis
AGRUPAMENTO	Particiona dados em segmentos previamente desconhecidos com características semelhantes.

MEDIDAS DE INTERESSE	DESCRIÇÃO
SUPORTE/ PREVALÊNCIA	Trata-se da frequência com que um conjunto de itens específicos ocorrem no banco de dados, isto é, o percentual de transações que contém todos os itens em um conjunto. Em termos matemáticos, a medida de suporte para uma regra X → Y é a frequência em que o conjunto de itens aparece nas transações do banco de dados. Um suporte alto nos leva a crer que os itens do conjunto X e Y costumam ser comprados juntos, pois ocorrem com alta frequência no banco (Ex: 70% das compras realizadas em um mercado contém arroz e refrigerante).

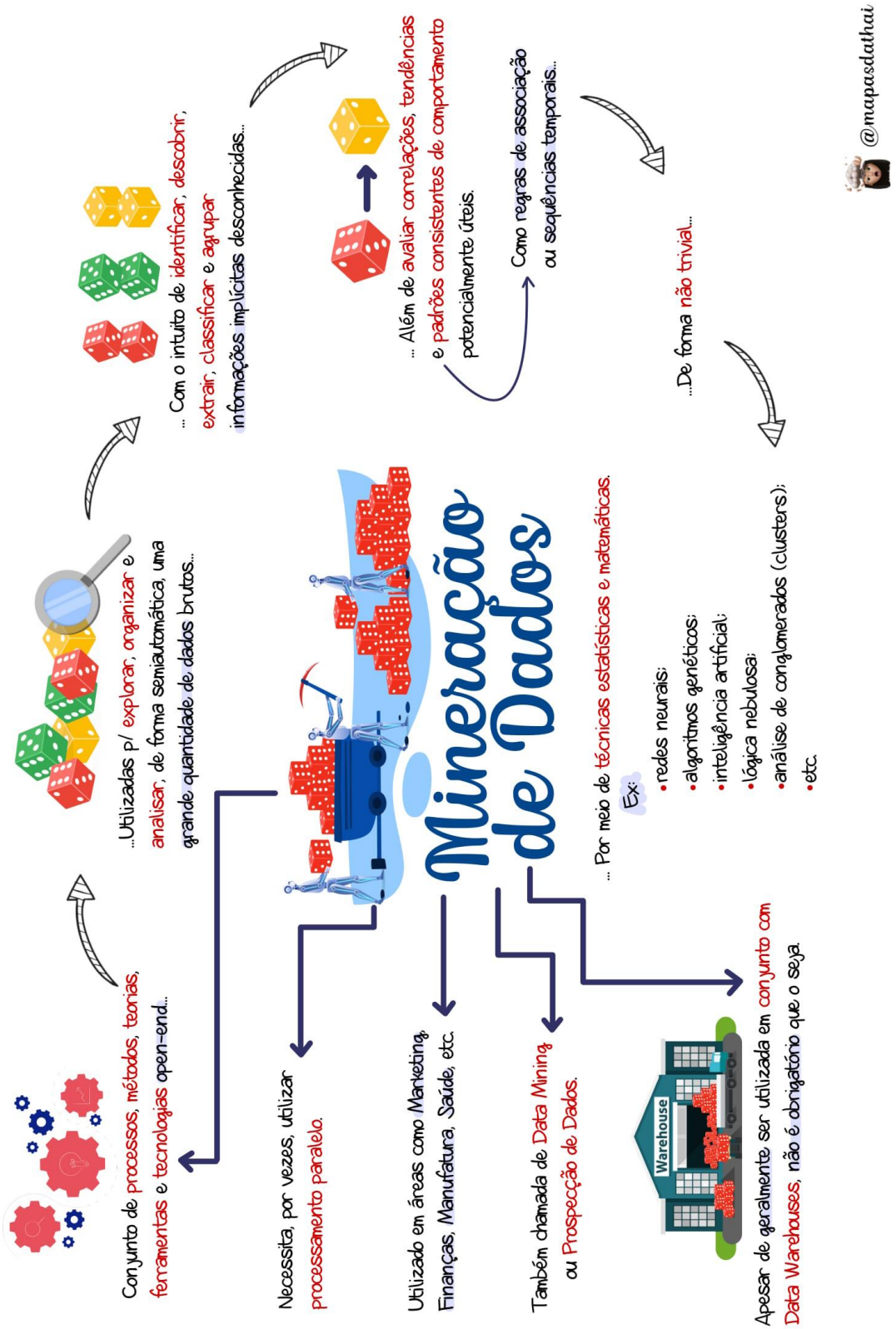
CONFIANÇA/ FORÇA	Trata-se da probabilidade de que exista uma relação entre itens. Em termos matemáticos, a medida de confiança para uma regra X → Y é a força com que essa regra funciona. Ela é calculada pela frequência dos itens Y serem comprados dado que os itens X foram comprados. Uma confiança alta nos leva a crer que exista uma alta probabilidade de que se X for comprado, Y também será (Ex: existe uma probabilidade de 70% de que clientes que compram fraldas também copprem cerveja).

CONCEITOS AVANÇADOS	DESCRIÇÃO
APRENDIZADO DE MÁQUINA	Trata-se de uma ferramenta poderosa para a aquisição automática de conhecimento por meio da imitação do comportamento de aprendizagem humano com foco em aprender a reconhecer padrões complexos e tomar decisões.
MINERAÇÃO DE TEXTO	Trata-se de um meio para encontrar padrões interessantes/úteis em um contexto de informações textuais não estruturadas, combinado com alguma tecnologia de extração e de recuperação da informação, processo de linguagem natural e de sumarização ou indexação de documentos.



CRISP-DM	DESCRIÇÃO
ENTENDIMENTO DO NEGÓCIO	Busca compreender das necessidades gerenciais e dos objetivos e requisitos de negócio que devem ser atendidos pela mineração de dados.
ENTENDIMENTO DOS DADOS	Busca identificar os dados relevantes das diferentes fontes de dados.
PREPARAÇÃO DOS DADOS	Busca carregar os dados identificados no passo anterior e prepará-los para análise por métodos de mineração de dados.
CONSTRUÇÃO DO MODELO	Busca selecionar e aplicar técnicas de modelagem a um conjunto de dados previamente preparado.
TESTE E AVALIAÇÃO	Busca testar e avaliar os modelos desenvolvidos.
IMPLANTAÇÃO	Busca organizar o conhecimento adquirido com a exploração dos dados de forma que o usuário possa compreendê-lo.

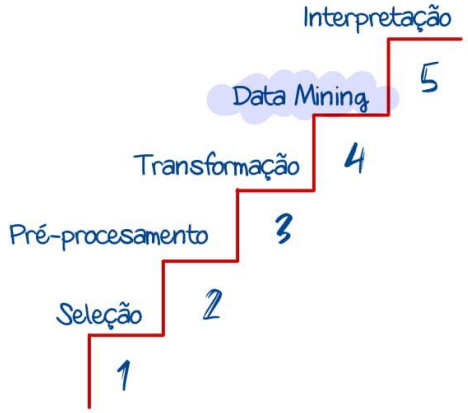
MAPA MENTAL


Mineração de Dados
Conjunto de processos, métodos, teorias, ferramentas e tecnologias open-end...
...Utilizadas p/ explorar, organizar e analisar, de forma semiautomática, uma grande quantidade de dados brutos...
... Com o intuito de identificar, descobrir, extrair, classificar e agrupar informações implícitas desconhecidas...
... Além de avaliar correlações, tendências e padrões consistentes de comportamento potencialmente úteis.
Como regras de associação ou sequências temporais...
...De forma não trivial...
@mapasdathai
... Por meio de técnicas estatísticas e matemáticas.
Ex:
•redes neurais;
•algoritmos genéticos;
•inteligência artificial;
•lógica nebulosa;
•análise de conglomerados (clusters);
•etc.
Apesar de geralmente ser utilizada em conjunto com Data Warehouses, não é obrigatório que o seja.
Warehouse
Também chamada de Data Mining ou Prospecção de Dados.
Utilizado em áreas como Marketing, Finanças, Manufatura, Saúde, etc.
Necessita, por vezes, utilizar processamento paralelo.

Mineração de Dados

Processo de Descoberta de Conhecimento

- Knowledge Discovery in Databases (KDD) - Descoberta de Conhecimento em Banco de Dados.
- A mineração de dados é uma das fases do KDD.



- **Seleção**: selecionar um conjunto de dados ou se concentrar em um subconjunto de variáveis ou amostras.
- **Limpeza e Pré-processamento**: remoção de erros, coleta de informações, etc.
- **Transformação**: os dados são transformados e consolidados em formas apropriadas à mineração (sumarizando-os ou agregando-os).
- **Mineração de dados**: algoritmos e técnicas p/ extrair possíveis padrões úteis de dados.
- **Interpretação**: os padrões encontrados são avaliados e interpretados.

Objetivos

PREVISÃO

- Prever comportamentos futuros com base em comportamentos passados.

IDENTIFICAÇÃO

- Identificar, através de padrões de dados, a existência de um item, um evento ou uma atividade.

CLASSIFICAÇÃO:

- Particionar os dados p/ que diferentes categorias possam ser identificadas com base em combinações de parâmetros.

OTIMIZAÇÃO:

- Otimizar o uso de recursos limitados (como tempo, espaço, dinheiro ou materiais) e maximizar variáveis de saída (como vendas e lucros), sob determinado conjunto de restrições.

@mapasdathai

Mineração de Dados
Técnicas e Tarefas

Predição
Ou Previsão

Busca descrever a natureza de ocorrências futuras de certos eventos com base nos acontecimentos passados.

- **Classificação**
  - Categorização de dados p/ a identificação de uma classe por meio de múltiplos atributos.
    - As classes são predefinidas antes da análise dos resultados (aprendizado supervisionado).
  - PRINCIPAIS FERRAMENTAS DE CLASSIFICAÇÃO:
    - Árvores de Decisão
      - Representação gráfica das regras de classificação, demonstrando visualmente as condições e probabilidades p/ se chegar a algum resultado.
    - Redes Neurais (Artificiais)
      - Algoritmos inspirados no cérebro de animais, com capacidade de aprender com experiências passadas, prever um comportamento ou reconhecer um padrão, desde que apresentados e um conjunto de dados estruturados.
- **Regressão**
  - Objetiva prever um número/valor.
    - Compreende a busca por uma função que mapeie os registros de um BD em um intervalo de valores reais.

Clusterização
Ou Agrupamento

Particiona dados em segmentos previamente desconhecidos com características semelhantes.

Associação

Busca descobrir relacionamentos entre variáveis correlacionando a presença de um item com uma faixa de valores p/ outro conjunto de variáveis.

- REGRAS DE ASSOCIAÇÃO: X -> Y
  - Visam descobrir o relacionamento entre variáveis de um banco de dados.
- PADRÕES SEQUENCIAIS
  - Buscam descobrir padrões sequencias de ações ou eventos de forma equivalente a certos relacionamentos temporais.
- PADRÕES TEMPORAIS
  - Semelhante à técnica dos Padrões Sequenciais, mas sempre envolve um fator temporal que permite diferenciá-los.
    - um intervalo regular de tempo

Uma regra de associação deve satisfazer alguma medida de interesse do analista de dados:

- SUPORTE/PREVALÊNCIA: frequência com que um conjunto de itens ocorre no BD.
- CONFIANÇA/FORÇA: probabilidade de que exista uma relação entre itens.


@mapasdathai

QUESTÕES COMENTADAS – CESPE

1. **(CESPE / MPO – 2024)** A regressão tem como objetivo a obtenção de uma equação que relacione uma variável de resposta a uma ou mais variáveis explicativas.

Comentários:

A regressão é uma técnica de aprendizado supervisionado que visa encontrar uma equação matemática que descreva a relação entre uma variável dependente (resposta) e uma ou mais variáveis independentes (explicativas). O objetivo é prever a variável de resposta com base nas variáveis explicativas. Exemplos incluem prever preços de imóveis com base em características como tamanho e localização, e prever a demanda por um produto com base em variáveis econômicas.

Gabarito: Correto

---

2. **(CESPE / ANTT – 2024)** Os algoritmos de regras de associação constroem regras com apenas uma única conclusão, ao contrário dos algoritmos de árvore de decisão, que tentam localizar muitas regras, cada uma delas com uma conclusão diferente.

Comentários:

Tanto algoritmos de regras de associação quanto árvores de decisão podem gerar múltiplas regras. A diferença está no propósito e na forma como essas regras são utilizadas. As regras de associação geram regras com uma única consequência por regra, mas um conjunto de dados pode resultar em muitas regras. As árvores de decisão também geram múltiplas regras, cada uma levando a uma conclusão de classificação ou previsão diferente.

Gabarito: Errado

---

3. **(CESPE / CTI – 2024)** Clustering é uma técnica de mineração de dados que agrupa dados não rotulados com base em suas semelhanças ou diferenças; os algoritmos de cluster podem ser categorizados em sobrepostos, hierárquicos ou probabilísticos.

Comentários:

Clustering é uma técnica de mineração de dados que agrupa dados não rotulados com base em suas semelhanças ou diferenças. Os algoritmos de clustering podem ser categorizados como sobrepostos (onde um dado pode pertencer a mais de um cluster), hierárquicos (que constroem uma hierarquia de clusters) e probabilísticos (que utilizam modelos estatísticos para agrupar dados com base em probabilidades).

Gabarito: Correto

---

4. **(CESPE / DATAPREV – 2023)** As técnicas de regressão utilizam um conjunto finito de hipóteses para, a partir dos atributos previsores, determinar a categoria de um objeto do conjunto de dados analisado.

Comentários:

As técnicas de regressão são utilizadas para prever valores contínuos, não para determinar a categoria de um objeto. A descrição fornecida se refere a técnicas de classificação, que utilizam um conjunto finito de hipóteses para categorizar objetos com base em seus atributos previsores. Regressão e classificação são duas abordagens distintas de aprendizado supervisionado.

Gabarito: Errado

---

5. **(CESPE / DATAPREV – 2023)** A regra de associação é uma técnica que busca relações de co-ocorrência entre objetos de uma base de dados.

Comentários:

A regra de associação é uma técnica que busca identificar relações de co-ocorrência entre objetos em uma base de dados. Ela é amplamente utilizada em mineração de dados para descobrir padrões interessantes, como produtos frequentemente comprados juntos em transações de mercado. Um exemplo clássico é a análise de cestas de compras.

Gabarito: Correto

---

6. **(CESPE / AGER-MT – 2023)** Em machine learning, quando algoritmos de aprendizado de máquina são usados para analisar e agrupar conjuntos de dados não rotulados, de forma tal que os algoritmos descobrem padrões ocultos sem a necessidade de intervenção humana, usa-se a forma de aprendizado do tipo:

a) não supervisionado.
b) supervisionado.
c) over fitting.
d) under fitting.
e) classificação.

Comentários:

(a) Correto. No aprendizado não supervisionado, os algoritmos analisam e agrupam conjuntos de dados não rotulados, descobrindo padrões ocultos sem a necessidade de intervenção humana;

(b) Errado. O aprendizado supervisionado envolve dados rotulados, onde o algoritmo é treinado com exemplos de entrada e saída para prever resultados em novos dados;

(c) Errado. Overfitting é um problema em machine learning onde o modelo se ajusta muito bem aos dados de treinamento, mas não generaliza bem para novos dados – não é um tipo de aprendizado;

(d) Errado. Underfitting ocorre quando um modelo é muito simples e não captura a complexidade dos dados – também não é um tipo de aprendizado;

(e) Errado. Classificação é uma tarefa dentro do aprendizado supervisionado, onde o objetivo é categorizar dados em classes predefinidas, não um tipo de aprendizado.

Gabarito: Letra A

---

7. **(CESPE / SEFIN de Fortaleza-CE – 2023)** Aprendizado de máquina é um subcampo da inteligência artificial que consiste no treinamento de modelos computacionais para que possam reconhecer padrões e, a partir de um conjunto de dados de entrada, prever o valor de uma variável de saída. Em relação ao aprendizado de máquina, julgue o item a seguir.

   Em aprendizado de máquina, as características de entrada e saída são definidas, respectivamente, como atributos previsores e atributos alvo ou meta.

Comentários:

Em aprendizado de máquina, as características de entrada são chamadas de atributos previsores ou features, pois são usadas para fazer previsões. A variável de saída é conhecida como atributo alvo ou meta (target), pois é o valor que o modelo tenta prever com base nos atributos previsores. Esse processo de mapeamento de atributos de entrada para a saída é fundamental para o treinamento de modelos de aprendizado supervisionado.

Gabarito: Correto

---

8. **(CESPE / SEFIN de Fortaleza-CE – 2023)** Nos algoritmos de aprendizado por reforço, o agente recebe uma recompensa atrasada na próxima etapa de tempo para avaliar sua ação anterior; seu objetivo, então, é maximizar a recompensa.

Comentários:

No aprendizado por reforço, o agente interage com o ambiente e, com base em suas ações, recebe recompensas (que podem ser atrasadas) que indicam o quão boa foi a ação tomada. O objetivo do agente é aprender uma política de ações que maximize a soma das recompensas ao longo do tempo, mesmo que algumas recompensas sejam recebidas em etapas futuras.

Gabarito: Correto

---

9. **(CESPE / DATAPREV – 2023)** Um sistema de aprendizado não supervisionado, dotado de um conjunto de dados de treinamento que foram classificados manualmente, tenta aprender, a partir desses dados de treinamento, uma forma de classificá-los, bem como de classificar novos dados, ainda não observados.

Comentários:

Em um sistema de aprendizado não supervisionado, o modelo é treinado com dados que não foram classificados manualmente. O objetivo do aprendizado não supervisionado é identificar padrões, agrupamentos ou relações inerentes nos dados sem ter rótulos ou classificações pré-definidos. Por outro lado, quando se utiliza um conjunto de dados de treinamento classificados manualmente para aprender e classificar novos dados, estamos lidando com aprendizado supervisionado. Esse é o método utilizado para prever ou classificar novos dados com base em exemplos de treinamento.

Gabarito: Errado

---

10. **(CESPE / CNMP - 2023)** O *data mining* é um processo usado para extrair e analisar informações que revelam padrões ou tendências estratégicas do negócio.

Comentários:

Perfeito! Data Mining é um processo usado para extrair e analisar informações que revelam padrões ou tendências estratégicas do negócio, fornecendo insights valiosos para auxiliar na tomada de decisões e no desenvolvimento de estratégias empresariais.

Gabarito: Correto

---

11. **(CESPE / TRT8 – 2022)** Acerca de modelos preditivos e descritivos, assinale a opção correta:

a) Com um modelo não supervisionado consegue-se construir um estimador a partir de exemplos rotulados.

b) Um modelo supervisionado refere-se à identificação de informações relevantes nos dados sem a presença de um elemento externo para orientar o aprendizado.

c) Com o uso de técnicas do modelo não supervisionado, consegue-se prever com exatidão o resultado de uma eleição utilizando pesquisas como parâmetro.

d) A análise de agrupamento pertence ao paradigma de aprendizado não supervisionado, em que o aprendizado é dirigido aos dados, não requerendo conhecimento prévio sobre as suas classes ou categorias.

e) Tendo como objetivo encontrar padrões ou tendências para auxiliar o entendimento dos dados, deve-se usar técnicas do modelo supervisionado.

Comentários:

(a) Errado, os modelos não supervisionados não utilizam exemplos rotulados; (b) Errado, esse seria o modelo não supervisionado; (c) Errado, modelos não preveem dados com exatidão; (d) Correto, a análise de agrupamento realmente utiliza o paradigma não supervisionado, isto é, aquele que não necessita de conhecimento prévio de classes, categorias ou rótulos; (e) Errado, é possível utilizar técnicas do modelo não supervisionado ou do modelo por esforço.

Gabarito: Letra D

---

12. **(CESPE / ISS-Aracaju – 2021)** Em um projeto de data mining, a coleta do dado que será garimpado ocorre no processo de:

a) mineração.
b) preparação.
c) aplicação.
d) associação.
e) classificação.

Comentários:

A etapa de preparação de dados é responsável por carregar os dados identificados e prepará-los para análise por métodos de mineração de dados.

Gabarito: Letra B

---

13. **(CESPE / ISS-Aracaju – 2021)** De acordo com o modelo CRSP-DM, a seleção das técnicas que serão aplicadas nos dados selecionados ocorre na fase de:

a) modelagem.
b) entendimento dos dados.
c) entendimento do negócio.
d) avaliação.
e) preparação dos dados.

Comentários:

CRISP-DM	DESCRIÇÃO
ENTENDIMENTO DO NEGÓCIO	Busca compreender das necessidades gerenciais e dos objetivos e requisitos de negócio que devem ser atendidos pela mineração de dados.

ENTENDIMENTO DOS DADOS	Busca identificar os dados relevantes das diferentes fontes de dados.
PREPARAÇÃO DOS DADOS	Busca carregar os dados identificados no passo anterior e prepará-los para análise por métodos de mineração de dados.
CONSTRUÇÃO DO MODELO	Busca selecionar e aplicar técnicas de modelagem a um conjunto de dados previamente preparado.
TESTE E AVALIAÇÃO	Busca testar e avaliar os modelos desenvolvidos.
IMPLANTAÇÃO	Busca organizar o conhecimento adquirido com a exploração dos dados de forma que o usuário possa compreendê-lo.

A etapa de Construção do Modelo (também chamada de Modelagem) é responsável por selecionar e aplicar técnicas de modelagem a um conjunto de dados previamente preparado. Detalhe: mais uma questão desleixada, visto que o nome do modelo é CRISP-DM e, não, CRSP-DM!

Gabarito: Letra A

---

14. (CESPE / ISS-Aracaju – 2021) O enriquecimento de dados da etapa de pré-processamento e preparação do data mining tem como objetivo:

a) a deduplicidade de registros.
b) a seleção de amostras.
c) a integração de bases diferentes.
d) o tratamento de valores nulos.
e) o acréscimo de dados à base já existentes.

Comentários:

De acordo com Navathe, as etapas de Descoberta do Conhecimento (KDD) são compostas por seis etapas, sendo que as quatro primeiras são agrupadas em uma etapa de pré-processamento.

1	SELEÇÃO DE DADOS	Dados sobre itens ou categorias são selecionados.
2	LIMPEZA DE DADOS	Dados são corrigidos ou eliminados dados incorretos.
3	ENRIQUECIMENTO DE DADOS	Dados são melhorados com fontes de informações adicionais.
4	TRANSFORMAÇÃO DE DADOS	Dados são reduzidos por meio de sumarizações, agregações e discretizações.
5	MINERAÇÃO DE DADOS	Padrões úteis são descobertos.
6	EXIBIÇÃO DE DADOS	Informações descobertas são exibidas ou relatórios são construídos.

Dito isso, o enriquecimento de dados tem o objetivo de melhorar os dados com fontes de informações adicionais, isto é, acrescer dados à base de dados já existente.

Gabarito: Letra E

---

15. (CESPE / PCDF – 2021) Uma das aplicações de Python é o aprendizado de máquina, que pode ser exemplificado por um programa de computador que aprende com a experiência de detectar imagens de armas e de explosivos em vídeos, tendo seu desempenho medido e melhorando por meio dos erros e de acertos decorrentes da experiência de detecção.

Comentários:

Nada melhor do que a justificativa da própria banca: *"O exemplo apresentado enquadra-se na definição atual de aprendizado de máquina. O que é aprendizado de máquina? Duas definições de aprendizado de máquina são oferecidas. Arthur Samuel o descreveu como: "o campo de estudo que dá aos computadores a capacidade de aprender sem serem explicitamente programados. Essa é uma definição mais antiga e informal. Na definição mais moderna, "um programa de computador aprende com a experiência E com relação a alguma classe de tarefas T e medida de desempenho P, se seu desempenho nas tarefas em T, conforme medido por P, melhora com a experiência E. Por exemplo, jogar damas. E = a experiência de jogar muitos jogos de damas; T = a tarefa de jogar damas. P = a probabilidade de o programa vencer o próximo jogo".*

Perfeito! Python é realmente muito utilizado com Aprendizado de Máquina (Machine Learning). O exemplo também está correto: à medida que a máquina vai aprendendo a identificar armas e explosivos em vídeos, seu desempenho vai sendo medido e melhorado.

Gabarito: Correto

---

16. (CESPE / PCDF – 2021) A detecção de novos tipos de fraudes é uma das aplicações comuns da técnica de modelagem descritiva da mineração de dados, a que viabiliza o mapeamento rápido e preciso de novos tipos de golpes por meio de modelos de classificação de padrões predefinidos de fraudes.

Comentários:

Nada melhor do que a justificativa da própria banca: *"A detecção de fraudes é uma das aplicações da técnica "detecção de anomalias" da mineração de dados.*

Detecção de anomalia

A detecção de anomalias pode ser vista como o outro lado do cluster — ou seja, encontrar instâncias de dados que são incomuns e não se enquadram em nenhum padrão estabelecido. A detecção de fraude é um exemplo de detecção de anomalias. Embora a detecção de fraude possa ser vista como um problema para a modelagem preditiva, a relativa raridade de transações fraudulentas e a velocidade com que os criminosos desenvolvem novos tipos de fraude significam que qualquer modelo preditivo provavelmente terá baixa precisão e se tornará rapidamente desatualizado. Assim, a detecção de anomalias se concentra em modelar o que é um comportamento normal para identificar transações incomuns.

Modelagem descritiva

A modelagem descritiva, ou clustering, também divide os dados em grupos. Com o agrupamento, no entanto, os grupos apropriados não são conhecidos com antecedência; os padrões descobertos pela análise dos dados são usados para determinar os grupos".

De fato, a questão trata de detecção de anomalia e, não, modelagem descritiva. O algoritmo identifica que determinado comportamento não se encaixa em nenhum padrão pré-estabelecido.

Gabarito: Errado

---

17. (CESPE / APEX – 2021) m data mining revela informações que consultas manuais não poderiam revelar efetivamente. Por exemplo, em data mining, o algoritmo de classificação permite:

a) dividir o banco de dados em segmentos cujos membros compartilhem características iguais por meio de redes neurais.

b) analisar os dados históricos armazenados em um banco de dados e gerar automaticamente um modelo que possa prever comportamentos futuros.

c) mapear dados por meio da técnica estatística e, assim, obter um valor de previsão a partir de técnicas de regressão linear e não linear.

d) estabelecer relações entre itens que estejam juntos em determinado registro, o que é conhecido como análise de cesta de compras.

Comentários:

(a) Errado. Na classificação, os dados já são previamente segmentados/rotulados – esse item trata de agrupamento; (b) Correto, a classificação nos leva à previsão; (c) Errado, esse item trata da regressão; (d) Errado, esse item trata da associação.

Gabarito: Letra B

---

18. (CESPE / TCE-RJ – 2021) 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.

Comentários:

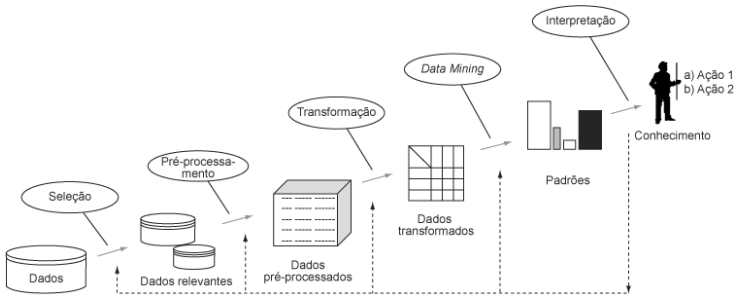
Perfeito! As fases do CRISP-DM são: (1) Entendimento do Negócio; (2) Entendimento dos Dados; (3) Preparação dos Dados; (4) Construção do Modelo; (5) Teste e Avaliação; e (6) Implantação/Implementação. Note que a implantação realmente deve ocorrer após a avaliação.

Gabarito: Correto

19.(CESPE / TCE-RJ – 2021) A descoberta de conhecimento em bases de dados, ou KDD (Knowledge-Discovery), é a etapa principal do processo de mineração de dados.

Comentários:

Bastava lembrar da nossa figurinha de KDD! A Mineração de Dados (Data Mining) é a etapa principal do processo de descoberta de conhecimento em base de dados e, não, o inverso.



Gabarito: Errado

20.(CESPE / TCE-RJ – 2021) Na mineração de dados preditiva, ocorre a geração de um conhecimento obtido de experiências anteriores para ser aplicado em situações futuras.

Comentários:

A Análise Preditiva combina técnicas de estatística, mineração de dados e aprendizagem de máquina (Machine Learning) para encontrar significado em grandes quantidades de dados, trabalhando com probabilidades, predições, entre outros para antecipar comportamentos futuros com base em eventos passados. Responde à pergunta: "O que vai acontecer?", portanto o conhecimento obtido de experiências passadas poderá ser aplicado em situações futuras.

Gabarito: Correto

21.(CESPE / TCE-RJ – 2021) As regras de associação adotadas em mineração de dados buscam padrões frequentes entre 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 potenciais produtos de interesse.

Comentários:

Na Mineração de Dados, uma regra de associação relaciona a presença de um conjunto de itens com outra faixa de valores de um outro conjunto de variáveis. Em outras palavras, essa técnica

permite buscar padrões frequentes para identificar hábitos de consumo – também conhecido como Análise de Cesta de Compras. Estudam-se os padrões e preferências de compra do consumidor por meio de coletas e análises de dados.

Gabarito: Correto

---

22. (CESPE / TCE-RJ – 2021) 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.

Comentários:

Opa... as fases do CRISP-DM são: (1) Entendimento do Negócio; (2) Entendimento dos Dados; (3) Preparação dos Dados; (4) Construção do Modelo; (5) Teste e Avaliação; e (6) Implantação/Implementação. Logo, a primeira fase é o entendimento do negócio e, não, o entendimento dos dados.

Gabarito: Errado

---

23. CESPE / TCE-RJ – 2021) No método de classificação para mineração de dados, a filiação dos objetos é obtida por meio de um processo não supervisionado de aprendizado, em que somente as variáveis de entrada são apresentadas para o algoritmo.

Comentários:

O método de classificação é supervisionado, logo as variáveis de saída são conhecidas de antemão, portanto também são apresentadas para o algoritmo.

Gabarito: Errado

---

24. (CESPE / TCE-RJ – 2021) No método de mineração de dados por agrupamento (clustering), são utilizados algoritmos com heurísticas para fins de descoberta de agregações naturais entre objetos.

Comentários:

Heurística é uma técnica que permite aprimorar progressivamente o processo de busca por soluções para um problema. Em outras palavras, algoritmos com heurísticas buscam resolver problemas por aproximações progressivas. Eles são bastante utilizados em análises de agrupamento, uma vez que essa é uma técnica não supervisionada, logo não possui categorias de saída pré-definidas – o próprio algoritmo deve aprender a encontrar similaridades sem a intervenção de um ser humano. Dito isso, algoritmos com heurísticas são extremamente adequados para descobrir agregações naturais entre objetos.

Gabarito: Correto

---

25. (CESPE / TCE-RJ – 2021) O fator de suporte e o fator de confiança são dois índices utilizados para definir o grau de certeza de uma regra de associação.

Comentários:

Perfeito! Ambas são medidas utilizadas para medir a qualidade ou grau de certeza de uma regra de associação.

Gabarito: Correto

---

26. (CESPE / TCE-RJ – 2021) Os principais métodos de análise de agrupamentos em mineração de dados incluem redes neurais, lógica difusa, métodos estatísticos e algoritmos genéticos.

Comentários:

Perfeito! Existem vários métodos de realizar a clusterização, tais como: redes neurais, lógica difusa, métodos estatísticos e algoritmos genéticos – notem que alguns métodos podem ser utilizados em diversas técnicas de mineração de dados.

Gabarito: Correto

---

27. (CESPE / Polícia Federal – 2021) A análise de *clustering* é uma tarefa que consiste em agrupar um conjunto de objetos de tal forma que estes, juntos no mesmo grupo, sejam mais semelhantes entre si que em outros grupos.

Comentários:

A análise de clustering, também conhecida como agrupamento, é uma técnica de aprendizado de máquina não supervisionado que tem como objetivo dividir um conjunto de objetos em grupos, ou clusters, de modo que os objetos em cada grupo sejam mais semelhantes entre si do que com aqueles em outros grupos.

Esta semelhança é geralmente medida com base em características ou distâncias definidas no espaço de atributos dos dados. Clustering é uma ferramenta comum em mineração de dados e análise de big data, utilizada em diversos campos como marketing, biologia, bibliotecas digitais, análise de redes sociais, entre outros.

Em suma: a análise de clustering (agrupamento) é uma técnica de mineração de dados que identifica e separa objetos semelhantes entre si e diferentes dos demais.

Gabarito: Correto

---

28. (CESPE / ME – 2020) Aprendizagem de máquina pode ajudar a clusterização na identificação de outliers, que são objetos completamente diferentes do padrão da amostra.

Comentários:

Outliers são dados que se diferenciam drasticamente de todos os outros. Em outras palavras, um outlier é um valor que foge da normalidade e que pode (e provavelmente irá) causar anomalias nos resultados obtidos por meio de algoritmos e sistemas de análise. Entender os outliers é fundamental em uma análise de dados por pelo menos dois aspectos:

1. Os outliers podem enviesar negativamente todo o resultado de uma análise;
2. O comportamento dos outliers pode ser justamente o que está sendo procurado.

O aprendizado de máquina pode ser usado para tratar outliers de um conjunto de dados de forma adequada.

Gabarito: Correto

---

29. (CESPE / ME – 2020) A técnica de associação é utilizada para indicar um grau de afinidade entre registros de eventos diferentes, para permitir o processo de data mining.

Comentários:

Regras de Associação realmente descobrem correlações entre variáveis indicando afinidades entre registros de eventos. Elas têm como premissa básica encontrar elementos que implicam na presença de outros elementos em uma mesma transação, ou seja, encontrar relacionamentos ou padrões frequentes entre conjuntos de dados.

Gabarito: Correto

---

30. (CESPE / ME – 2020) No modelo CRISP-DM, a fase na qual se planejam todas as atividades para carga dos dados é denominada entendimento dos dados.

Comentários:

Opa... o entendimento das atividades é feito na fase de Entendimento do Negócio.

Gabarito: Errado

---

31. (CESPE / ME – 2020) Na etapa de mineração do data mining, ocorre a seleção dos conjuntos de dados que serão utilizados no processo de mining.

Comentários:

Na etapa de Mineração do Data Mining ocorre a seleção dos conjuntos de dados que serão utilizados no processo de mining? Essa frase não faz nenhum sentido – Data Mining é somente a tradução de Mineração de Dados.

Gabarito: Errado

---

32. (CESPE / Ministério da Economia – 2020) A técnica de agregação na mineração de dados atua em conjunto de registros que tenham sido previamente classificados.

Comentários:

Opa... a técnica de agregação (também chamada de agrupamentos, clusters, grupos, aglomerados ou segmentos) atua em conjunto de registros que não tenham sido previamente classificados, uma vez que utilizam uma abordagem de aprendizagem não supervisionada.

Gabarito: Errado

---

33. (CESPE / Ministério da Economia – 2020) O objetivo da etapa de pré-processamento é diminuir a quantidade de dados que serão analisados, por meio da aplicação de filtros e de eliminadores de palavras.

Comentários:

A etapa de pré-processamento diz respeito a um conjunto de atividades preparatórias que utiliza algoritmos de naturezas diferentes, mas com objetivos maiores em comum: diminuir a quantidade de dados a ser processado, diminuir a ambiguidade das expressões linguísticas e estruturar as informações como tuplas. Eu diria que reduzir a quantidade de dados a serem analisados é um objetivo secundário, mas – de acordo com alguns autores – ainda é um dos objetivos da etapa de pré-processamento. Questão complicada...

Gabarito: Correto

---

34. (CESPE / Ministério da Economia – 2020) Modelagem preditiva é utilizada para antecipar comportamentos futuros, por meio do estudo da relação entre duas ou mais variáveis.

Comentários:

Perfeito! A técnica de previsão é utilizada para modelar e encontrar padrões que utilizam dados históricos para realizar previsões de tendências, padrões de comportamentos ou eventos futuros. Ela o faz por meio do estudo probabilístico entre duas ou mais variáveis.

Gabarito: Correto

---

35. (CESPE / ME – 2020) Outlier ou anomalias são padrões nos dados que não estão de acordo com uma noção bem definida de comportamento normal.

Comentários:

outliers são valores extremos que se desviam de outras observações nos dados – eles podem indicar uma variabilidade em uma medição, erros experimentais ou uma novidade. Em outras palavras, um outlier é uma observação que diverge de um padrão geral em uma amostra.

Gabarito: Correto

---

36. (CESPE / ME – 2020) A análise de regressão em mineração de dados tem como objetivos a sumarização, a predição, o controle e a estimação.

Comentários:

A análise de regressão consiste na realização de uma análise estatística com o objetivo de verificar a existência de uma relação funcional entre uma variável dependente com uma ou mais variáveis independentes. De maneira geral, a análise de regressão tem como objetivos a sumarização, a predição, o controle, a estimação, a seleção de variáveis e a inferência.

Gabarito: Correto

---

37. (CESPE / TJ-AM – 2019) A técnica machine learning pode ser utilizada para apoiar um processo de data mining.

Comentários:

Perfeito! A base da mineração de dados compreende três disciplinas científicas entrelaçadas que existem há tempos: Estatística (o estudo numérico das relações entre dados), Inteligência Artificial (inteligência exibida por softwares e/ou máquinas, que se assemelha à humana) e Machine Learning (algoritmos que podem aprender com dados para realizar previsões).

Gabarito: Correto

---

38. (CESPE / POLÍCIA FEDERAL – 2018) Pode-se definir mineração de dados como o processo de identificar, em dados, padrões válidos, novos, potencialmente úteis e, ao final, compreensíveis.

Comentários:

De fato, a mineração de dados identifica em um grande conjunto de dados, padrões válidos, novos potencialmente úteis e – ao final – compreensíveis. Perfeito!

Gabarito: Correto

---

39. (CESPE / FUB – 2018) No Data Mining, uma regra de associação relaciona a presença de um conjunto de itens com outra faixa de valores de um outro conjunto de variáveis.

Comentários:

Perfeito! Na Mineração de Dados, uma regra de associação relaciona a presença de um conjunto de itens com outra faixa de valores de um outro conjunto de variáveis. Uma regra de associação pode ser vista como uma expressão da forma X → Y, onde há a relação dos valores de X e Y em um certo conjunto de valores (Ex: {fralda} → {cerveja}).

Gabarito: Correto

---

40. (CESPE / Polícia Federal – 2018) A mineração de dados se caracteriza especialmente pela busca de informações em grandes volumes de dados, tanto estruturados quanto não estruturados, alicerçados no conceito dos 4V's: volume de mineração, variedade de algoritmos, velocidade de aprendizado e veracidade dos padrões.

Comentários:

Opa... a questão trata de Big Data e, não, Mineração de Dados!

Gabarito: Errado

---

41. (CESPE / Polícia Federal – 2018) Descobrir conexões escondidas e prever tendências futuras é um dos objetivos da mineração de dados, que utiliza a estatística, a inteligência artificial e os algoritmos de aprendizagem de máquina.

Comentários:

Descobrir conexões escondidas? Prever tendências futuras? Ambas são objetivos comuns da mineração de dados. Além disso, ela realmente utiliza estatística, inteligência artificial e algoritmos de aprendizagem de máquina. O processo de minerar dados para descobrir conexões escondidas e prever tendências futuras tem uma longa história. Sua base compreende três disciplinas científicas entrelaçadas que existem há tempos: Estatística (o estudo numérico das relações entre dados), Inteligência Artificial (inteligência exibida por softwares e/ou máquinas, que se assemelha à humana) e Machine Learning (algoritmos que podem aprender com dados para realizar previsões).

Gabarito: Correto

---

42. (CESPE / Polícia Federal – 2018) Situação hipotética: Na ação de obtenção de informações por meio de aprendizado de 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 aprendizagem supervisionados. Assertiva: Nessa situação, a ação em realização está relacionada ao processo de classificação.

Comentários:

Perfeito! Examinar características de determinado objeto e atribuir-lhe uma ou mais classes é um exemplo típico do processo de classificação (que realmente é um algoritmo de aprendizagem supervisionado).

Gabarito: Correto

CPF
NOME
DATA DE NASCIMENTO
NOME DO PAI
NOME DA MAE
TELEFONE
CEP
NUMERO

As informações anteriormente apresentadas correspondem aos campos de uma tabela de um banco de dados, a qual é acessada por mais de um sistema de informação e também por outras tabelas. Esses dados são utilizados para simples cadastros, desde a consulta até sua alteração, e também para prevenção à fraude, por meio de verificação dos dados da tabela e de outros dados em diferentes bases de dados ou outros meios de informação.

Considerando essas informações, julgue o item que se segue.

43. (CESPE / Polícia Federal – 2018) Se um sistema de informação correlaciona os dados da tabela em questão com outros dados não estruturados, então, nesse caso, ocorre um processo de mineração de dados.

Comentários:

Um processo de mineração de dados busca explorar grande quantidade de dados para extração não-trivial de informação implícita desconhecida. Eu não vejo como a simples correlação entre dados estruturados de uma tabela com dados não estruturados possa ser considerado mineração de dados (talvez possa ser considerada a forma mais primitiva de mineração), no entanto o gabarito definitivo da banca foi verdadeiro.

Gabarito: Correto

44.(CESPE / EBSERH – 2018) A descoberta de novas regras e padrões em conjuntos de dados fornecidos, ou aquisição de conhecimento indutivo, é um dos objetivos de data mining.

Comentários:

Perfeito! O conhecimento normalmente é classificado como indutivo ou dedutivo. O conhecimento dedutivo deduz novas informações com base na aplicação de regras lógicas previamente especificadas de dedução sobre o dado indicado. Já a mineração de dados enfoca o conhecimento indutivo, que descobre novas regras e padrões com base nos dados fornecidos.

Gabarito: Correto

45.(CESPE / IPHAN – 2018) Na busca de padrões no data mining, é comum a utilização do aprendizado não supervisionado, em que um agente externo apresenta ao algoritmo alguns conjuntos de padrões de entrada e seus correspondentes padrões de saída, comparando-se a resposta fornecida pelo algoritmo com a resposta esperada.

Comentários:

Se o agente externo apresenta ao algoritmo alguns conjuntos de padrões de entrada, então ele os conhece de antemão. Dessa forma, trata-se de um aprendizado supervisionado e, não, um aprendizado não supervisionado.

Gabarito: Errado

46.(CESPE / EBSERH – 2018) A descoberta de novas regras e padrões em conjuntos de dados fornecidos, ou aquisição de conhecimento indutivo, é um dos objetivos de data mining.

Comentários:

Perfeita definição de mineração de dados...

Gabarito: Correto

47.(CESPE / STJ – 2018) O processo de mineração de dados está intrinsecamente ligado às dimensões e a fato, tendo em vista que, para a obtenção de padrões úteis e relevantes, é necessário que esse processo seja executado dentro dos data warehouses.

Comentários:

Vocês não precisam entender o que são dimensões e fatos – basta lembrar que não é necessário que o processo de mineração de dados ocorra dentro de um data warehouse.

Gabarito: Errado

48.(CESPE / TCM-BA – 2018) A respeito das técnicas e(ou) métodos de mineração de dados, assinale a opção correta.

a) O agrupamento (ou clustering) realiza identificação de grupos de dados que apresentam coocorrência.

b) A classificação realiza o aprendizado de uma função que pode ser usada para mapear os valores associados aos dados em um ou mais valores reais.

c) A regressão ou predição promove o aprendizado de uma função que pode ser usada para mapear dados em uma de várias classes discretas definidas previamente, bem como encontrar tendências que possam ser usadas para entender e explorar padrões de comportamento dos dados.

d) As regras de associação identificam grupos de dados, em que os dados têm características semelhantes aos do mesmo grupo e os grupos têm características diferentes entre si.

e) Os métodos de classificação supervisionada podem ser embasados em separabilidade (entropia), utilizando árvores de decisão e variantes, e em particionamento, utilizando SVM (support vector machines).

Comentários:

(a) Errado, ele realiza a identificação de grupos de dados que apresentam similaridade – coocorrência é utilizada nas regras de associação; (b) Errado, essa seria a descrição de regressão e, não, classificação; (c) Errado, essa seria a descrição de classificação e, não, regressão; (d) Errado, essa seria a definição de agrupamento (ou *clustering*); (e) Correto, porém com uma ressalva. Eu nunca encontrei a fonte dessa questão, mas – na minha visão – a separabilidade (entropia) está relacionada tanto a Árvores de Decisão quanto ao SVM.

Gabarito: Letra E

49.(CESPE / TCM-BA – 2018) Assinale a opção correta a respeito do CRISP-DM.

a) CRISP-DM é uma suíte de ferramentas proprietárias que vem se tornando um padrão da indústria para mineração de dados, uma vez que fornece um plano completo e tecnologias para a realização de um projeto de mineração de dados.

b) A verificação da qualidade dos dados é uma atividade da fase de entendimento dos dados.

c) Durante a fase de preparação dos dados, é realizado um inventário de requisitos, suposições e restrições de recursos.

d) Na fase de avaliação dos dados, são realizadas as atividades de identificar valores especiais dos dados e catalogar seu significado.

e) Na fase de preparação dos dados, são realizadas as atividades de analisar o potencial de implantação de cada resultado e estimar o potencial de melhoria do processo atual.

Comentários:

(a) Errado, não se trata de uma suíte de ferramentas, mas de um modelo de referência; (b) Correto, de fato na fase de entendimento dos dados está a atividade de verificação da qualidade dos dados; (c) Errado, essa é uma atividade da fase de Entendimento do Negócio; (d) Errado, o nome correto da fase seria Teste e Avaliação. Além disso, essas atividades ocorreriam na fase de entendimento dos dados; (e) Errado, essa é uma atividade da fase de Teste e Avaliação.

Gabarito: Letra B

---

50. (CESPE / SEDF – 2017) Agrupar registros em grupos, de modo que os registros em um grupo sejam semelhantes entre si e diferentes dos registros em outros grupos é uma maneira de descrever conhecimento descoberto durante processos de mineração de dados.

Comentários:

Agrupar registros em grupos? Semelhantes entre si e diferentes de outros grupos? Um bocado de palavra-chave para identificar que se trata de uma maneira de descrever conhecimento descoberto durante processos de mineração de dados por meio da análise de agrupamento (ou clusterização).

Gabarito: Correto

---

51. (CESPE / FUNPRESP-EXE – 2016) Na implementação de mineração de dados (data mining), a utilização da técnica de padrões sequenciais pode ser útil para a identificação de tendências.

Comentários:

Algoritmos de padrões sequenciais identificam tipos de padrões sequenciais em restrições mínimas especificadas pelo usuário. Esta técnica procura por compras ou eventos que ocorrem em uma sequência através do tempo. Por exemplo: uma loja pode descobrir que consumidores que compram TVs tendem também a comprar filmadoras de 8mm em 50% das vezes. Em outras palavras, pode identificar tendências de compras.

Gabarito: Correto

---

52. (CESPE / TJ/SE – 2016) DataMining pode ser considerado uma etapa no processo de descoberta de conhecimento em base de dados, consistindo em análise de conjuntos de dados cujo objetivo é descobrir padrões úteis para tomada de decisão.

Comentários:

Ele realmente é considerado uma etapa em um processo maior para descobrir padrões úteis.

Gabarito: Correto

---

53. (CESPE / FUNPRESP/JUD – 2016) Em DataMining, as árvores de decisão podem ser usadas com sistemas de classificação para atribuir informação de tipo.

Comentários:

As árvores de decisão são variações da técnica de classificação, logo podem – sim – ser usadas com sistemas de classificação para atribuir informação do tipo.

Gabarito: Correto

---

54. (CESPE / TRT-18ª Região – 2016) Acerca de data mining, assinale a opção correta.

a) A fase de preparação para implementação de um projeto de data mining consiste, entre outras tarefas, em coletar os dados que serão garimpados, que devem estar exclusivamente em um data warehouse interno da empresa.

b) As redes neurais são um recurso matemático/computacional usado na aplicação de técnicas estatísticas nos processos de data mining e consistem em utilizar uma massa de dados para criar e organizar regras de classificação e decisão em formato de diagrama de árvore, que vão classificar seu comportamento ou estimar resultados futuros.

c) As aplicações de data mining utilizam diversas técnicas de natureza estatística, como a análise de conglomerados (cluster analysis), que tem como objetivo agrupar, em diferentes conjuntos de dados, os elementos identificados como semelhantes entre si, com base nas características analisadas.

d) As séries temporais correspondem a técnicas estatísticas utilizadas no cálculo de previsão de um conjunto de informações, analisando-se seus valores ao longo de determinado período. Nesse caso, para se obter uma previsão mais precisa, devem ser descartadas eventuais sazonalidades no conjunto de informações.

e) Os processos de data mining e OLAP têm os mesmos objetivos: trabalhar os dados existentes no data warehouse e realizar inferências, buscando reconhecer correlações não explícitas nos dados do data warehouse.

Comentários:

(a) Errado. Não devem estar necessariamente em um Data Warehouse; (b) Errado. Isso é função das Árvores de Decisão e, não, Redes Neurais; (c) Correto. As aplicações de data mining utilizam diversas técnicas de natureza estatística, como a análise de conglomerados (cluster analysis), que tem como objetivo agrupar, em diferentes conjuntos de dados, os elementos identificados como semelhantes entre si, com base nas características analisadas; (d) Errado. Devem ser consideradas eventuais sazonalidades no conjunto de informações; (e) Errado. Esses não são objetivos de uma Ferramenta OLAP.

Gabarito: Letra C

---

55. (CESPE / TCE-PA – 2016) No contexto de data mining, o processo de descoberta de conhecimento em base de dados consiste na extração não trivial de conhecimento previamente desconhecido e potencialmente útil.

Comentários:

De acordo com Fayyad, a Descoberta de Conhecimento em Base de Dados (KDD) é um processo não trivial de identificação de novos padrões válidos, úteis e compreensíveis. A ideia é tentar adquirir conhecimento sobre algo a partir de uma base de dados. *Por que é não-trivial?* Porque não é extrair qualquer conhecimento – é extrair conhecimento que não era conhecido e que pode ser útil para a tomada de decisão de uma organização.

Gabarito: Correto

---

56. (CESPE / MEC – 2015) O conhecimento obtido no processo de data mining pode ser classificado como uma regra de associação quando, em um conjunto de eventos, há uma hierarquia de tuplas sequenciais.

Comentários:

Nooooope! A questão trata de classificação e, não, associação. Vamos revisar: Regras de Associação procuram descobrir relacionamentos entre variáveis em bancos de dados; Classificação procura identificar a existência de hierarquia em um conjunto pré-existente de eventos ou transações. A hierarquia existe na classificação e, não, na associação.

Gabarito: Errado

---

57. (CESPE / MEC – 2015) Situação hipotética: Após o período de inscrição para o vestibular de determinada universidade pública, foram reunidas informações acerca do perfil dos candidatos, cursos inscritos e concorrências. Ademais, que, por meio das soluções de BI e DW que integram outros sistemas, foram realizadas análises para a detecção de relacionamentos sistemáticos entre as informações registradas. Assertiva: Nessa situação, tais análises podem ser consideradas como data mining, pois agregam valor às decisões do MEC e sugerem tendências, como, por exemplo, o aumento no número de escolas privadas e a escolha de determinado curso superior.

Comentários:

Perfeito! Análises que agregam valor às decisões de gestores e que sugerem tendências são típicas características de Mineração de Dados (Data Mining).

Gabarito: Correto

---

58. (CESPE / MEC – 2015) Os objetivos do Data Mining incluem identificar os tipos de relacionamentos que se estabelecem entre informações armazenadas em um grande repositório.

Comentários:

Perfeito! Identificar tipos de relacionamentos que se estabelecem entre informações armazenadas em um grande repositório para descobrir padrões, tendências, anomalias, entre outros é um exemplo típico de Data Mining.

Gabarito: Correto

---

59. (CESPE / MEC – 2015) Algoritmo genético é uma das ferramentas do *data mining* que utiliza mecanismos de biologia evolutiva, como hereditariedade, recombinação, seleção natural e mutação, para solucionar e agrupar problemas.

Comentários:

Algoritmos genéticos utilizam mecanismos de biologia evolutiva, como hereditariedade, recombinação, seleção natural e mutação, para solucionar e agrupar problemas. De acordo com Navathe, eles são uma classe de procedimentos de pesquisa aleatórios capazes de realizar pesquisa adaptativa e robusta por uma grande faixa de topologias de espaço de busca. Eu costumo dizer que é uma maneira de resolver problemas de otimização ao simular um processo de seleção natural.

Gabarito: Correto

60.(CESPE / MEC – 2015) A predição em algoritmos de *data mining* objetiva modelar funções sobre valores para apresentar o comportamento futuro de determinados atributos.

Comentários:

Perfeito! A predição busca descrever a natureza de ocorrências futuras de certos eventos com base nos acontecimentos passados.

Gabarito: Correto

61.(CESPE / MEC – 2015) Selecionar uma amostra e determinar os conjuntos de itens frequentes dessa amostra para formar a lista de previsão de subconjunto são as principais características do algoritmo de previsão.

Comentários:

Essa é a principal característica de um algoritmo de amostragem e, não, de previsão – o algoritmo de amostragem busca selecionar uma amostra pequena e determinar os conjuntos de itens frequentes com base nessa amostra a fim de prever regras de associação. Algoritmos de previsão ou predição buscam avaliar o valor de uma variável ainda não identificada, baseando-se em dados adquiridos por meio do comportamento desta variável no passado.

Gabarito: Errado

62.(CESPE / TCU – 2015) A finalidade do uso do data mining em uma organização é subsidiar a produção de afirmações conclusivas acerca do padrão de comportamento exibido por agentes de interesse dessa organização.

Comentários:

Perfeito! Alguns alunos argumentam que a mineração de dados não produz afirmações conclusivas acerca de um padrão de comportamento, no entanto não é isso que diz a questão. Ela afirma que a mineração de dados busca subsidiar a produção de afirmações conclusivas acerca do padrão de um comportamento – são coisas totalmente diferentes.

Gabarito: Correto

63.(CESPE / TCU – 2015) Quem utiliza o data mining tem como objetivo descobrir, explorar ou minerar relacionamentos, padrões e vínculos significativos presentes em grandes massas documentais registradas em arquivos físicos (analógicos) e arquivos lógicos (digitais).

Comentários:

Opa... a mineração de dados é um processo eminentemente digital, não podendo ser realizado em arquivos físicos/analógicos. Como uma piada que eu vi recentemente, mineração de dados em arquivos físicos só se for realizada pelo estagiário!

Gabarito: Errado

64.(CESPE / TCU – 2015) O uso prático de data mining envolve o emprego de processos, ferramentas, técnicas e métodos oriundos da matemática, da estatística e da computação, inclusive de inteligência artificial.

Comentários:

Perfeito, perfeito, perfeito! Mineração de Dados é um conjunto de processos, métodos, teorias, ferramentas e tecnologias open-end utilizadas para explorar, organizar e analisar de forma automática ou semi-automática uma grande quantidade de dados brutos com o intuito de identificar, descobrir, extrair, classificar e agrupar informações implícitas desconhecidas, além de avaliar correlações, tendências e padrões consistentes de comportamento potencialmente úteis – como regras de associação ou sequências temporais – de forma não-trivial por meio de técnicas estatísticas e matemáticas, como redes neurais, algoritmos genéticos, inteligência artificial, lógica nebulosa, análise de conglomerados (clusters), entre outros.

Gabarito: Correto

65.(CESPE / DEPEN – 2015) Os objetivos do *datamining* incluem identificar os tipos de relacionamentos que se estabelecem entre informações armazenadas em um grande repositório.

Comentários:

Perfeito! O tamanho do repositório é irrelevante para o conceito de mineração, apesar de costumar ocorrer em grandes repositórios de dados.

Gabarito: Correto

---

66. (CESPE / ANTAQ – 2014) Em um processo de descoberta do conhecimento, um Data Mining executado para atingir uma meta pode falhar nas classes de predição, de identificação, de classificação e de otimização.

Comentários:

Questão polêmica! Essa questão foi feita baseada no livro de Elmasri & Navathe. Em sua 5ª Edição, o livro foi traduzido da seguinte forma:

"Data mining é normalmente executada para alguma meta ou aplicação. De forma geral, esses propósitos falham nas seguintes classes: predição, identificação, classificação e otimização".

Na 6ª Edição, o livro alterou a tradução para:

"A mineração de dados costuma ser executada com alguns objetivos finais ou aplicações. De um modo geral, esses objetivos se encontram nas seguintes classes: previsão, identificação, classificação e otimização".

Dito isso, vamos analisar a questão: Data Mining realmente é uma das etapas do processo de descoberta de conhecimento. Além disso, nós acabamos de ver que ele é normalmente executado para atingir alguma meta ou aplicação. Conforme apresentado na primeira tradução, esses propósitos podem falhar nas classes de predição, identificação, classificação ou otimização.

De certa forma, ambas as traduções fazem sentido: se esses são os quatro objetivos da mineração de dados, então eventuais falhas podem ocorrer em algumas dessas classes.

Gabarito: Correto

---

67. (CESPE / TCDF – 2014) Com o uso da classificação como técnica de Data Mining, busca-se a identificação de uma classe por meio de múltiplos atributos. Essa técnica também pode ser usada em conjunto com outras técnicas de mineração de dados.

Comentários:

Realmente essa técnica busca identificar uma classe por meio de múltiplos atributos e pode ser usada – sim – em conjunto com outras técnicas de mineração de dados.

Gabarito: Correto

---

68. (CESPE / TJ-SE – 2014) O uso de agrupamento (clustering) em DataMining exige que os registros sejam previamente categorizados, tendo por finalidade aproximar registros similares para predizer valores de variáveis.

Comentários:

Análise de Agrupamentos é utilizado quando nenhum grupo foi definido e nenhum registro foi previamente categorizado – a questão trata da Classificação.

Gabarito: Errado

---

69. **(CESPE / TJ-SE – 2014)** Assim como o DataMining, os DataMarts são voltados para a obtenção de informações estratégicas de maneira automática, ou seja, com o mínimo de intervenção humana a partir da análise de dados oriundos de DataWarehouses.

Comentários:

Data Mining é um processo de descobrir padrões, associações, mudanças, anomalias e estruturas em grandes quantidades de dados armazenados e Data Marts são simplesmente repositórios departamentais de dados de uma organização. Data Marts geralmente não são voltados para obtenção de informações estratégicas justamente por serem departamentais e, em regra, informações estratégias permeiam diversos departamentos de uma organização.

No entanto, o erro da questão está em afirmar que o Data Mining funciona de maneira automática com o mínimo de intervenção humana. Se ele necessita de intervenção humana, então ele não é automático – é semiautomático.

Gabarito: Errado

---

70. **(CESPE / ANATEL – 2014)** No processo de Data Mining (mineração de dados), é indispensável o uso de técnica conhecida como Data Warehousing, uma vez que a mineração de dados deve ocorrer necessariamente em estruturas não normalizadas (FNo).

Comentários:

Opa... não tem nada de indispensável! Guardem essa informação: a mineração de dados prescinde do Data Warehouse, isto é, não precisa desse repositório de dados para ocorrer! E o final da questão viaja completamente...

Gabarito: Errado

---

71. **(CESPE / TJ-SE – 2014)** Os principais processos de DataMining são a identificação de variações embasado em normas, a detecção e análise de relacionamentos, a paginação de memória e o controle de periféricos.

Comentários:

Examinador viajando... ele realmente envolve identificações de variações/anomalias embasado em normas e detecção/análise de relacionamentos. No entanto, ele não possui nenhuma relação com paginação de memória e controle de periféricos – que seriam atividades de um sistema operacional.

Gabarito: Errado

---

72. **(CESPE / TJ-CE – 2014)** Assinale a opção correta acerca de datamining:

a) A informação acerca dos resultados obtidos no processo de mineração é apresentada apenas de forma gráfica.

b) A classificação, uma das principais tecnologias da mineração de dados, caracteriza-se por possuir um conjunto de transações, sendo cada uma delas relacionada a um itemset.

c) É possível realizar mineração de dados em documentos textuais como, por exemplo, uma página da Internet.

d) A grande desvantagem de um datamining consiste no fato de que a identificação de um padrão, para a geração do conhecimento, só é possível por meio da análise em pequenas quantidades de dados.

e) Durante a fase de reconhecimento de padrões, para cada banco de dados, é permitido um único tipo de padrão.

Comentários:

(a) Errado, pode ser apresentada de diversas outras formas; (b) Errado, as palavras-chave itemset e transação nos remetem à técnica de regras de associação e, não, de classificação; (c) Correto, chama-se mineração de texto; (d) Errado, em geral, é necessário utilizar uma grande quantidade de dados para identificar padrões; (e) Errado, pode-se utilizar diversos padrões diferentes.

Gabarito: Letra C

---

73. (CESPE / MPOG – 2013) ETL é definido como o processo de descobrir padrões, associações, mudanças, anomalias e estruturas em grandes quantidades de dados armazenados ou em repositórios de informação gerais dentro do data mining.

Comentários:

~~ETL~~ Data Mining é definido como o processo de descobrir padrões, associações, mudanças, anomalias e estruturas em grandes quantidades de dados armazenados ou em repositórios de informação gerais dentro geralmente dentro de um ~~data mining~~ Data Warehouse.

Gabarito: Errado

---

74. (CESPE / SERPRO – 2013) Datamining é a tecnologia por intermédio da qual os processos são automatizados mediante racionalização e potencialização por meio de dois componentes: organização e tecnologia.

Comentários:

Que viagem! Data Mining não tem nenhuma relação com automatização de processos.

Gabarito: Errado

---

75. (CESPE / TJ-RO – 2012) A técnica de associação em data mining verifica se há controle ou influência entre atributos ou valores de atributos, no intuito de verificar, mediante a análise de probabilidades condicionais, dependências entre esses atributos.

Comentários:

Essa é a definição formal da técnica de regras de associação.

Gabarito: Correto

---

76. (CESPE / PEFOCE – 2012) O data mining tem por objetivo a extração de informações úteis para tomadas de decisão com base nos grandes volumes de dados armazenados nas organizações. Os dados para o data mining são originados restritamente dos data warehouses, pois estes são os que aglomeram enorme quantidade de dados não voláteis e organizados por assunto.

Comentários:

Opa... não são restritos aos Data Warehouses! Os dados podem ser de diversas fontes diferentes – jamais caiam nessa pegadinha!

Gabarito: Errado

---

77. (CESPE / TJ-AC – 2012) O data mining possibilita analisar dados para obtenção de resultados estatísticos que poderão gerar novas oportunidades ao negócio.

Comentários:

Perfeito! O grande objetivo da mineração de dados é obter informações que possam gerar vantagens competitivas e oportunidades de negócio.

Gabarito: Correto

---

78. (CESPE / SEDUC-AM - 2011) A mineração de dados (data mining) é um método computacional que permite extrair informações a partir de grande quantidade de dados.

Comentários:

A Mineração de Dados realmente permite extrair informações a partir de uma grande quantidade de dados – em geral, extraídas de Data Warehouses.

Gabarito: Correto

---

79. (CESPE / MEC – 2011) A exploração, no sentido de utilizar as informações contidas em um datawarehouse, é conhecida como data mining.

Comentários:

A redação não ficou muito bacana! A exploração, no sentido de buscar novos padrões interessantes para o negócio, é conhecida como Data Mining. Já a exploração, no sentido de simplesmente utilizar informações contidas em um DW, não é necessariamente Data Mining. Acredito que caberia recurso! Por fim, não é obrigatório que ocorra em um Data Warehouse, mas é o mais comum...

Gabarito: Correto

---

80. (CESPE / Correios – 2011) Um dos métodos de classificação do datamining é o de análise de agrupamento (cluster), por meio do qual são determinadas características sequenciais utilizando-se dados que dependem do tempo, ou seja, extraindo-se e registrando-se desvios e tendências no tempo.

Comentários:

A questão se refere ao padrão temporal das regras de associação (e, não, análise de agrupamento), isto é, similaridades podem ser detectadas dentro de posições de uma série temporal de dados, que é uma sequência de dados tomados em intervalos regulares. Seu objetivo é modelar o estado do processo extraindo e registrando desvios e tendências no tempo.

Gabarito: Errado

---

81. (CESPE / TJ-ES – 2011) Mineração de dados, em seu conceito pleno, consiste na realização, de forma manual, de sucessivas consultas ao banco de dados com o objetivo de descobrir padrões úteis, mas não necessariamente novos, para auxílio à tomada de decisão.

Comentários:

Vamos corrigir a redação do item: *Mineração de dados, em seu conceito pleno, consiste na realização, de forma ~~manual~~* ***automática ou semiautomática****, de sucessivas consultas ao banco de dados com o objetivo de descobrir padrões úteis, ~~mas não necessariamente~~* ***novos****, para auxílio à tomada de decisão*.

Gabarito: Errado

82.(CESPE / PREVIC – 2011) Um banco de dados pode conter objetos de dados que não sigam o padrão dos dados armazenados. Nos métodos de mineração de dados, esses objetos de dados são tratados como exceção, para que não induzirem a erros na mineração.

Comentários:

Objetos de dados que não seguem um padrão são conhecidos como anomalias (outliers). No contexto de mineração de dados, esses objetos de dados realmente são tratados como exceções ou distorções, mas não necessariamente para não induzir a erros de mineração. Na verdade, é comum que os dados anormais sejam justamente a razão para a mineração dos dados. Dados fora do padrão podem indicar uma variabilidade em uma medição, erros experimentais ou justamente um desvio de comportamento que está sendo procurado.

Gabarito: Errado

83.(CESPE / SERPRO – 2010) A mineração de dados (datamining) é uma atividade de processamento analítico não trivial, que, por isso, deve ser realizada por especialistas em ferramentas de desenvolvimento de software e em repositórios de dados históricos orientados a assunto (datawarehouse).

Comentários:

Em primeiro lugar, a mineração de dados não deve ser realizada em Data Warehouses – apesar de ser comum; em segundo lugar, ela não deve ser realizada por especialistas em ferramentas de desenvolvimento de software (isto é, programadores) e, sim, por usuários finais de negócio.

Gabarito: Errado

84.(CESPE / TRT-RN – 2010) O data mining é um processo automático de descoberta de padrões, de conhecimento em bases de dados, que utiliza, entre outros, árvores de decisão e métodos bayesianos como técnicas para classificação de dados.

Comentários:

Ele pode ser um processo automático ou semiautomático de descoberta de padrões, de conhecimento em bases de dados, que utiliza, entre outros, árvores de decisão e métodos bayesianos como técnicas para classificação de dados. Logo, não há erro na questão!

Gabarito: Correto

85.(CESPE / EMBASA – 2010) Data mining é o processo de extração de conhecimento de grandes bases de dados, sendo estas convencionais ou não, e que faz uso de técnicas de inteligência artificial.

Comentários:

Eu não vejo absolutamente nenhum erro nessa questão! Ele pode ou não ocorrer em grandes bases de dados, podem ser bases convencionais ou não e pode utilizar técnicas de inteligência artificial ou não, mas a banca a considerou falsa.

Gabarito: Errado

---

86. (CESPE / SECONT/ES – 2009) A mineração de dados (data mining) é uma das etapas do processo de descoberta de conhecimento em banco de dados. Nessa etapa, pode ser executada a técnica previsão, que consiste em repartir os dados de modo que diferentes classes ou categorias possam ser identificadas com base em combinações de parâmetros.

Comentários:

É realmente uma das etapas do KDD! No entanto, a técnica de repartir os dados de modo que diferentes classes ou categorias possam ser identificadas com base em combinações de parâmetros é, na verdade, a técnica de classificação e, não, previsão.

Gabarito: Errado

---

87. (CESPE / IPEA – 2008) O data mining é um processo utilizado para a extração de dados de grandes repositórios para tomada de decisão, mas sua limitação é não conseguir analisar dados de um data warehouse.

Comentários:

Data Mining não é utilizado para extração de dados e, sim, conhecimentos ou insights de grandes repositórios para tomada de decisão. Além disso, ele não só consegue como frequentemente é utilizado para analisar dados de um Data Warehouse. A mineração de dados pode ser usada junto com um Data Warehouse para ajudar com certos tipos de decisões. Para bancos de dados muito grandes, que rodam terabytes ou até petabytes de dados, o uso bem-sucedido das aplicações de mineração de dados dependerá, em geral, da construção de um Data Warehouse.

Gabarito: Errado

---

88. (CESPE / SERPRO – 2008) A data mining apóia a descoberta de regras e padrões em grandes quantidades de dados. Em data mining, um possível foco é a descoberta de regras de associação. Para que uma associação seja de interesse, é necessário avaliar o seu suporte, que se refere à freqüência com a qual a regra ocorre no banco de dados.

Comentários:

Uma regra de associação deve satisfazer alguma medida de interesse do analista de dados. As duas principais medidas de interesse são: Suporte e Confiança. O suporte é justamente a frequência com que um conjunto de itens específico ocorre no banco de dados, isto é, o percentual de transações que contém todos os itens em um conjunto (Ex: 50% das compras realizadas em um mercado contém arroz e refrigerante).

Gabarito: Correto

---

89. (CESPE / SERPRO – 2008) A etapa de Mineração de Dados (DM – Data Mining) tem como objetivo buscar efetivamente o conhecimento no contexto da aplicação de KDD (Knowledge Discovery in Databases – Descoberta de Conhecimento em Base de Dados). Alguns autores referem-se à Mineração de Dados e à Descoberta de Conhecimento em Base de Dados como sendo sinônimos. Na etapa de Mineração de Dados são definidos os algoritmos e/ou técnicas que serão utilizados para resolver o problema apresentado. Podem ser usados Redes Neurais, Algoritmo Genéticos, Modelos Estatísticos e Probabilísticos, entre outros, sendo que esta escolha irá depender do tipo de tarefa de KDD que será realizado. "Uma dessas tarefas compreende a busca por uma função que mapeie os registros de um banco de dados em um intervalo de valores reais". Trata-se de:

a) Regressão.
b) Sumarização.
c) Agrupamento.
d) Detecção de desvios.

Comentários:

Quem busca mapear registros de um banco de dados em um intervalo de valores reais é a Regressão.

Gabarito: Letra A

---

90. (CESPE / TCU – 2007) No datamining, o agrupamento e a classificação funcionam de maneira similar: o agrupamento reconhece os padrões que descrevem o grupo ao qual um item pertence, examinando os itens existentes; a classificação é aplicada quando nenhum grupo foi ainda definido.

Comentários:

A questão inverteu os conceitos de agrupamento e classificação.

Gabarito: Errado

QUESTÕES COMENTADAS – FCC

91.(FCC / AL-AP – 2020) 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.

Comentários:

Logo no início do enunciado, afirma-se que pré-existem duas categorias de problemas: clientes adimplentes e clientes inadimplentes. Logo, trata-se de uma técnica de aprendizado supervisionado. A aplicação de mineração de dados buscará descobrir uma função que mapeie corretamente os clientes, a partir de seus dados, em uma destas categorias, podendo ser usada para prever (predição) o comportamento de novos clientes que desejem contrair empréstimos junto à financeira. *Que técnica utiliza aprendizado supervisionado para mapear classes/categorias pré-existentes a fim de predizer comportamentos?* Classificação!

Gabarito: Letra C

92.(FCC / TRF4 – 2019) Um Tribunal pretende analisar fatos (fatores ambientais e perfis profissionais, entre outros) que esclareçam por que alguns colaboradores se destacam profissionalmente enquanto outros não se desenvolvem e acabam por se desligar do órgão. 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 prazos. 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 descrevam tendências futuras, denominado:

a) snowflake.
b) drill over.

c) star schema.
d) slice accross.
e) data mining.

Comentários:

O Tribunal pretende analisar fatos para esclarecer uma questão e, para tal, deseja utilizar um auxílio tecnológico que indique quais características apresentam razões positivas que justifiquem investimentos mais robustos no treinamento de colaboradores que tendem a se destacar a médio e longo prazos. *Galera... por que eu gosto tanto dessa questão?* Porque esse é um exemplo prático de como pode ser utilizada a mineração de dados.

Ora, para que o Tribunal gaste dinheiro em treinamento com seus colaboradores, é necessário que ela entenda quais serão os benefícios. O país está em crise e essas iniciativas são importantíssimas: todas as decisões do alto escalão devem ser tomadas baseadas em dados. Para tal, um analista realmente pode utilizar um processo de análise científica preditiva com base em dados estruturados, que consiste na obtenção de padrões que expliquem e descrevam tendências futuras.

Estamos falando de... Mineração de Dados (Data Mining). As outras opções não fazem qualquer sentido lógico, portanto não se preocupem com isso!

Gabarito: Letra E

---

93. **(FCC / SEFAZ/BA – 2019)** *Além dos indicadores reativos que, uma vez implantados, automaticamente detectam as ocorrências com base nos indicadores mapeados, existem também os controles proativos, que requerem que os gestores os promovam periodicamente. Uma das técnicas que os gestores podem usar requer que sejam selecionadas, exploradas e modeladas grandes quantidades de dados para revelar padrões, tendências e relações que podem ajudar a identificar casos de fraude e corrupção. Relações ocultas entre pessoas, entidades e eventos são identificadas e as relações suspeitas podem ser encaminhadas para apuração específica. As anomalias apontadas por esse tipo de técnica não necessariamente indicam a ocorrência de fraude e corrupção, mas eventos singulares que merecem avaliação individualizada para a exclusão da possibilidade de fraude e corrupção e, no caso da não exclusão, uma investigação.*

(Adaptado de: TCU - Tribunal de Contas da União)

O texto se refere à técnica de:

a) data mart.
b) data warehousing.
c) big data.
d) OLAP.
e) data mining.

Comentários:

Mais uma vez, um exemplo prático! A questão traz um bocado de palavras-chave que ajudam a entender do que ela trata: *grandes quantidades de dados? Relevar padrões? Tendências? Relações? Relações ocultas? Anomalias?* Trata-se de... Mineração de Dados (Data Mining).

Gabarito: Letra E

---

94. (FCC / SANASA – 2019) Considere que a SANASA busca realizar a gestão de recursos hídricos subterrâneos com base em parâmetros conhecidos que determinam a 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.

Comentários:

Regras para realizar o aprendizado supervisionado? Pronto, já sabemos que é classificação!

Gabarito: Letra B

---

95. (FCC / SANASA Campinas – 2019) Considere que a SANASA busca realizar a gestão de recursos hídricos subterrâneos com base em parâmetros conhecidos que determinam a 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.

Comentários:

Regras para realizar aprendizado supervisionado é uma característica da técnica de classificação, que mapeia dados em uma de várias classes discretas definidas previamente.

Gabarito: Letra B

---

96. (FCC / SABESP – 2018) O conceito de Data Mining descreve:

a) o uso de teorias, métodos, processos e tecnologias para organizar uma grande quantidade de dados brutos para identificar padrões de comportamentos em determinados públicos.

b) o conjunto de métodos, tecnologias e estratégias para atração voluntária de visitantes, buscando a conversão consistente de leads em clientes (realização de compra).

c) as atividades coordenadas de modo sistemático por uma determinada organização para relacionamento com os seus distintos públicos, bem como com outras organizações, sejam públicas, privadas ou não governamentais.

d) o conjunto de tarefas e processos, organizados e sistematizados, normalmente como uso de uma plataforma tecnológica (hardware e software, ou até mesmo em cloud computing) para a gestão do relacionamento com clientes.

e) o trabalho de produzir levantamento sobre os hábitos de consumo de mídia de um determinado público, identificando horários, tempo gasto etc., associando ao perfil socioeconômico, potencial de consumo, persuasão etc.

Comentários:

(a) Correto. Identifica padrões de comportamentos em uma grande quantidade de dados brutos; (b) Errado. *Einh?* Esse item não faz nenhum sentido; (c) Errado. Não se trata de uma atividade para relacionamento com públicos; (d) Errado. Não se trata de gestão de relacionamento com clientes; (e) Errado. Não se trata de levantamento de hábitos.

Gabarito: Letra A

---

97. (FCC / SEFAZ-SC – 2018) Para responder à questão, considere o seguinte caso hipotético:

Um Auditor da Receita Estadual pretende descobrir, após denúncia, elementos que possam caracterizar e fundamentar a possível existência de fraudes, tipificadas como sonegação tributária, que vêm ocorrendo sistematicamente na arrecadação do ICMS.

A denúncia é que, frequentemente, caminhões das empresas Org1, Org2 e Org3 não são adequadamente fiscalizados nos postos de fronteiras. Inobservâncias de procedimentos podem ser avaliadas pelo curto período de permanência dos caminhões dessas empresas na operação de pesagem, em relação ao período médio registrado para demais caminhões.

Para caracterizar e fundamentar a existência de possíveis fraudes, o Auditor deverá coletar os registros diários dos postos por, pelo menos, 1 ano e elaborar demonstrativos para análises mensais, trimestrais e anuais.

A aplicação de técnicas de mineração de dados (data mining) pode ser de grande valia para o Auditor. No caso das pesagens, por exemplo, uma ação típica de mining, que é passível de ser tomada com o auxílio de instrumentos preditivos, é:

a) quantificar as ocorrências de possíveis pesagens fraudulentas ocorridas durante todo o trimestre que antecede a data da análise, em alguns postos selecionados, mediante parâmetros comparativos preestabelecidos.

b) analisar o percentual de ocorrências das menores permanências de caminhões nos postos, no último ano, em relação ao movimento total.

c) relacionar os postos onde ocorreram, nos últimos seis meses, as menores permanências das empresas suspeitas e informar o escalão superior para a tomada de decisão.

d) realizar uma abordagem surpresa em determinado posto, com probabilidade significativa de constatar ocorrência fraudulenta.

e) reportar ao escalão superior as características gerais das pesagens e permanências de todos os caminhões, nos cinco maiores postos do Estado, no mês que antecede a data de análise.

Comentários:

(a) Errado, se é do trimestre anterior, então não é uma predição; (b) Errado, se é do ano anterior, então não é uma predição; (c) Errado, se é dos últimos seis meses, então não é uma predição; (d) Correto, a mineração de dados pode identificar um determinado posto com maior probabilidade de ocorrência de fraudes de modo que essa informação possa ser utilizada para realizar uma abordagem surpresa especificamente nesse posto; (e) Errado, se é do mês anterior, então não é uma predição.

Eu vou elaborar um pouquinho mais: os auditores receberam a denúncia de que caminhões de determinadas empresas não estão sendo adequadamente fiscalizados nos postos de fronteiras. No entanto, existe uma infinidade de postos de fronteira e uma abordagem aleatória em cada um deles seria pouco eficaz. No entanto, se o auditor coletar registros diários de cada um dos postos por um

longo período, ele poderá utilizar técnicas de mineração de dados para identificar aqueles postos mais suspeitos de não estarem realizando as fiscalizações.

Um dos dados passados que ele pode observar são os períodos de permanência dos caminhões dessa empresa nas operações de pesagem em relação ao tempo médio dos caminhões de outras empresas. Se a mineração de dados encontrar desvios nesses dados (também chamados de anomalias), ela poderá identificar os postos de fronteira que estão com dados anormais e programar uma abordagem surpresa naqueles postos de forma mais eficiente baseado em uma probabilidade estatística maior em vez de tentar fiscalizar postos aleatoriamente.

Gabarito: Letra D

---

98. (FCC / DPE-RS - 2017) Uma das técnicas bastante utilizadas em sistemas de apoio à decisão é o Data Mining, que se constitui em uma técnica:

a) para a exploração e análise de dados, visando descobrir padrões e regras, a princípio ocultos, importantes à aplicação.

b) para se realizar a criptografia inteligente de dados, objetivando a proteção da informação.

c) que visa sua distribuição e replicação em um cluster de servidores, visando aprimorar a disponibilidade de dados.

d) de compactação de dados, normalmente bastante eficiente, permitindo grande desempenho no armazenamento de dados.

e) de transmissão e recepção de dados que permite a comunicação entre servidores, em tempo real.

Comentários:

(a) Correto, é uma técnica para a exploração e análise de dados, visando descobrir padrões e regras, a princípio ocultos, importantes à aplicação; (b) Errado, não tem nenhuma relação com criptografia inteligente de dados; (c) Errado, não tem nenhuma relação com distribuição e replicação ou disponibilidade de dados; (d) Errado, não tem nenhuma relação com compactação de dados; (e) Errado, não tem nenhuma relação com transmissão e recepção de dados.

Gabarito: Letra A

---

99. (FCC / AL-MS – 2016) Um famoso site de vendas sempre envia ao cliente que acabou de comprar um item X, ou o está analisando, a seguinte frase: Pessoas que compraram o item X também compraram o Y. Para isso, o site deve estar aplicando a técnica de Data Mining denominada:

a) profiling.
b) coocorrência.
c) regressão múltipla.
d) regressão logística.
e) classificação.

Comentários:

A questão trata da técnica de coocorrência (conhecida como Análise de Cesta de Compra/Mercado), cujo objetivo é identificar combinações de itens que ocorrem com frequência significativa em bancos de dados e podem caracterizar, por exemplo, hábitos de consumo de clientes em um supermercado.

Gabarito: Letra B

---

100. (FCC / CNMP – 2015) Em relação às ferramentas de Data Discovery e os fundamentos de Data Mining, é correto afirmar:

a) As ferramentas de Data Mining permitem ao usuário avaliar tendências e padrões não conhecidos entre os dados. Esses tipos de ferramentas podem utilizar técnicas avançadas de computação como redes neurais, algoritmos genéticos e lógica nebulosa, dentre outras.

b) Data Mining é o processo de descobrir conhecimento em banco de dados, que envolve várias etapas. O KDD – Knowledge Discovery in Database é uma destas etapas, portanto, a mineração de dados é um conceito que abrange o KDD.

c) A etapa de KDD do Data Mining consiste em aplicar técnicas que auxiliem na busca de relações entre os dados. De forma geral, existem três tipos de técnicas: Estatísticas, Exploratórias e Intuitivas. Todas são devidamente experimentadas e validadas para o processo de mineração.

d) Os dados podem ser não estruturados (bancos de dados, CRM, ERP), estruturados (texto, documentos, arquivos, mídias sociais, cloud) ou uma mistura de ambos (emails, SOA/web services, RSS). As ferramentas de Data Discovery mais completas possuem conectividade para todas essas origens de dados de forma segura e controlada.

e) Estima-se que, atualmente, em média, 80% de todos os dados disponíveis são do tipo estruturado. Existem diversas ferramentas open source e comerciais de Data Discovery. Dentre as open source está a InfoSphere Data Explorer e entre as comerciais está a Vivisimo da IBM.

Comentários:

(a) Correto. As ferramentas de Data Mining permitem ao usuário avaliar tendências e padrões não conhecidos entre os dados. Esses tipos de ferramentas podem utilizar técnicas avançadas de

computação como redes neurais, algoritmos genéticos e lógica nebulosa, dentre outras; (b) Errado. Na verdade, o Data Mining é uma das etapas do KDD e, não, o contrário; (c) Errado. Na verdade, o Data Mining é uma das etapas do KDD e, não, o contrário; (d) Errado. Na verdade, são ferramentas de Data Mining e, não, Data Discovery; (e) Errado. Na verdade, a maioria é do tipo não-estruturado.

Gabarito: Letra A

---

101. (FCC / TRF-3R – 2014) Mineração de dados é a investigação de relações e padrões globais que existem em grandes bancos de dados, mas que estão ocultos no grande volume de dados. Com base nas funções que executam, há diferentes técnicas para a mineração de dados, dentre as quais estão:

I. identificar afinidades existentes entre um conjunto de itens em um dado grupo de registros. Por exemplo: 75% dos envolvidos em processos judiciais ligados a ataques maliciosos a servidores de dados também estão envolvidos em processos ligados a roubo de dados sigilosos.

II. identificar sequências que ocorrem em determinados registros. Por exemplo: 32% de pessoas do sexo feminino após ajuizarem uma causa contra o INSS solicitando nova perícia médica ajuízam uma causa contra o INSS solicitando ressarcimento monetário.

III. as categorias são definidas antes da análise dos dados. Pode ser utilizada para identificar os atributos de um determinado grupo que fazem a discriminação entre 3 tipos diferentes, por exemplo, os tipos de processos judiciais podem ser categorizados como infrequentes, ocasionais e frequentes.

Os tipos de técnicas referenciados em I, II e III, respectivamente, são:

a) I - Padrões sequenciais
II - Redes Neurais
III - Árvore de decisão

b) I - Redes Neurais
II - Árvore de decisão
III - Padrões sequenciais

c) I - Associação
II - Padrões sequenciais
III - Classificação

d) I - Classificação
II - Associação
III - Previsão

e) I - Árvore de decisão

II - Classificação
III - Associação

Comentários:

(I) Identificar afinidades existentes entre um conjunto de itens em um dado grupo de registros só pode estar relacionado à Associação; (II) Identificar sequências que ocorrem em determinados registros está relacionado a padrões sequenciais; (III) As categorias são definidas antes da análise dos dados.

Gabarito: Letra C

---

102. (FCC / TCE-RS – 2014) A revista da CGU – Controladoria Geral da União, em sua 8ª edição, publicou um artigo que relata que foram aplicadas técnicas de exploração de dados, visando a descoberta de conhecimento útil para auditoria, em uma base de licitações extraída do sistema ComprasNet, em que são realizados os pregões eletrônicos do Governo Federal. Dentre as técnicas preditivas e descritivas utilizadas, estão a classificação, clusterização e regras de associação. Como resultado, grupos de empresas foram detectados em que a média de participações juntas e as vitórias em licitações levavam a indícios de conluio. As técnicas aplicadas referem-se a:

a) On-Line Analytical Processing.
b) Data Mining.
c) Business Process Management.
d) Extraction, Transformation and Load.
e) Customer Churn Trend Analysis.

Comentários:

Técnicas preditivas e descritiva? Classificação, clusterização e regras de associação? Foram encontradas informações novas e úteis que levaram a indícios de conluio? Todas essas características nos levam ao conceito de Data Mining.

Gabarito: Letra B

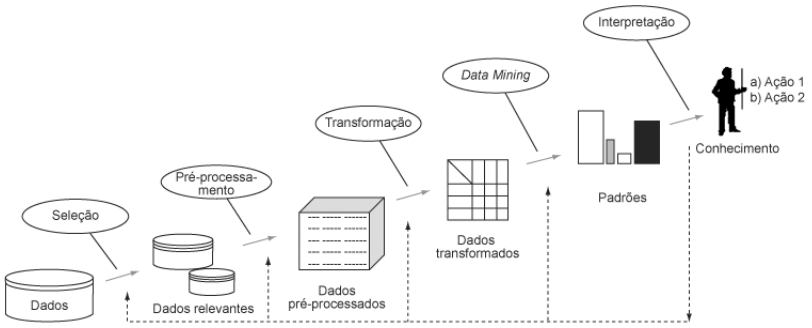
---

103. (FCC / BANESE – 2012) 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.

Comentários:

Data Mining é parte de um processo maior denominado Knowledge Discovery in Database (KDD) ou Descoberta de Conhecimento em Bancos de Dados.



Gabarito: Letra C

---

104. (FCC / TRT/14ª Região – 2011) No contexto de DW, é uma categoria de ferramentas de análise denominada open-end e que permite ao usuário avaliar tendências e padrões não conhecidos entre os dados. Trata-se de:

a) slice.
b) star schema.
c) ODS.
d) ETL.
e) data mining.

Comentários:

Ferramentas de análise denominada open-end e que permite ao usuário avaliar tendências e padrões não conhecidos entre os dados é o Data Mining. *O que é uma ferramenta de análise open-end?* Ao invés de fazerem perguntas, os usuários entregam para a ferramenta grandes quantidades de dados em busca de tendências ou agrupamentos dos dados. Ferramentas de data mining utilizam-se das mais modernas técnicas de computação, como redes neurais, descoberta por regra, detecção de desvio, programação genética, para extrair padrões e associações de dados.

Gabarito: Letra E

---

105. (FCC / INFRAERO – 2011) No âmbito da descoberta do conhecimento (KDD), a visão geral das etapas que constituem o processo KDD (Fayyad) e que são executadas de forma interativa e iterativa apresenta a seguinte sequência de etapas:

a) seleção, pré-processamento, transformação, data mining e interpretação/avaliação.
b) seleção, transformação, pré-processamento, interpretação/avaliação e data mining.
c) data warehousing, star modeling, ETL, OLAP e data mining.

d) ETL, data warehousing, pré-processamento, transformação e star modeling.
e) OLAP, ETL, star modeling, data mining e interpretação/avaliação.

Comentários:

Seleção, Pré-processamento, Transformação, Data Mining, e Interpretação e Avaliação.

Gabarito: Letra A

---

106. (FCC / TRT/4ª Região – 2010) Sobre data mining, é correto afirmar:

a) É o processo de descoberta de novas correlações, padrões e tendências entre as informações de uma empresa, por meio da análise de grandes quantidades de dados armazenados em bancos de dados usando técnicas de reconhecimento de padrões, estatísticas e matemáticas.

b) Não requer interação com analistas humanos, pois os algoritmos utilizados conseguem determinar de forma completa e eficiente o valor dos padrões encontrados.

c) Na mineração de dados, encontrar padrões requer que os dados brutos sejam sistematicamente "simplificados", de forma a desconsiderar aquilo que é genérico e privilegiar aquilo que é específico.

d) É um grande banco de dados voltado para dar suporte necessário nas decisões de usuários finais, geralmente gerentes e analistas de negócios.

e) O processo de descobrimento realizado pelo data mining só pode ser utilizado a partir de um data warehouse, onde os dados já estão sem erros, sem duplicidade, são consistentes e habilitam descobertas abrangentes e precisas.

Comentários:

(a) Correto. É o processo de descoberta de novas correlações, padrões e tendências entre as informações de uma empresa, por meio da análise de grandes quantidades de dados armazenados em bancos de dados usando técnicas de reconhecimento de padrões, estatísticas e matemáticas; (b) Errado. Eles funcionam de forma automática ou semiautomática, mas nem sempre conseguem determinar de forma completa e eficiente o valor dos padrões encontrados; (c) Errado. Informações precisam ser pré-processadas e, não, simplificadas – nada de desconsiderar informações genéricas; (d) Errado. Data Mining não é um banco de dados; (e) Errado. Essa é uma falácia comum – não é obrigatório utilizar um Data Warehouse, apesar de ser o mais comum.

Gabarito: Letra A

---

107. (FCC / TCE-SP – 2010) NÃO é um objetivo da mineração de dados (mining), na visão dos diversos autores,

a) garantir a não redundância nos bancos transacionais.
b) conhecer o comportamento de certos atributos no futuro.
c) possibilitar a análise de determinados padrões de eventos.
d) categorizar perfis individuais ou coletivos de interesse comercial.
e) apoiar a otimização do uso de recursos limitados e/ou maximizar variáveis de resultado para a empresa.

Comentários:

(a) Errado, esse é o objetivo do processo de normalização – não há nenhuma relação com mineração de dados; (b) Correto, esse é objetivo de Previsão; (c) Correto, esse é o objetivo de identificação; (d) Correto, esse é o objetivo de classificação; (e) Correto, esse é o objetivo de Otimização.

Gabarito: Letra A

---

108. (FCC / TCE-SP – 2010) Considere uma dada população de eventos ou novos itens que podem ser particionados (segmentados) em conjuntos de elementos similares, tal como, por exemplo, uma população de dados sobre uma doença que pode ser dividida em grupos baseados na similaridade dos efeitos colaterias produzidos. Como um dos modos de descrever o conhecimento descoberto durante a data mining este é chamado de:

a) associação.
b) otimização.
c) classificação.
d) clustering.
e) temporização.

Comentários:

O modo de descrever conhecimento descoberto é o clustering. Por que não pode ser classificação? Porque o enunciado dá a entender que os grupos não são previamente conhecidos e serão descobertos no decorrer do processo.

Gabarito: Letra D

---

109. (FCC / TCM-PA – 2010) Especificamente, um data mining onde as tendências são modeladas conforme o tempo, usando dados conhecidos, e as tendências futuras são obtidas com base no modelo possui a forma de mining:

a) textual.

b) flocos de neve.
c) espacial.
d) estrela.
e) preditivo.

Comentários:

Utilização de dados conhecidos para modelar tendências futuras é um exemplo típico de mineração preditiva, isto é, a combinação de técnicas para encontrar significado em grandes quantidades de dados a fim de antecipar comportamentos futuros com base em eventos passados.

Gabarito: Letra E

QUESTÕES COMENTADAS – FGV

110. (FGV / Câmara dos Deputados – 2023) CRISP-DM (Cross Industry Standard Process for Data Mining) é uma metodologia utilizada em projetos de Ciência dos Dados. De acordo com esta metodologia, a definição do problema que será investigado por meio de técnicas de mineração de dados ocorre na etapa:

a) *modeling*.
b) *evaluation*.
c) *data preparation*.
d) *data understanding*.
e) *business understanding.*

Comentários:

O CRISP-DM é um modelo de processo padrão na indústria de mineração de dados que descreve abordagens comuns usadas por especialistas em mineração de dados. A definição do problema que será investigado por meio de técnicas de mineração de dados ocorre na etapa de Entendimento do Negócio (*Business Understanding*).

Esta fase inicial concentra-se na compreensão dos objetivos e requisitos do projeto a partir de uma perspectiva de negócios, e então na conversão deste conhecimento em uma definição do problema de mineração de dados e um plano preliminar para alcançá-lo. A definição do problema que será investigado é estabelecida nesta fase.

Gabarito: Letra E

---

111. (FGV / Câmara dos Deputados – 2023) O Coeficiente Silhouette é utilizado na análise de agrupamentos, principalmente para examinar:

a) a separação e a coesão dos agrupamentos.
b) a preservação de pequenos agrupamentos.
c) a completude e a interseção dos agrupamentos.
d) a heterogeneidade dos agrupamentos.
e) a forma convexa dos agrupamentos.

Comentários:

O Coeficiente Silhouette é uma medida usada para interpretar e validar a consistência interna de dados em um modelo de análise de cluster. Este coeficiente oferece uma perspectiva sobre o quão bem um objeto foi classificado e quão bem se encaixa em seu cluster. Os aspectos principais que o Coeficiente Silhouette avalia são:

- **Separação:** refere-se a quão distintos ou bem separados os clusters estão. Um bom coeficiente indica que os clusters não estão apenas separados, mas também são distintos;

- **Coesão:** refere-se a quão próximos os membros de um cluster estão entre si. Em um cluster ideal, os membros estão próximos uns dos outros, indicando alta coesão.

Logo, o Coeficiente Silhouette é utilizado para examinar a separação e a coesão dos agrupamentos.

Gabarito: Letra A

---

112. (FGV / Câmara dos Deputados – 2023) Uma escola está planejando um sistema de acompanhamento temporal de seus alunos, de modo a classificá-los em relação ao desempenho em português e em matemática ao longo de cada ano.

Na escola há uma base de dados históricos que anualmente armazena, para cada aluno, em cada série, a nota final de cada uma dessas duas disciplinas. Essa nota é um valor decimal, entre 0 e 10. Note-se que essa escola, como em outras, há professores que aplicam diferentes graus de exigência nas suas avaliações, uns sendo mais "generosos" e outros, mais "rigorosos".

Três estratégias de transformação de dados foram discutidas, à luz das ideias da Ciência de Dados, como descritas a seguir.

I. Agrupar os alunos a partir de intervalos de notas finais, do tipo "0 até 2,0", "2,1 até 4,0", ..., "8,1 até 10".

II. Rotular grupos de desempenho, "Aprovado" e "Reprovado" e agrupar os alunos de acordo com os critérios de aprovação vigentes em cada situação.

III. Rotular grupos de desempenho, do tipo "Grupo A", "Grupo B", ..., "Grupo E", e agrupar separadamente os alunos de cada conjunto ano/série/disciplina/professor de acordo com a distribuição relativa das notas em cada conjunto.

À luz da ciência de dados e do exposto acima, assinale a afirmativa correta.

a) A primeira estratégia é a melhor para a escola, pois manipula unicamente números que produzem conclusões irrefutáveis.

b) As estratégias II e III complementam-se, pois uma classifica os alunos a partir de parâmetro importante e, a outra, permite uma análise que tenta isolar o grau de exigência de cada professor, e as nuances didáticas de cada disciplina.

c) A segunda estratégia é a melhor para a escola, pois, no fundo, a nota de aprovação adotada em uma escola é a verdadeira medida que reflete o aproveitamento nas disciplinas referidas, independentemente dos critérios do professor.

d) Embora as notas sejam todas numéricas, não existem algoritmos que criem os agrupamentos da estratégia III que sejam diferentes dos agrupamentos que seriam obtidos na estratégia I.

e) As estratégias I e III lidam diretamente com as notas e é impossível gerar novos conhecimentos que alterem a interpretação preconizada pelas notas.

Comentários:

As estratégias propostas visam transformar dados de desempenho escolar para análise e classificação dos alunos. Vamos analisar cada uma:

I. Esta abordagem simplifica os dados, mas pode não capturar nuances importantes, como a variação na dificuldade de avaliação entre diferentes professores. Além disso, ela não considera a subjetividade do grau de exigência de cada professor;

II. Enquanto esta estratégia destaca a distinção fundamental entre aprovação e reprovação, pode não refletir adequadamente as diferenças no desempenho dentro desses grupos ou as variações na exigência dos professores;

III. Esta estratégia tenta ajustar a análise ao contexto específico de cada turma e professor, o que pode oferecer uma visão mais precisa do desempenho dos alunos em relação aos seus pares, considerando a variabilidade nos critérios de avaliação dos professores;

Avaliando as opções:

(A) Errado. A primeira estratégia, embora baseada em números, não aborda a variabilidade na exigência dos professores e pode levar a conclusões imprecisas;

(B) Correto. As estratégias II e III se complementam, pois uma foca na distinção fundamental entre aprovação e reprovação, enquanto a outra permite uma análise mais detalhada do desempenho, considerando as variações didáticas e de exigência de cada professor;

(C) Errado. A segunda estratégia, focada apenas em aprovação e reprovação, pode não capturar sutilezas importantes do desempenho dos alunos.

(D) Errado. As estratégias I e III diferem significativamente em como abordam e agrupam os dados, com a III levando em conta a distribuição relativa das notas no contexto específico de cada turma e professor.

(E) Errado. As estratégias I e III tratam os dados de maneiras diferentes e podem gerar insights distintos sobre o desempenho dos alunos.

Gabarito: Letra B

---

113. (FGV / SMF-RJ – 2023) O fiscal de rendas Renan está explorando a base de dados sobre a situação fiscal de empresas que atuam no Rio de Janeiro, e encontrou os seguintes padrões:

• TIPO_EMPRESA = "MEI", RENDA _ANO = "NIVEL A", -> QUANTIDADE_SOCIOS = 1, SITUACAO_FISCAL = "INADIMPLENTE" (suporte = 50%, confiança = 70%)
• TIPO_EMPRESA = "Simples", RENDA_ANO = "NIVEL B" -> QUANTIDADE_SOCIOS = 2, SITUACAO_FISCAL = "REGULAR" (suporte 30%, confiança = 80%)

A técnica de Mineração de dados que Renan aplicou para descobrir elementos que ocorrem em comum dentro de um determinado conjunto de dados foi:

a) análise de cluster;
b) modelos preditivos;
c) árvores de decisão;
d) regras de associação;
e) técnicas de amostragem.

Comentários:

MEDIDAS DE INTERESSE	DESCRIÇÃO
SUPORTE/ PREVALÊNCIA	Trata-se da frequência com que um conjunto de itens específicos ocorrem no banco de dados, isto é, o percentual de transações que contém todos os itens do conjunto. Em termos matemáticos, a medida de suporte para uma regra X → Y é a frequência em que o conjunto de itens aparece nas transações do banco de dados. Um suporte alto nos leva a crer que os itens do conjunto X e Y costumam ser comprados juntos, pois ocorrem com alta frequência no banco
CONFIANÇA/ FORÇA	Trata-se da probabilidade de que exista uma relação entre itens. Em termos matemáticos, a medida de confiança para uma regra X → Y é a força com que essa regra funciona. Ela é calculada pela frequência dos itens Y serem comprados dado que os itens X foram comprados. Uma confiança alta nos leva a crer que exista uma alta probabilidade de que se X for comprado, Y também será.

Nesse caso, Renan usou regras de associação para identificar padrões com base em atributos como tipo de empresa, renda anual, quantidade de sócios e situação fiscal. Essas regras ajudam a encontrar associações significativas dentro dos dados.

Gabarito: Letra D

---

114. (FGV / SMF-RJ – 2023) Observe a seguinte estrutura do conjunto de dados PESSOA que contém dados sobre pessoas e a sua renda anual.

Coluna	Tipo	Descrição
Idade	Contínua	Idade em anos
Ganho_capital	Contínua	Ganho de capital
Anos_estudo	Contínua	Anos de estudo
Horas_trabalhadas	Contínua	Horas trabalhadas
Sexo	Categórica	Sexo
Raça / Etnia	Categórica	Raça / Etnia
Educação	Categórica	Educação
Ocupação	Categórica	Ocupação
Classe_trabalho	Categórica	Classe de trabalho
Classe	Categórica	Renda (> 50 mil, <= 50 mil)

O conjunto de dados PESSOA será usado para a tarefa de aprendizagem supervisionada de classificação com a finalidade de prever se a renda (Classe) de uma pessoa excede 50 mil por ano. Para isso, a operação de pré-processamento de dados que deve ser executada no conjunto de dados PESSOA é:

a) exclusão da coluna do tipo categórica "Classe" que possui outlier;

b) discretização das colunas do tipo categórica "Sexo, Raça / Etnia e Educação";

c) normalização por padronização das colunas do tipo categórica "Ocupação e Classe_trabalho";

d) normalização das colunas do tipo continua "Idade, Ganho_capital, Anos_estudo e Horas_trabalhadas";

e) imputação de valores com base na média dos valores existentes na coluna do tipo categórica "Sexo" que possui valores faltantes.

Comentários:

(a) Errado. A presença de outliers em uma coluna categórica não justifica a exclusão da coluna inteira. Além disso, Classe é a variável de destino que queremos prever, então não faz nenhum sentido excluí-la;

(b) Errado. *Como assim?* As colunas Sexo, Raça/Etnia e Educação – em regra – já são categóricas por natureza, logo a discretização não é necessária;

(c) Errado. A normalização por padronização é uma técnica de pré-processamento de dados que consiste em redimensionar todas as variáveis de um conjunto de dados para que tenham a mesma média e o mesmo desvio padrão. Ela geralmente é aplicada a dados contínuos e, não, a dados categóricos. Além disso, a normalização não é necessária para colunas categóricas;

(d) Correto. A normalização é geralmente aplicada a dados contínuos para que todas as características estejam na mesma escala, o que pode ser importante para algoritmos sensíveis à escala. Logo, normalizar as colunas contínuas (Idade, Ganho_capital, Anos_estudo e Horas_trabalhadas) é realmente uma etapa apropriada de pré-processamento;

(e) Errado. A imputação é uma técnica usada para preencher valores faltantes em dados com base na média dos valores de uma coluna. No entanto, é comum ser utilizada para imputar valores faltantes em colunas categóricas usando técnicas como preenchimento com o valor mais frequente (moda) – não é comum para colunas contínuas

Gabarito: Letra D

---

115. (FGV / EPPGG - 2023) A mineração de dados ou data mining é uma disciplina interdisciplinar e multidisciplinar que envolve diversas áreas de conhecimento. Assinale a alternativa que enumera corretamente dois tipos de modelagem para análise de dados:

a) Preditivas e Descritivas.
b) Baseadas em Dados e Baseadas em Informação.
c) Matemáticas e Não-Numéricas.
d) Estatísticas e Visualização.
e) Extração e Processamento.

Comentários:

A única alternativa que apresenta dois tipos de modelagem para análise de dados é a letra (a), dado que a modelagem preditiva visa identificar padrões nos dados para prever o futuro e a modelagem descritiva visa entender os dados e suas características.

Gabarito: Letra A

---

116. (FGV / Receita Federal - 2023) A Análise de Componentes Principais (PCA) é uma técnica de transformação de dados que tem como objetivo encontrar as direções de maior variação nos dados, geralmente representadas pelos chamados componentes principais, e gerar novas representações dos dados.

Assinale o objetivo principal dessa técnica.

a) Discretização dos dados.
b) Redução da dimensionalidade dos dados.
c) Normalização dos dados.
d) Padronização dos dados.
e) Cálculo de distâncias entre os dados.

Comentários:

O objetivo principal da Análise de Componentes Principais (PCA) é a redução da dimensionalidade dos dados, mantendo a maior quantidade possível de informações contidas neles.

Através da identificação dos componentes principais, que são as direções de maior variação nos dados, é possível projetar os dados em um espaço de menor dimensão, sem perder informações relevantes. As outras alternativas mencionadas (discretização, normalização, padronização e cálculo de distâncias) podem ser etapas complementares da análise de dados, mas não são o objetivo principal da técnica de PCA.

Gabarito: Letra B

---

117. (FGV / TCE-TO – 2022) Ao analisar um grande volume de dados, João encontrou algumas anomalias, por exemplo: pessoas com mais de 200 anos de idade e salário de engenheiro menor que salário de pedreiro.

A operação de limpeza da fase de preparação de dados para tratar os pontos extremos existentes em uma série temporal a ser executada por João é:

a) Normalização;
b) Discretização;
c) Classificação;
d) Tratamento de outlier;
e) Redução de dimensionalidade.

Comentários:

Se João detectou anomalias, ele deve tratá-las por meio da técnica de Tratamento de Outliers.

Gabarito: Letra D

---

118. (FGV / TJDFT – 2022) Maria está explorando a seguinte tabela da base de dados de vendas do mercado HortVega:

IDvenda	*ItensComprados*
1	Cacau, castanha, cogumelo, chia
2	Cacau, chia
3	Cacau, aveia
4	Castanha, cogumelo, tâmara

Utilizando técnicas de Mineração de Dados, Maria encontrou a seguinte informação:

Se um cliente compra Cacau, a probabilidade de ele comprar chia é de 50%. Cacau => Chia, suporte = 50% e confiança = 66,7%.

Para explorar a base de dados do HortVega, Maria utilizou a técnica de Mineração de Dados:

a) normalização;
b) classificação;
c) regra de associação;
d) clusterização;
e) redução de dimensionalidade.

Comentários:

Falou em suporte e confiança, devemos lembrar de Regras de Associação. Como uma grande quantidade de regras de associação pode ser derivada a partir de uma base de dados, mesmo que pequena, normalmente se objetiva a derivação de regras que suportem um grande número de transações e que possuam uma confiança razoável para as transações às quais elas são aplicáveis. Esses requisitos estão associados a dois conceitos centrais em mineração de regras de associação:

- **Suporte:** o suporte, ou cobertura, de uma regra de associação é o número de transações para as quais ela faz a predição correta. Também pode ser entendida como a utilidade de uma dada regra.

- **Confiança:** a confiança, ou acurácia, de uma regra é o número de transações que ela prediz corretamente proporcionalmente às transações para as quais ela se aplica. Também pode ser entendida como a certeza de uma dada regra.

Gabarito: Letra C

---

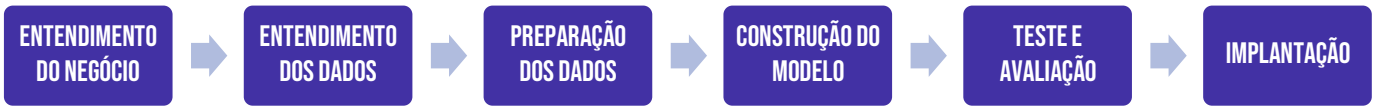
119. (FGV / SEFAZ-AM – 2022) Leia o fragmento a seguir. "CRISP-DM é um modelo de referência não proprietário, neutro, documentado e disponível na Internet, sendo amplamente utilizado para descrever o ciclo de vida de projetos de Ciência de Dados. O modelo é composto por seis fases:

1. entendimento do negócio;
2. _____;
3. _____;
4. Modelagem;
5. _____ ; e
6. implantação".

Assinale a opção cujos itens completam corretamente as lacunas do fragmento acima, na ordem apresentada.

a) modelagem do negócio – limpeza de dados – testagem.
b) modelagem de requisitos – raspagem de dados – execução.
c) modelagem do negócio – mineração de dados – reexecução.
d) compreensão dos dados – preparação dos dados – avaliação.
e) mapeamento de metadados – mineração de dados – testagem.

Comentários:



Os nomes variam um pouco, mas temos que: (2) Entendimento/Compreensão de Dados; (3) Preparação dos Dados; (5) Teste e Avaliação.

Gabarito: Letra D

---

120. (FGV / SEFAZ-AM – 2022) O tipo de aprendizado máquina, que consiste em treinar um sistema a partir de dados que não estão rotulados e/ou classificados e utilizar algoritmos que buscam descobrir padrões ocultos que agrupam as informações de acordo com semelhanças ou diferenças, é denominado:

a) dinâmico.
b) sistêmico.
c) por reforço.
d) supervisionado.
e) não supervisionado.

Comentários:

O tipo de aprendizado máquina, que consiste em treinar um sistema a partir de dados que não estão rotulados e/ou classificados é chamado de aprendizado não supervisionado. Ora, se os dados não são previamente rotulados, então o aprendizado é não-supervisionado.

Gabarito: Letra E

---

121. (FGV / SEFAZ-AM – 2022) Leia o fragmento a seguir.

"A tarefa de detecção de anomalias é um caso particular de problema de _____, onde a quantidade de objetos da classe alvo (anomalia) é muito inferior à quantidade de objetos da classe normal e, adicionalmente, o custo da não detecção de uma anomalia (_____) é normalmente muito maior do que identificar um objeto normal como uma anomalia (_____)".

Assinale a opção cujos itens completam corretamente as lacunas do fragmento acima, na ordem apresentada.

a) aumento de dimensionalidade – redundância – conflito.
b) redução de dimensionalidade – ruído – desvio padrão.
c) análise associativa – discretização – inconsistência.
d) classificação binária – falso negativo – falso positivo.
e) análise probabilística – conflito – ruído.

Comentários:

A tarefa de detecção de anomalias é um caso particular de problema de **classificação binária** onde a quantidade de objetos da classe alvo (anomalia) é muito inferior à quantidade de objetos da classe normal e, adicionalmente, o custo da não detecção de uma anomalia (**falso negativo**) é normalmente muito maior do que identificar um objeto normal como uma anomalia (**falso positivo**).

A detecção de anomalias em bases de dados é essencialmente um problema de classificação binária, no qual se deseja determinar se um ou mais objetos pertencem à classe normal ou à classe anômala. Assim, esse processo é muito similar ao fluxo convencional da tarefa de predição. Quanto ao custo de não detecção, vejamos um exemplo:

para uma operadora de cartão de crédito, autorizar uma transação fraudulenta é um falso negativo e constitui um prejuízo financeiro maior do que identificar uma transação normal como fraudulenta (falso positivo) e evitá-la ou contatar o cliente para confirmação. O mesmo vale para uma falha numa turbina de avião, pois não a detectar pode custar a vida de muitas pessoas; todavia, um alarme falso (falso positivo) pode causar atrasos e prejuízos financeiros, mas não custa vidas.

Gabarito: Letra D

---

122. (FGV / SEFAZ-ES – 2021) Maria está preparando um relatório sobrea as empresas de serviços de um município, de modo a identificar e estudar o porte dessas empresas com vistas ao estabelecimento de políticas públicas de previsões de arrecadações. Maria pretende criar nove grupos empresas, de acordo com os valores de faturamento, e recorreu às técnicas usualmente empregadas em procedimentos de data mining para estabelecer as faixas de valores de cada grupo. Assinale a opção que apresenta a técnica diretamente aplicável a esse tipo de classificação:

a) Algoritmos de associação.
b) Algoritmos de clusterização.
c) Árvores de decisão.
d) Modelagem de dados.
e) Regressão linear.

Comentários:

Ora, se ela quer criar nove grupos e empresas de acordo com os valores de faturamento e recorreu às técnicas de mineração de dados para estabelecer as faixas de valores de cada grupo significa que ela não tinha as classes previamente definidas. Logo, ela utilizou um algoritmo de aprendizado não supervisionado. No caso, trata-se claramente de algoritmos de clusterização.

Gabarito: Letra B

---

123. (FGV / DETRAN-RN – 2010) Sobre Data Mining, pode-se afirmar que:

a) Refere-se à implementação de banco de dados paralelos.
b) Consiste em armazenar o banco de dados em diversos computadores.
d) Relaciona-se à capacidade de processar grande volume de tarefas em um mesmo intervalo de tempo.
e) Permite-se distinguir várias entidades de um conjunto.
e) Refere-se à busca de informações relevantes a partir de um grande volume de dados.

Comentários:

Nenhum dos itens têm qualquer relação com Data Mining, exceto o último. O Data Mining (Mineração de Dados) se refere à busca de informações relevantes a partir de um grande volume de dados.

Gabarito: Letra E

---

124. (FGV / Senado Federal – 2008 – Letra A) Em Regras de Associação, confiança refere-se a quantas vezes uma regra de associação se verifica no conjunto de dados analisado.

Comentários:

Na verdade, a questão trata de Suporte e, não, Confiança.

Gabarito: Errado

QUESTÕES COMENTADAS – DIVERSAS BANCAS

125. (FEPESE / ISS-Criciúma – 2022) Quais tipos de conhecimento podem ser descobertos empregando técnicas clássicas de mineração de dados?

1. Regras de associação
2. Hierarquias de classificação
3. Padrões sequenciais ou de série temporal
4. Conhecimento implícito, emergente e não estruturado
5. Agrupamentos e segmentações.

Assinale a alternativa que indica todas as afirmativas **corretas.**

a) São corretas apenas as afirmativas 3 e 5.
b) São corretas apenas as afirmativas 1, 2, 3 e 4.
c) São corretas apenas as afirmativas 1, 2, 3 e 5.
d) São corretas apenas as afirmativas 2, 3, 4 e 5.
e) São corretas as afirmativas 1, 2, 3, 4 e 5.

Comentários:

Questão polêmica! Sendo rigoroso, a única alternativa que apresenta um tipo de conhecimento é a (4), quando fala de conhecimento implícito.

Analisando a questão com menos rigor técnico, podemos afirmar que – por meio de técnicas clássicas de mineração de dados – podemos descobrir conhecimentos, tais como: regras de associação (por meio da própria técnica de Regras de Associação), hierarquias de classificação (por meio da própria técnica de Classificação), padrões sequenciais ou de série temporal (por meio da própria técnica de Regras de Associação) e agrupamentos e segmentações (por meio da própria técnica de Análise de Agrupamento).

Por que não o conhecimento implícito, emergente e não estruturado? Porque se trata justamente de um conhecimento que não foi documentado, logo não pode ser minerado.

Gabarito: Letra C

---

126. (FEPESE / ISS-Criciúma – 2022) São técnicas de Inteligência Artificial de Data Mining:

1. Estatística.
2. Reconhecimento de Padrões.
3. Representação do Conhecimento.
4. Regras de Associação.

Assinale a alternativa que indica todas as afirmativas **corretas.**

a) São corretas apenas as afirmativas 2 e 3.
b) São corretas apenas as afirmativas 1, 2 e 3.
c) São corretas apenas as afirmativas 1, 2 e 4.
d) São corretas apenas as afirmativas 1, 3 e 4.
e) São corretas as afirmativas 1, 2, 3 e 4.

Comentários:

Questão polêmica! O gabarito preliminar considerou como técnica de IA apenas o reconhecimento de padrões (que permite reconhecer e avaliar padrões de acordo com o seu valor de interesse à medida que representam algum tipo de conhecimento relevante) e a representação do conhecimento (que permite representar e visualizar os dados extraídos e transformados, possibilitando extrair dados relevantes).

No entanto, a questão não considerou estatística e regras de associação como técnicas de inteligência artificial na mineração de dados. Eu não vejo como isso é possível! Na minha visão, ambas podem ser consideradas técnicas de inteligência artificial (em sentido amplo) utilizada na mineração de dados. Discordo do gabarito preliminar e aguardo uma retificação.

Gabarito: Letra A

---

127. (CESGRANRIO / BB – 2021) Um banco decidiu realizar uma ação de marketing de um novo produto. Buscando apoiar essa ação, a área de TI decidiu estabelecer um mecanismo para identificar quais clientes estariam mais inclinados a adquirir esse produto. Esse mecanismo partia de uma base histórica de clientes que receberam a oferta do produto, e tinha várias colunas com dados sobre os clientes e a oferta, além de uma coluna registrando se eles haviam efetuado ou não a compra do tal produto. Para isso, decidiram ser mais adequado usar um processo de mineração de dados baseado na noção de:

a) agrupamento
b) aprendizado não supervisionado
c) classificação
d) regressão linear
e) suavização

Comentários:

Vamos destacar as palavras-chave do enunciado: "base **histórica** de clientes", "coluna **registrando** se eles haviam efetuado ou não a compra do tal produto". Ora, temos uma base histórica de clientes e um registro se eles fizeram ou não a compra. Agora queremos prever se novos clientes vão

adquirir o produto em função das características registradas nas colunas. Como já temos, de antemão, quais serão as variáveis-alvo, trata-se de aprendizado supervisionado.

(a) Errado, trata-se de um método não-supervisionado; (b) Errado, o método que buscamos é supervisionado; (e) Errado, isso não é um método de mineração de dados. Logo, sobram duas opções de aprendizado supervisionado: classificação e regressão linear. No entanto, lembrem-se que a regressão linear retorna um valor numérico contínuo e, não, uma categoria discreta. Logo, trata-se do algoritmo de classificação, dado que o intuito é prever se o cliente vai ou não comprar o produto.

Gabarito: Letra C

---

128. (AOCP / MJSP – 2020) Dentre os métodos de mineração de dados, existem aqueles que são supervisionados e os não supervisionados. Assinale a alternativa que apresenta corretamente um dos métodos supervisionados mais comuns para a aplicação da mineração de dados que é voltado às tarefas frequentes do dia a dia:

a) Regras de associação.
b) Bubble sort.
c) Clusterização.
d) Classificação.
e) Formulação.

Comentários:

(a) Errado, trata-se de um método não-supervisionado; (b) Errado, trata-se de um método de ordenação; (c) Errado, trata-se de um método não-supervisionado; (d) Correto; (e) Errado, esse item não tem qualquer sentido no contexto de aprendizado de máquina.

Gabarito: Letra D

---

129. (IBADE / Prefeitura de Vila Velha – 2020) O processo de explorar grandes quantidades de dados a procura de padrões consistentes, para detectar relacionamentos sistemáticos entre variáveis, detectando assim novos subconjuntos de dados é chamado de:

a) Data Lake.
b) Big Data.
c) Data Query.
d) Data Warehouse.
e) Data Mining.

Comentários:

Vejam o tanto de palavras-chave para nos ajudar: explorar dados; procurar padrões consistentes; detectar relacionamentos sistemáticos; novos subconjuntos de dados – tudo isso nos remete ao processo de Mineração de Dados (Data Mining).

Gabarito: Letra E

---

130. (NC-UFPR / Itaipu – 2019) Os algoritmos de Mineração de Dados podem ser classificados quanto a seus objetivos, sendo alguns a classificação, o agrupamento e a identificação de regras de associação. A respeito dessas classificações e seus algoritmos, assinale a alternativa correta.

a) Algoritmos de agrupamento podem ser utilizados para classificação não supervisionada.
b) Algoritmos de agrupamento são também chamados de algoritmos supervisionados.
c) Algoritmos de classificação têm como resultado um modelo descritivo dos dados de entrada.
d) Algoritmos de identificação de regras são também conhecidos como algoritmos preditivos.
e) Algoritmos de agrupamento são equivalentes a algoritmos de identificação de anomalias.

Comentários:

(a) Correto, algoritmos de agrupamento realmente podem ser utilizados para classificação não supervisionada – como as classes não são previamente definidas, trata-se de um aprendizado não supervisionado; (b) Errado, algoritmos de agrupamento são não-supervisionados, enquanto os de classificação são supervisionados; (c) Errado, algoritmos de classificação têm como resultado um modelo preditivo dos dados de entrada; (d) Errado, algoritmos de identificação de regras são descritivos e, não, preditivos; (e) Errado, identificação de anomalias visa descobrir padrões em dados com um comportamento diferente do esperado, não sendo equivalentes aos algoritmos de agrupamento.

Gabarito: Letra A

---

131. (CESGRANRIO / BANCO DA AMAZÔNIA – 2018) As ferramentas e técnicas de mineração de dados (data mining) têm por objetivo:

a) preparar dados para serem utilizados em um "data warehouse" (DW).
b) permitir a navegação multidimensional em um DW.
c) projetar, de forma eficiente, o registro de dados transacionais.
d) buscar a classificação e o agrupamento (clusterização) de dados, bem como identificar padrões.
e) otimizar o desempenho de um gerenciador de banco de dados.

Comentários:

(a) Errado. Data Mining utiliza dados extraídos (geralmente) de Data Warehouses e, não, o contrário; (b) Errado. Esse é um objetivo de Ferramentas OLAP; (c) Errado, esse não é um objetivo

do Data Mining; (d) Correto. Ele realmente busca classificar e agrupar dados com o intuito de identificar padrões; (e) Errado, esse não é um objetivo do Data Mining.

Gabarito: Letra D

---

132. (COPESE / UFT – 2018 – Item III) Diversos modelos de Redes Neurais Artificiais podem ser utilizados na implementação de métodos de Mineração de Dados.

Comentários:

Perfeito! Redes Neurais (Artificiais) é realmente uma das técnicas de Mineração de Dados.

Gabarito: Correto

---

133. (FAURGS / FAURGS – 2018) Uma nuvem de palavras é um recurso gráfico (usado principalmente na internet) para descrever os termos mais frequentes de um determinado texto. O tamanho da fonte em que a palavra é apresentada é uma função da frequência da palavra no texto: palavras mais frequentes são desenhadas em fontes de tamanho maior, palavras menos frequentes são desenhadas em fontes de tamanho menor. Qual é a técnica de análise de dados descrita pelo texto acima?

a) Processamento de Linguagem Natural.
b) Agrupamento.
c) Classificação.
d) Redes Neurais.
e) Regressão Linear.

Comentários:

Nuvem de Palavras (Word Cloud) é um tipo de técnica de análise de dados de Processamento de Linguagem Natural (PLN).

Gabarito: Letra A

---

134. (CESGRANRIO / Petrobrás – 2018) Dois funcionários de uma empresa de crédito discutiam sobre quais algoritmos deveriam usar para ajudar a classificar seus clientes como bons ou maus pagadores. A empresa possui, para todos os empréstimos feitos no passado, um registro formado pelo conjunto de informações pessoais sobre o cliente e de como era composta a dívida inicial. Todos esses registros tinham classificações de bons ou maus pagadores, de acordo com o perfil de pagamento dos clientes. A partir desses dados, os funcionários querem construir um modelo, por meio de aprendizado de máquina, que classifique os novos clientes, que serão descritos por registros com o mesmo formato. A melhor opção, nesse caso, é usar um algoritmo:

a) supervisionado, como SVM.
b) supervisionado, como K-means.
c) não supervisionado, como regressão linear.
d) não supervisionado, como árvores de decisão.
e) semi-supervisionado, como redes bayesianas.

Comentários:

Vamos destacar as palavras-chave do enunciado: "**classificar** seus clientes como **bons ou maus pagadores**", "todos os empréstimos feitos no **passado**", "classifique **novos clientes**". Note que o objetivo dos funcionários da empresa é classificar um novo cliente como bom ou mau pagador com base em informações históricas de outros clientes. Como já conhecemos de antemão quais serão as classes que guiarão o aprendizado de máquina, trata-se de aprendizado supervisionado.

O algoritmo de aprendizado de máquina poderá inferir uma função a partir dos dados de treinamento previamente classificados de modo que seja possível prever uma classificação de saída (bom ou mau pagador) com base nos dados de entrada. Como já vimos, as principais técnicas de aprendizado supervisionado são: árvores de decisão, regressão linear, regressão logística, redes neurais, K-Nearest Neighbors, **Support Vector Machines (SVM)**, etc.

Dica: há muito mais algoritmos supervisionados do que algoritmos não-supervisionados, logo sugiro memorizar apenas que os principais algoritmos não-supervisionados são: K-Means, Agrupamento Hierárquico e Regras de Associação. Logo, se uma banca trouxer outro nome de algoritmo que não seja um desses dois, é muito provável que se trate de um algoritmo de aprendizado supervisionado porque essa classe de algoritmos tem bem mais opções.

Gabarito: Letra A

---

135. (ESAF / STN – 2018) Uma técnica de classificação em Mineração de Dados é uma abordagem sistemática para:

a) construção de controles de ordenação a partir de um conjunto de acessos.
b) construção de modelos de classificação a partir de um conjunto de dados de entrada.
c) construção de modelos de dados a partir de um conjunto de algoritmos.
d) construção de controles de ordenação independentes dos dados de entrada.
e) construção de modelos de sistemas de acesso a partir de um conjunto de algoritmos.

Comentários:

Uma técnica de classificação em Mineração de Dados é uma abordagem sistemática para a construção de modelos de classificação a partir de um conjunto de dados de entrada.

Gabarito: Letra B

---

136. (CESGRANRIO / TRANSPETRO – 2018) Um desenvolvedor recebeu um conjunto de dados representando o perfil de um grupo de clientes, sem nenhuma informação do tipo de cada cliente, onde cada um era representado por um conjunto fixo de atributos, alguns contínuos, outros discretos. Exemplos desses atributos são: idade, salário e estado civil. Foi pedido a esse desenvolvedor que, segundo a similaridade entre os clientes, dividisse os clientes em grupos, sendo que clientes parecidos deviam ficar no mesmo grupo. Não havia nenhuma informação que pudesse ajudar a verificar se esses grupos estariam corretos ou não nos dados disponíveis para o desenvolvedor. Esse é um problema de data mining conhecido, cuja solução mais adequada é um algoritmo:

a) de regressão
b) não supervisionado
c) por reforço
d) semissupervisionado
e) supervisionado

Comentários:

A questão nos dá diversas dicas para identificar a resposta. É um conjunto de dados que representa o perfil de um grupo, sendo que **não há informação do tipo de cada cliente**. Foi pedido ao desenvolvedor que, segundo a similaridade entre os clientes, dividisse em grupos sendo que clientes parecidos deviam ficar no mesmo grupo. O enunciado ainda insiste que não havia nenhuma informação que pudesse ajudar a verificar se esses grupos estariam corretos ou não nos dados disponíveis para o desenvolvedor. Ora, se é para dividir um grupo de clientes sem nenhuma informação prévia, trata-se de um algoritmo de aprendizado não supervisionado.

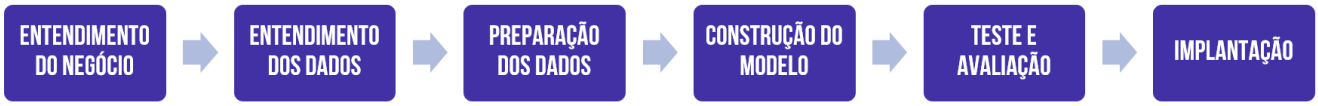
Gabarito: Letra B

---

137. (FEPESE / CIASC – 2017) Assinale a alternativa que contém as principais fases do processo de Data Minning CRISP-DM.

a) Amostragem; Exploração; Modificação; Modelagem; Execução; Avaliação.
b) Amostragem; Exploração; Modelagem; Modificação; Avaliação; Implementação.
c) Compreensão do negócio; Compreensão dos dados; Preparação dos dados; Modelagem; Avaliação; implementação.
d) Compreensão dos dados; Amostragem; Preparação dos dados; Implementação; Avaliação.
e) Compreensão do negócio; Exploração dos dados; Modificação dos dados; Implementação; Avaliação.

Comentários:



Em ordem, temos: Compreensão do negócio; Compreensão dos Dados; Preparação dos Dados; Modelagem; Avaliação; Implementação. Apesar de alguns nomes um pouco diferentes, essa é a opção correta.

Gabarito: Letra C

---

138. (NC-UFPR / Itaipu Binacional – 2015) Qual é a funcionalidade do Oracle Data Mining que encontra aglomerados de objetos de dados semelhantes em algum sentido entre si?

a) Aprior
b) Associação
c) Classificação
d) Clustering
e) Regressão

Comentários:

Encontrar aglomerados de objetos de dados semelhantes é chamado de *Clustering*.

Gabarito: Letra D

---

139. (AOCP / TCE-PA – 2014) O processo de explorar grandes quantidades de dados à procura de padrões consistentes com o intuito de detectar relacionamentos sistemáticos entre variáveis e novos subconjuntos de dados, é conhecido como:

a) Data Mart.
b) Data Exploring.
c) Objeto Relacional.
d) Relacionamento.
e) Data Mining.

Comentários:

O processo de explorar grandes quantidades de dados à procura de padrões consistentes com o intuito de detectar relacionamentos sistemáticos entre variáveis e novos subconjuntos de dados é conhecido como Data Mining.

Gabarito: Letra E

---

140. (VUNESP / TJ-PA – 2014) Uma das tarefas implementadas por uma ferramenta de Data Mining consiste em realizar a determinação de um valor futuro de determinada característica ou atributo de um registro ou conjunto de registros. Tal tarefa corresponde à:

a) normalização.
b) indexação.
c) análise de afinidade.
d) predição.
e) análise de equivalência

Comentários:

Tarefa de mineração de dados que permite determinar o valor futuro de determinada característica ou atributo de um registro é a predição, isto é, a técnica que busca descrever a natureza de ocorrências futuras de certos eventos com base nos acontecimentos passados.

Gabarito: Letra D

---

141. (FUNDEP / IFN/MG – 2014) Ao se utilizar a técnica de data mining (mineração de dados), como é conhecido o resultado dessa mineração, em que, por exemplo, se um cliente compra equipamento de vídeo, ele pode também comprar outros equipamentos eletrônicos?

a) Regras de associação
b) Padrões sequenciais
c) Árvores de classificação
d) Padrões de aquisição

Comentários:

O resultado de uma mineração de dados como o exemplo da questão é uma Regra de Associação.

Gabarito: Letra A

---

142. (FAURGS / TJ-RS – 2014) O resultado da mineração de dados pode ser a descoberta de tipos de informação "nova". Supondo-se que um cliente compre uma máquina fotográfica e que, dentro de três meses, compre materiais fotográficos, há probabilidade de que, dentro dos próximos seis meses, ele comprará um acessório. Um cliente que compra mais que duas vezes, em um período de baixa, deverá estar propenso a comprar, pelo menos uma vez, no período do Natal. Esse tipo de informação pode ser verificado através de:

a) predição de links.
b) regras de associação.
c) árvores de classificação.

d) árvores de decisão.
e) padrões sequenciais.

Comentários:

Trata-se de padrões sequenciais. *Por que não pode ser Regras de Associação?* Porque a questão afirma claramente que há uma sequência. *Por que não pode ser Padrões Temporais?* Se houvesse essa opção, ela seria a mais correta.

Gabarito: Letra E

---

143. (CESGRANRIO / LIQUIGÁS – 2014) As empresas possuem grandes quantidades de dados. Em geral, a maioria delas é incapaz de aproveitar plenamente o valor que eles têm. Com o intuito de melhorar essa situação, surgiu o data mining, que se caracteriza por:

a) desenhar padrões já conhecidos
b) extrair padrões ocultos nos dados.
c) tomar decisões para os gestores.
d) não trabalhar com tendências.
e) não trabalhar com associações.

Comentários:

(a) Errado, ele se caracteriza por desenhar padrões desconhecidos ou ocultos; (b) Correto; (c) Errado, ele se caracteriza por auxiliar a tomada de gestão por parte dos gestores; (d) Errado, ele se caracteriza por trabalhar com tendências; (e) Errado, ele se caracteriza por trabalhar com associações.

Gabarito: Letra B

---

144. CCV-UFC / UFC / 2013) Sobre Mineração de Dados, assinale a alternativa correta.

a) É uma técnica de organização de grandes volumes de dados.

b) É um conjunto de técnicas avançadas para busca de dados complexos.

c) É o processo de explorar grande quantidade de dados para extração não-trivial de informação implícita desconhecida.

d) É um processo automatizado para a recuperação de informações caracterizadas por registros com grande quantidade de atributos.

e) É um processo de geração de conhecimento que acontece durante o projeto de banco de dados. Os requisitos dos usuários são analisados e minerados para gerar as abstrações que finalmente são representadas em um modelo de dados.

Comentários:

(a) Errado, não se trata de organização de grandes volumes de dados; (b) Errado, não se trata de busca de dados complexos; (c) Correto, pode ser definido como o processo de explorar grande quantidade de dados para extração não-trivial de informação implícita desconhecida – isto é, busca de insights em uma grande quantidade de dados; (d) Errado, não se trata de um processo automatizado, mas semi-automatizado – além disso, não se trata de um processo de recuperação de informações, mas de descobertas de informações; (e) Errado, na verdade ele faz parte de um processo de geração de conhecimento, sendo uma de suas fases.

Gabarito: Letra C

---

145. (FMP CONCURSOS / MPE-AC – 2013) Processo de explorar grandes quantidades de dados à procura de padrões consistentes, como regras de associação ou sequências temporais, para detectar relacionamentos sistemáticos entre variáveis, detectando assim novos subconjuntos de dados é conhecido como:

a) datawarehouse.
b) SGBD.
c) mineração de dados (data mining).
d) modelagem relacional de dados.
e) mineração de textos (text mining).

Comentários:

Processo de explorar grandes quantidades de dados à procura de padrões consistentes, como regras de associação ou sequências temporais, para detectar relacionamentos sistemáticos entre variáveis, detectando assim novos subconjuntos de dados é conhecido como Mineração de Dados (Data Mining).

Gabarito: Letra C

---

146. (IBFC / EBSERH – 2013) Processo de explorar grandes quantidades de dados à procura de padrões consistentes, como regras de associação ou sequências temporais:

a) Data Warehouse
b) Data Mining
c) Tunning
d) APS (Application Platform Suite)

Comentários:

O processo de explorar grandes quantidades de dados à procura de padrões consistentes, como regras de associação ou sequências temporais é denominado Data Mining.

Gabarito: Letra B

---

147. (FUNRIO / MPOG – 2013) Qual o tipo de descoberta de conhecimento através de mineração de dados (do inglês "data mining"), em que se relaciona a presença de conjuntos de itens diversos, como por exemplo: "Quando uma mulher compra uma bolsa em uma loja, ela está propensa a comprar sapatos"?

a) Hierarquias de classificação.
b) Padrões sequenciais.
c) Regras de associação.
d) Séries temporais.
e) Agrupamentos por similaridade.

Comentários:

Regras de Associação são regras que correlacionam a presença de um conjunto de itens com outra faixa de valores para um conjunto de variáveis diverso. A correção entre a compra de bolsas e sapatos é uma regra de associação.

Gabarito: Letra C

---

148. (ESPP / MPE-PR – 2013) Data Mining refere-se à busca de informações relevantes, ou "à descoberta de conhecimento", a partir de um grande volume de dados. Assim como a descoberta de conhecimento no ramo da inteligência artificial, a extração de dados tenta descobrir automaticamente modelos estatísticos a partir dos dados. O conhecimento obtido a partir de um banco de dados pode ser representado em regras. Duas importantes classes de problemas de extração de dados são as:

a) regras de indexação e regras de população.
b) regras de validação e regras de otimização.
c) regras de interpolação e regras de valoração.
d) regras de maximização e regras de generalização.
e) regras de classificação e regras de associação.

Comentários:

Essa redação é um pouco estranha, mas podemos inferir que a questão trata de Regras de Classificação e Regras de Associação.

Gabarito: Letra E

---

149. (ESAF / MF – 2013) A Mineração de Dados requer uma adequação prévia dos dados através de técnicas de pré-processamento. Entre elas estão as seguintes técnicas:

a) Agrupamento. Amostragem. Redução de dimensionalidade. Seleção de subconjuntos de recursos. Recursos pontuais. Polarização. Redução de variáveis.

b) Agregação. Classificação. Redução de faixas de valores. Seleção de subconjuntos de recursos. Redução de recursos. Terceirização e discretização. Transformação de variáveis.

c) Agrupamento. Classificação. Redução de dimensionalidade. Seleção de subconjuntos de usuários. Criação de recursos. Binarização e discretização. Transformação de conjuntos.

d) Agregação. Amostragem. Redução de dimensionalidade. Seleção de subconjuntos de usuários. Criação de recursos. Polarização. Transformação de conjuntos.

e) Agregação. Amostragem. Redução de dimensionalidade. Seleção de subconjuntos de recursos. Criação de recursos. Binarização e discretização. Transformação de variáveis.

Comentários:

Trata-se da Agregação, Amostragem, Redução de dimensionalidade, Seleção de subconjuntos de recursos, Criação de recursos, Binarização e discretização, e Transformação de variáveis.

Gabarito: Letra E

---

150. (IADES / EBSERH – 2012) Existem algumas técnicas utilizadas em Data mining, para fins de estatísticas. A técnica que permite lidar com a previsão de um valor, em vez de uma classe, é denominada:

a) associação.
b) exploração.
c) classificação.
d) regressão.
e) árvore de decisão.

Comentários:

A técnica que permite lidar com a previsão de um valor real em vez de uma classe é a Regressão.

Gabarito: Letra D

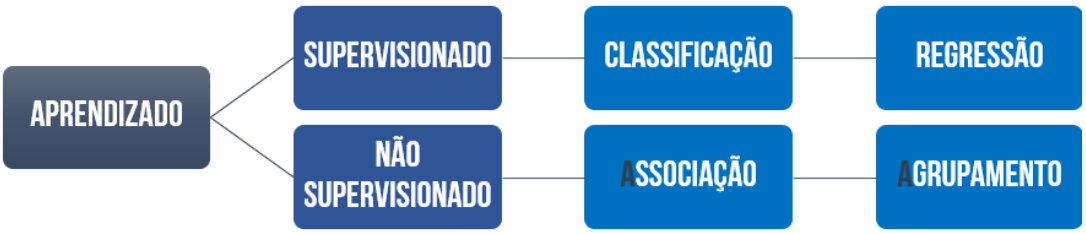
---

151. (CESGRANRIO / EPE – 2012) As técnicas de mineração de dados podem ser categorizadas em supervisionadas e não supervisionadas. As técnicas de árvores de decisão, agrupamento e regras de associação são categorizadas, respectivamente, como:

a) não supervisionada, não supervisionada, não supervisionada
b) não supervisionada, supervisionada e não supervisionada
c) supervisionada, não supervisionada e não supervisionada
d) supervisionada, não supervisionada e supervisionada
e) supervisionada, supervisionada e supervisionada

Comentários:

Árvore de Decisão (que é um tipo de Classificação) é supervisionada; Agrupamento é não supervisionada; e regras de associação é não supervisionada.



Gabarito: Letra C

---

152. (FMP CONCURSOS / TCE-RS – 2011) Mineração de dados consiste em:

a) explorar um conjunto de dados visando a extrair ou a ajudar a evidenciar padrões, como regras de associação ou sequências temporais, para detectar relacionamentos entre estes.

b) acessar um banco de dados para realizar consultas de forma genérica, buscando recuperar informações (registros) que atendam um mesmo critério de pesquisa.

c) recuperar informações de um banco de dados específico, voltado a representar e armazenar dados relacionados com companhias de exploração petrolífera e de recursos mineralógicos.

d) um banco de dados específico voltado à gestão de negócios usando tecnologia de informação (TI) como, por exemplo, a área de BI (Business Inteligence).

e) representar informações de um banco de dados mediante vários modelos hierárquicos como, por exemplo, o de entidade-relacionamento (ER).

Comentários:

(a) Correto. Explorar um conjunto de dados visando a extrair ou a ajudar a evidenciar padrões, como regras de associação ou sequências temporais, para detectar relacionamentos entre estes; (b) Errado. Realizam-se pesquisas de forma específica e, não, genérica; (c) Errado. *Companhias de exploração petrolífera e recursos mineralógicos?* Hahaha! Galera, mineração não tem nada a ver com minas em nosso contexto; (d) Errado. Data Mining não é um banco de dados; (e) Errado. Esse item não tem qualquer relação com Data Mining.

Gabarito: Letra A

---

153. (FUMARC / PRODEMGE – 2011) Analise as afirmativas abaixo em relação às técnicas de mineração de dados.

I. Regras de associação podem ser usadas, por exemplo, para determinar, quando um cliente compra um produto X, ele provavelmente também irá comprar um produto Y.

II. Classificação é uma técnica de aprendizado supervisionado, no qual se usa um conjunto de dados de treinamento para aprender um modelo e classificar novos dados.

III. Agrupamento é uma técnica de aprendizado supervisionado que particiona um conjunto de dados em grupos.

Assinale a alternativa VERDADEIRA:

a) Apenas as afirmativas I e II estão corretas.
b) Apenas as afirmativas I e III estão corretas.
c) Apenas as afirmativas II e III estão corretas.
d) Todas as afirmativas estão corretas.

Comentários:

(I) Correto. As regras de associação consistem em identificar fatos que possam ser direta ou indiretamente associados; (II) Correto. Trata-se de uma técnica de aprendizado supervisionado, isto é, as classes são pré-definidas antes da análise dos resultados. Além disso, essa técnica realmente usa um conjunto de dados de treinamento para aprender um modelo e classificar novos dados; (III) Errado. Agrupamento é uma técnica de aprendizado não-supervisionado.

Gabarito: Letra A

---

154. (FMP CONCURSOS / TCE/RS – 2011 – Letra B) Mineração de Dados é parte de um processo maior de pesquisa chamado de Busca de Conhecimento em Banco de Dados (KDD).

Comentários:

Ele realmente é parte de um processo maior de pesquisa chamado de busca ou descoberta de conhecimento em bancos de dados.

Gabarito: Correto

---

155. (ESAF / CVM – 2010) Mineração de Dados é:

a) o processo de atualizar de maneira semi-automática grandes bancos de dados para encontrar versões úteis.

b) o processo de analisar de maneira semi-automática grandes bancos de dados para encontrar padrões úteis.

c) o processo de segmentar de maneira semi-automática bancos de dados qualitativos e corrigir padrões de especificação.

d) o programa que depura de maneira automática bancos de dados corporativos para mostrar padrões de análise.

e) o processo de automatizar a definição de bancos de dados de médio porte de maior utilidade para os usuários externos de rotinas de mineração.

Comentários:

(a) Errado. Mineração de Dados é o processo de ~~atualizar~~ analisar de maneira semi-automática grandes bancos de dados para encontrar versões úteis;

(b) Correto. Mineração de Dados é o processo de analisar de maneira semi-automática grandes bancos de dados para encontrar versões úteis;

(c) Errado. Mineração de Dados é o processo de ~~segmentar~~ analisar de maneira semi-automática grandes bancos de dados para encontrar versões úteis;

(d) Errado. Mineração de Dados não realiza depurações (que é o processo de encontrar defeitos em um software ou hardware);

(e) Errado. Mineração de Dados não tem nenhuma relação com automatizar a definição de bancos de dados de médio porte de maior utilidade para os usuários externos de rotinas de mineração.

Gabarito: Letra B

---

156. (ESAF / MPOG – 2010) Mineração de Dados:

a) é uma forma de busca sequencial de dados em arquivos.
b) é o processo de programação de todos os relacionamentos e algoritmos existentes nas bases de dados.
c) por ser feita com métodos compiladores, método das redes neurais e método dos algoritmos gerativos.
d) engloba as tarefas de mapeamento, inicialização e clusterização.
e) engloba as tarefas de classificação, regressão e clusterização.

Comentários:

Nenhum dos itens faz qualquer sentido, exceto o último. A mineração de dados realmente engloba as tarefas de classificação, regressão e clusterização.

Gabarito: Letra E

---

157. (CESGRANRIO / ELETROBRÁS – 2010) Em uma reunião sobre prospecção de novos pontos de venda, um analista de TI afirmou que técnicas OLAP de análise de dados são orientadas a oferecer informações, assinalando detalhes intrínsecos e facilitando a agregação de valores, ao passo que técnicas de data mining tem como objetivo:

a) captar, organizar e armazenar dados colecionados a partir de bases transacionais, mantidas por sistemas OLTP.

b) facilitar a construção de ambientes de dados multidimensionais, através de tabelas fato e dimensionais.

c) melhorar a recuperação de dados organizados de forma não normalizada em uma base relacional conhecida como data warehouse.

d) extrair do data warehouse indicadores de controle (BSC) para apoio à tomada de decisão por parte da diretoria da empresa.

e) identificar padrões e recorrência de dados, oferecendo conhecimento sobre o comportamento dos dados analisados.

Comentários:

(a) Errado, esse não é um dos objetivos de mineração de dados – isso se parece mais com ETL; (b) Errado, esse não é um dos objetivos de mineração de dados – isso se parece mais com modelagem multidimensional; (c) Errado, esse não é um dos objetivos de mineração de dados; (d) Errado, mineração de dados não ocorre necessariamente em um Data Warehouse – além disso, ela não tem o objetivo de extrair indicadores de controle; (e) Correto, o principal objetivo da mineração de dados

é identificar padrões ocultos e recorrência de dados, oferecendo conhecimento sobre o comportamento dos dados analisados que possam auxiliar na tomada de decisão por parte dos gestores de uma organização.

Gabarito: Letra E

---

158. (UFF / UFF – 2009) O conjunto de técnicas que, envolvendo métodos matemáticos e estatísticos, algoritmos e princípios de inteligência artificial, tem o objetivo de descobrir relacionamentos significativos entre dados armazenados em repositórios de grandes volumes e concluir sobre padrões de comportamento de clientes de uma organização é conhecido como:

a) Datawarehouse;
b) Metadados;
c) Data Mart;
d) Data Mining;
e) Sistemas Transacionais.

Comentários:

O conjunto de técnicas que, envolvendo métodos matemáticos e estatísticos, algoritmos e princípios de inteligência artificial, tem o objetivo de descobrir relacionamentos significativos entre dados armazenados em repositórios de grandes volumes e concluir sobre padrões de comportamento de clientes de uma organização é conhecido como Data Mining.

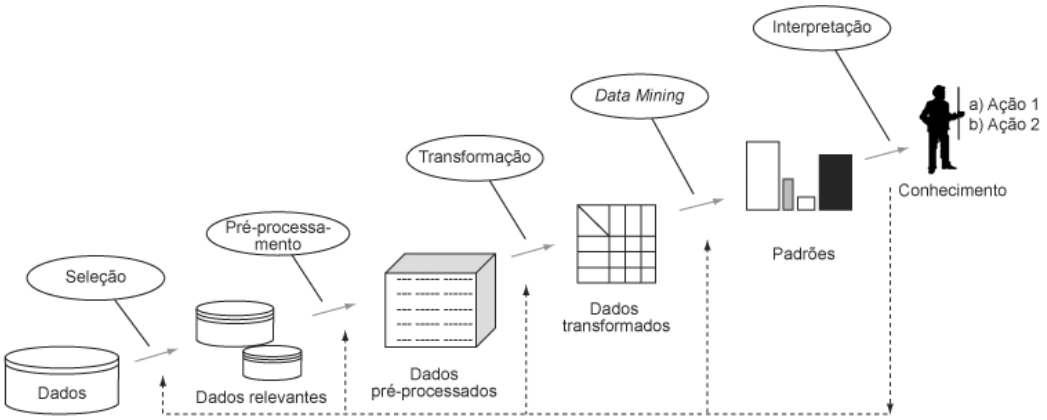
Gabarito: Letra D

---

159. (COSEAC / DATAPREV – 2009) "*Mining é parte de um processo maior de conhecimento, que o processo consiste, fundamentalmente, na estruturação do banco de dados; na seleção, preparação e pré-processamento dos dados; na transformação, adequação e redução da dimensionalidade dos dados; e nas análises, assimilações, interpretações e uso do conhecimento extraído do banco de dados*". O processo maior citado no início do texto é denominado:

a) Data mining
b) Data mart
c) Data warehouse
d) KDD
e) Segmentação de dados

Comentários:



Parte de um processo maior de conhecimento? Consiste na estruturação, seleção, preparação, pré-processamento de dados? Batava lembrar da nossa figurinha sobre o Processo de Descoberta do Conhecimento em Bancos de Dados (KDD).

Gabarito: Letra D

QUESTÕES COMENTADAS – CESPE

1. **(CESPE / MPO – 2024)** A regressão tem como objetivo a obtenção de uma equação que relacione uma variável de resposta a uma ou mais variáveis explicativas.

2. **(CESPE / ANTT – 2024)** Os algoritmos de regras de associação constroem regras com apenas uma única conclusão, ao contrário dos algoritmos de árvore de decisão, que tentam localizar muitas regras, cada uma delas com uma conclusão diferente.

3. **(CESPE / CTI – 2024)** Clustering é uma técnica de mineração de dados que agrupa dados não rotulados com base em suas semelhanças ou diferenças; os algoritmos de cluster podem ser categorizados em sobrepostos, hierárquicos ou probabilísticos.

4. **(CESPE / DATAPREV – 2023)** As técnicas de regressão utilizam um conjunto finito de hipóteses para, a partir dos atributos previsores, determinar a categoria de um objeto do conjunto de dados analisado.

5. **(CESPE / DATAPREV – 2023)** A regra de associação é uma técnica que busca relações de co-ocorrência entre objetos de uma base de dados.

6. **(CESPE / AGER-MT – 2023)** Em machine learning, quando algoritmos de aprendizado de máquina são usados para analisar e agrupar conjuntos de dados não rotulados, de forma tal que os algoritmos descobrem padrões ocultos sem a necessidade de intervenção humana, usa-se a forma de aprendizado do tipo:

   a) não supervisionado.
   b) supervisionado.
   c) over fitting.
   d) under fitting.
   e) classificação.

7. **(CESPE / SEFIN de Fortaleza-CE – 2023)** Aprendizado de máquina é um subcampo da inteligência artificial que consiste no treinamento de modelos computacionais para que possam reconhecer padrões e, a partir de um conjunto de dados de entrada, prever o valor de uma variável de saída. Em relação ao aprendizado de máquina, julgue o item a seguir.

   Em aprendizado de máquina, as características de entrada e saída são definidas, respectivamente, como atributos previsores e atributos alvo ou meta.

8. **(CESPE / SEFIN de Fortaleza-CE – 2023)** Nos algoritmos de aprendizado por reforço, o agente recebe uma recompensa atrasada na próxima etapa de tempo para avaliar sua ação anterior; seu objetivo, então, é maximizar a recompensa.

9. **(CESPE / DATAPREV – 2023)** Um sistema de aprendizado não supervisionado, dotado de um conjunto de dados de treinamento que foram classificados manualmente, tenta aprender, a partir desses dados de treinamento, uma forma de classificá-los, bem como de classificar novos dados, ainda não observados.

10. **(CESPE / CNMP - 2023)** O *data mining* é um processo usado para extrair e analisar informações que revelam padrões ou tendências estratégicas do negócio.

11. **(CESPE / TRT8 – 2022)** Acerca de modelos preditivos e descritivos, assinale a opção correta:

a) Com um modelo não supervisionado consegue-se construir um estimador a partir de exemplos rotulados.

b) Um modelo supervisionado refere-se à identificação de informações relevantes nos dados sem a presença de um elemento externo para orientar o aprendizado.

c) Com o uso de técnicas do modelo não supervisionado, consegue-se prever com exatidão o resultado de uma eleição utilizando pesquisas como parâmetro.

d) A análise de agrupamento pertence ao paradigma de aprendizado não supervisionado, em que o aprendizado é dirigido aos dados, não requerendo conhecimento prévio sobre as suas classes ou categorias.

e) Tendo como objetivo encontrar padrões ou tendências para auxiliar o entendimento dos dados, deve-se usar técnicas do modelo supervisionado.

12. **(CESPE / ISS-Aracaju – 2021)** Em um projeto de data mining, a coleta do dado que será garimpado ocorre no processo de:

a) mineração.
b) preparação.
c) aplicação.
d) associação.
e) classificação.

13. **(CESPE / ISS-Aracaju – 2021)** De acordo com o modelo CRSP-DM, a seleção das técnicas que serão aplicadas nos dados selecionados ocorre na fase de:

a) modelagem.
b) entendimento dos dados.
c) entendimento do negócio.
d) avaliação.
e) preparação dos dados.

14. **(CESPE / ISS-Aracaju – 2021)** O enriquecimento de dados da etapa de pré-processamento e preparação do data mining tem como objetivo:

a) a deduplicidade de registros.
b) a seleção de amostras.
c) a integração de bases diferentes.
d) o tratamento de valores nulos.
e) o acréscimo de dados à base já existentes.

15. **(CESPE / PCDF – 2021)** Uma das aplicações de Python é o aprendizado de máquina, que pode ser exemplificado por um programa de computador que aprende com a experiência de detectar imagens de armas e de explosivos em vídeos, tendo seu desempenho medido e melhorando por meio dos erros e de acertos decorrentes da experiência de detecção.

16. **(CESPE / PCDF – 2021)** A detecção de novos tipos de fraudes é uma das aplicações comuns da técnica de modelagem descritiva da mineração de dados, a que viabiliza o mapeamento rápido e preciso de novos tipos de golpes por meio de modelos de classificação de padrões predefinidos de fraudes.

17. **(CESPE / APEX – 2021)** m data mining revela informações que consultas manuais não poderiam revelar efetivamente. Por exemplo, em data mining, o algoritmo de classificação permite:

a) dividir o banco de dados em segmentos cujos membros compartilhem características iguais por meio de redes neurais.

b) analisar os dados históricos armazenados em um banco de dados e gerar automaticamente um modelo que possa prever comportamentos futuros.

c) mapear dados por meio da técnica estatística e, assim, obter um valor de previsão a partir de técnicas de regressão linear e não linear.

d) estabelecer relações entre itens que estejam juntos em determinado registro, o que é conhecido como análise de cesta de compras.

18. **(CESPE / TCE-RJ – 2021)** 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.

19. **(CESPE / TCE-RJ – 2021)** A descoberta de conhecimento em bases de dados, ou KDD (Knowledge-Discovery), é a etapa principal do processo de mineração de dados.

20. **(CESPE / TCE-RJ – 2021)** Na mineração de dados preditiva, ocorre a geração de um conhecimento obtido de experiências anteriores para ser aplicado em situações futuras.

21. **(CESPE / TCE-RJ – 2021)** As regras de associação adotadas em mineração de dados buscam padrões frequentes entre 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 potenciais produtos de interesse.

22. **(CESPE / TCE-RJ – 2021)** 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.

23. **CESPE / TCE-RJ – 2021)** No método de classificação para mineração de dados, a filiação dos objetos é obtida por meio de um processo não supervisionado de aprendizado, em que somente as variáveis de entrada são apresentadas para o algoritmo.

24. **(CESPE / TCE-RJ – 2021)** No método de mineração de dados por agrupamento (clustering), são utilizados algoritmos com heurísticas para fins de descoberta de agregações naturais entre objetos.

25. **(CESPE / TCE-RJ – 2021)** O fator de suporte e o fator de confiança são dois índices utilizados para definir o grau de certeza de uma regra de associação.

26. **(CESPE / TCE-RJ – 2021)** Os principais métodos de análise de agrupamentos em mineração de dados incluem redes neurais, lógica difusa, métodos estatísticos e algoritmos genéticos.

27. **(CESPE / Polícia Federal – 2021)** A análise de *clustering* é uma tarefa que consiste em agrupar um conjunto de objetos de tal forma que estes, juntos no mesmo grupo, sejam mais semelhantes entre si que em outros grupos.

28. **(CESPE / ME – 2020)** Aprendizagem de máquina pode ajudar a clusterização na identificação de outliers, que são objetos completamente diferentes do padrão da amostra.

29. **(CESPE / ME – 2020)** A técnica de associação é utilizada para indicar um grau de afinidade entre registros de eventos diferentes, para permitir o processo de data mining.

30. **(CESPE / ME – 2020)** No modelo CRISP-DM, a fase na qual se planejam todas as atividades para carga dos dados é denominada entendimento dos dados.

31. **(CESPE / ME – 2020)** Na etapa de mineração do data mining, ocorre a seleção dos conjuntos de dados que serão utilizados no processo de mining.

32. **(CESPE / Ministério da Economia – 2020)** A técnica de agregação na mineração de dados atua em conjunto de registros que tenham sido previamente classificados.

33. **(CESPE / Ministério da Economia – 2020)** O objetivo da etapa de pré-processamento é diminuir a quantidade de dados que serão analisados, por meio da aplicação de filtros e de eliminadores de palavras.

34. **(CESPE / Ministério da Economia – 2020)** Modelagem preditiva é utilizada para antecipar comportamentos futuros, por meio do estudo da relação entre duas ou mais variáveis.

35. **(CESPE / ME – 2020)** Outlier ou anomalias são padrões nos dados que não estão de acordo com uma noção bem definida de comportamento normal.

36. **(CESPE / ME – 2020)** A análise de regressão em mineração de dados tem como objetivos a sumarização, a predição, o controle e a estimação.

37. **(CESPE / TJ-AM – 2019)** A técnica machine learning pode ser utilizada para apoiar um processo de data mining.

38. **(CESPE / POLÍCIA FEDERAL – 2018)** Pode-se definir mineração de dados como o processo de identificar, em dados, padrões válidos, novos, potencialmente úteis e, ao final, compreensíveis.

39. **(CESPE / FUB – 2018)** No Data Mining, uma regra de associação relaciona a presença de um conjunto de itens com outra faixa de valores de um outro conjunto de variáveis.

40. **(CESPE / Polícia Federal – 2018)** A mineração de dados se caracteriza especialmente pela busca de informações em grandes volumes de dados, tanto estruturados quanto não estruturados, alicerçados no conceito dos 4V's: volume de mineração, variedade de algoritmos, velocidade de aprendizado e veracidade dos padrões.

41. **(CESPE / Polícia Federal – 2018)** Descobrir conexões escondidas e prever tendências futuras é um dos objetivos da mineração de dados, que utiliza a estatística, a inteligência artificial e os algoritmos de aprendizagem de máquina.

42. **(CESPE / Polícia Federal – 2018)** Situação hipotética: Na ação de obtenção de informações por meio de aprendizado de 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 aprendizagem supervisionados. Assertiva: Nessa situação, a ação em realização está relacionada ao processo de classificação.

CPF
NOME
DATA DE NASCIMENTO
NOME DO PAI
NOME DA MAE
TELEFONE

CEP
NUMERO

As informações anteriormente apresentadas correspondem aos campos de uma tabela de um banco de dados, a qual é acessada por mais de um sistema de informação e também por outras tabelas. Esses dados são utilizados para simples cadastros, desde a consulta até sua alteração, e também para prevenção à fraude, por meio de verificação dos dados da tabela e de outros dados em diferentes bases de dados ou outros meios de informação.

Considerando essas informações, julgue o item que se segue.

43. (CESPE / Polícia Federal – 2018) Se um sistema de informação correlaciona os dados da tabela em questão com outros dados não estruturados, então, nesse caso, ocorre um processo de mineração de dados.

44. (CESPE / EBSERH – 2018) A descoberta de novas regras e padrões em conjuntos de dados fornecidos, ou aquisição de conhecimento indutivo, é um dos objetivos de data mining.

45. (CESPE / IPHAN – 2018) Na busca de padrões no data mining, é comum a utilização do aprendizado não supervisionado, em que um agente externo apresenta ao algoritmo alguns conjuntos de padrões de entrada e seus correspondentes padrões de saída, comparando-se a resposta fornecida pelo algoritmo com a resposta esperada.

46. (CESPE / EBSERH – 2018) A descoberta de novas regras e padrões em conjuntos de dados fornecidos, ou aquisição de conhecimento indutivo, é um dos objetivos de data mining.

47. (CESPE / STJ – 2018) O processo de mineração de dados está intrinsecamente ligado às dimensões e a fato, tendo em vista que, para a obtenção de padrões úteis e relevantes, é necessário que esse processo seja executado dentro dos data warehouses.

48. (CESPE / TCM-BA – 2018) A respeito das técnicas e(ou) métodos de mineração de dados, assinale a opção correta.

a) O agrupamento (ou clustering) realiza identificação de grupos de dados que apresentam coocorrência.

b) A classificação realiza o aprendizado de uma função que pode ser usada para mapear os valores associados aos dados em um ou mais valores reais.

c) A regressão ou predição promove o aprendizado de uma função que pode ser usada para mapear dados em uma de várias classes discretas definidas previamente, bem como encontrar tendências que possam ser usadas para entender e explorar padrões de comportamento dos dados.

d) As regras de associação identificam grupos de dados, em que os dados têm características semelhantes aos do mesmo grupo e os grupos têm características diferentes entre si.

e) Os métodos de classificação supervisionada podem ser embasados em separabilidade (entropia), utilizando árvores de decisão e variantes, e em particionamento, utilizando SVM (support vector machines).

49. **(CESPE / TCM-BA – 2018)** Assinale a opção correta a respeito do CRISP-DM.

a) CRISP-DM é uma suíte de ferramentas proprietárias que vem se tornando um padrão da indústria para mineração de dados, uma vez que fornece um plano completo e tecnologias para a realização de um projeto de mineração de dados.

b) A verificação da qualidade dos dados é uma atividade da fase de entendimento dos dados.

c) Durante a fase de preparação dos dados, é realizado um inventário de requisitos, suposições e restrições de recursos.

d) Na fase de avaliação dos dados, são realizadas as atividades de identificar valores especiais dos dados e catalogar seu significado.

e) Na fase de preparação dos dados, são realizadas as atividades de analisar o potencial de implantação de cada resultado e estimar o potencial de melhoria do processo atual.

50. **(CESPE / SEDF – 2017)** Agrupar registros em grupos, de modo que os registros em um grupo sejam semelhantes entre si e diferentes dos registros em outros grupos é uma maneira de descrever conhecimento descoberto durante processos de mineração de dados.

51. **(CESPE / FUNPRESP-EXE – 2016)** Na implementação de mineração de dados (data mining), a utilização da técnica de padrões sequenciais pode ser útil para a identificação de tendências.

52. **(CESPE / TJ/SE – 2016)** DataMining pode ser considerado uma etapa no processo de descoberta de conhecimento em base de dados, consistindo em análise de conjuntos de dados cujo objetivo é descobrir padrões úteis para tomada de decisão.

53. **(CESPE / FUNPRESP/JUD – 2016)** Em DataMining, as árvores de decisão podem ser usadas com sistemas de classificação para atribuir informação de tipo.

54. **(CESPE / TRT-18ª Região – 2016)** Acerca de data mining, assinale a opção correta.

a) A fase de preparação para implementação de um projeto de data mining consiste, entre outras tarefas, em coletar os dados que serão garimpados, que devem estar exclusivamente em um data warehouse interno da empresa.

b) As redes neurais são um recurso matemático/computacional usado na aplicação de técnicas estatísticas nos processos de data mining e consistem em utilizar uma massa de dados para criar e organizar regras de classificação e decisão em formato de diagrama de árvore, que vão classificar seu comportamento ou estimar resultados futuros.

c) As aplicações de data mining utilizam diversas técnicas de natureza estatística, como a análise de conglomerados (cluster analysis), que tem como objetivo agrupar, em diferentes conjuntos de dados, os elementos identificados como semelhantes entre si, com base nas características analisadas.

d) As séries temporais correspondem a técnicas estatísticas utilizadas no cálculo de previsão de um conjunto de informações, analisando-se seus valores ao longo de determinado período. Nesse caso, para se obter uma previsão mais precisa, devem ser descartadas eventuais sazonalidades no conjunto de informações.

e) Os processos de data mining e OLAP têm os mesmos objetivos: trabalhar os dados existentes no data warehouse e realizar inferências, buscando reconhecer correlações não explícitas nos dados do data warehouse.

55. **(CESPE / TCE-PA – 2016)** No contexto de data mining, o processo de descoberta de conhecimento em base de dados consiste na extração não trivial de conhecimento previamente desconhecido e potencialmente útil.

56. **(CESPE / MEC – 2015)** O conhecimento obtido no processo de data mining pode ser classificado como uma regra de associação quando, em um conjunto de eventos, há uma hierarquia de tuplas sequenciais.

57. **(CESPE / MEC – 2015)** Situação hipotética: Após o período de inscrição para o vestibular de determinada universidade pública, foram reunidas informações acerca do perfil dos candidatos, cursos inscritos e concorrências. Ademais, que, por meio das soluções de BI e DW que integram outros sistemas, foram realizadas análises para a detecção de relacionamentos sistemáticos entre as informações registradas. Assertiva: Nessa situação, tais análises podem ser consideradas como data mining, pois agregam valor às decisões do MEC e sugerem tendências, como, por exemplo, o aumento no número de escolas privadas e a escolha de determinado curso superior.

58. **(CESPE / MEC – 2015)** Os objetivos do Data Mining incluem identificar os tipos de relacionamentos que se estabelecem entre informações armazenadas em um grande repositório.

59. **(CESPE / MEC – 2015)** Algoritmo genético é uma das ferramentas do *data mining* que utiliza mecanismos de biologia evolutiva, como hereditariedade, recombinação, seleção natural e mutação, para solucionar e agrupar problemas.

60. **(CESPE / MEC – 2015)** A predição em algoritmos de *data mining* objetiva modelar funções sobre valores para apresentar o comportamento futuro de determinados atributos.

61. **(CESPE / MEC – 2015)** Selecionar uma amostra e determinar os conjuntos de itens frequentes dessa amostra para formar a lista de previsão de subconjunto são as principais características do algoritmo de previsão.

62. **(CESPE / TCU – 2015)** A finalidade do uso do data mining em uma organização é subsidiar a produção de afirmações conclusivas acerca do padrão de comportamento exibido por agentes de interesse dessa organização.

63. **(CESPE / TCU – 2015)** Quem utiliza o data mining tem como objetivo descobrir, explorar ou minerar relacionamentos, padrões e vínculos significativos presentes em grandes massas documentais registradas em arquivos físicos (analógicos) e arquivos lógicos (digitais).

64. **(CESPE / TCU – 2015)** O uso prático de data mining envolve o emprego de processos, ferramentas, técnicas e métodos oriundos da matemática, da estatística e da computação, inclusive de inteligência artificial.

65. **(CESPE / DEPEN – 2015)** Os objetivos do *datamining* incluem identificar os tipos de relacionamentos que se estabelecem entre informações armazenadas em um grande repositório.

66. **(CESPE / ANTAQ – 2014)** Em um processo de descoberta do conhecimento, um Data Mining executado para atingir uma meta pode falhar nas classes de predição, de identificação, de classificação e de otimização.

67. **(CESPE / TCDF – 2014)** Com o uso da classificação como técnica de Data Mining, busca-se a identificação de uma classe por meio de múltiplos atributos. Essa técnica também pode ser usada em conjunto com outras técnicas de mineração de dados.

68. **(CESPE / TJ-SE – 2014)** O uso de agrupamento (clustering) em DataMining exige que os registros sejam previamente categorizados, tendo por finalidade aproximar registros similares para predizer valores de variáveis.

69. **(CESPE / TJ-SE – 2014)** Assim como o DataMining, os DataMarts são voltados para a obtenção de informações estratégicas de maneira automática, ou seja, com o mínimo de intervenção humana a partir da análise de dados oriundos de DataWarehouses.

70. **(CESPE / ANATEL – 2014)** No processo de Data Mining (mineração de dados), é indispensável o uso de técnica conhecida como Data Warehousing, uma vez que a mineração de dados deve ocorrer necessariamente em estruturas não normalizadas (FNo).

71. **(CESPE / TJ-SE – 2014)** Os principais processos de DataMining são a identificação de variações embasado em normas, a detecção e análise de relacionamentos, a paginação de memória e o controle de periféricos.

72. **(CESPE / TJ-CE – 2014)** Assinale a opção correta acerca de datamining:

a) A informação acerca dos resultados obtidos no processo de mineração é apresentada apenas de forma gráfica.

b) A classificação, uma das principais tecnologias da mineração de dados, caracteriza-se por possuir um conjunto de transações, sendo cada uma delas relacionada a um itemset.

c) É possível realizar mineração de dados em documentos textuais como, por exemplo, uma página da Internet.

d) A grande desvantagem de um datamining consiste no fato de que a identificação de um padrão, para a geração do conhecimento, só é possível por meio da análise em pequenas quantidades de dados.

e) Durante a fase de reconhecimento de padrões, para cada banco de dados, é permitido um único tipo de padrão.

73. **(CESPE / MPOG – 2013)** ETL é definido como o processo de descobrir padrões, associações, mudanças, anomalias e estruturas em grandes quantidades de dados armazenados ou em repositórios de informação gerais dentro do data mining.

74. **(CESPE / SERPRO – 2013)** Datamining é a tecnologia por intermédio da qual os processos são automatizados mediante racionalização e potencialização por meio de dois componentes: organização e tecnologia.

75. **(CESPE / TJ-RO – 2012)** A técnica de associação em data mining verifica se há controle ou influência entre atributos ou valores de atributos, no intuito de verificar, mediante a análise de probabilidades condicionais, dependências entre esses atributos.

76. **(CESPE / PEFOCE – 2012)** O data mining tem por objetivo a extração de informações úteis para tomadas de decisão com base nos grandes volumes de dados armazenados nas organizações. Os dados para o data mining são originados restritamente dos data warehouses, pois estes são os que aglomeram enorme quantidade de dados não voláteis e organizados por assunto.

77. **(CESPE / TJ-AC – 2012)** O data mining possibilita analisar dados para obtenção de resultados estatísticos que poderão gerar novas oportunidades ao negócio.

78. **(CESPE / SEDUC-AM - 2011)** A mineração de dados (data mining) é um método computacional que permite extrair informações a partir de grande quantidade de dados.

79. **(CESPE / MEC – 2011)** A exploração, no sentido de utilizar as informações contidas em um datawarehouse, é conhecida como data mining.

80. **(CESPE / Correios – 2011)** Um dos métodos de classificação do datamining é o de análise de agrupamento (cluster), por meio do qual são determinadas características sequenciais utilizando-se dados que dependem do tempo, ou seja, extraindo-se e registrando-se desvios e tendências no tempo.

81. **(CESPE / TJ-ES – 2011)** Mineração de dados, em seu conceito pleno, consiste na realização, de forma manual, de sucessivas consultas ao banco de dados com o objetivo de descobrir padrões úteis, mas não necessariamente novos, para auxílio à tomada de decisão.

82. **(CESPE / PREVIC – 2011)** Um banco de dados pode conter objetos de dados que não sigam o padrão dos dados armazenados. Nos métodos de mineração de dados, esses objetos de dados são tratados como exceção, para que não induzirem a erros na mineração.

83. **(CESPE / SERPRO – 2010)** A mineração de dados (datamining) é uma atividade de processamento analítico não trivial, que, por isso, deve ser realizada por especialistas em ferramentas de desenvolvimento de software e em repositórios de dados históricos orientados a assunto (datawarehouse).

84. **(CESPE / TRT-RN – 2010)** O data mining é um processo automático de descoberta de padrões, de conhecimento em bases de dados, que utiliza, entre outros, árvores de decisão e métodos bayesianos como técnicas para classificação de dados.

85. **(CESPE / EMBASA – 2010)** Data mining é o processo de extração de conhecimento de grandes bases de dados, sendo estas convencionais ou não, e que faz uso de técnicas de inteligência artificial.

86. **(CESPE / SECONT/ES – 2009)** A mineração de dados (data mining) é uma das etapas do processo de descoberta de conhecimento em banco de dados. Nessa etapa, pode ser executada a técnica previsão, que consiste em repartir os dados de modo que diferentes classes ou categorias possam ser identificadas com base em combinações de parâmetros.

87. **(CESPE / IPEA – 2008)** O data mining é um processo utilizado para a extração de dados de grandes repositórios para tomada de decisão, mas sua limitação é não conseguir analisar dados de um data warehouse.

88. **(CESPE / SERPRO – 2008)** A data mining apóia a descoberta de regras e padrões em grandes quantidades de dados. Em data mining, um possível foco é a descoberta de regras de associação. Para que uma associação seja de interesse, é necessário avaliar o seu suporte, que se refere à freqüência com a qual a regra ocorre no banco de dados.

89. **(CESPE / SERPRO – 2008)** A etapa de Mineração de Dados (DM – Data Mining) tem como objetivo buscar efetivamente o conhecimento no contexto da aplicação de KDD (Knowledge Discovery in Databases – Descoberta de Conhecimento em Base de Dados). Alguns autores referem-se à Mineração de Dados e à Descoberta de Conhecimento em Base de Dados como sendo sinônimos. Na etapa de Mineração de Dados são definidos os algoritmos e/ou técnicas que serão utilizados para resolver o problema apresentado. Podem ser usados Redes Neurais, Algoritmo Genéticos, Modelos Estatísticos e Probabilísticos, entre outros, sendo que esta escolha irá depender do tipo de tarefa de KDD que será realizado. "Uma dessas tarefas compreende a busca por uma função que mapeie os registros de um banco de dados em um intervalo de valores reais". Trata-se de:

a) Regressão.
b) Sumarização.
c) Agrupamento.
d) Detecção de desvios.

90. **(CESPE / TCU – 2007)** No datamining, o agrupamento e a classificação funcionam de maneira similar: o agrupamento reconhece os padrões que descrevem o grupo ao qual um item pertence, examinando os itens existentes; a classificação é aplicada quando nenhum grupo foi ainda definido.

QUESTÕES COMENTADAS – FCC

91. **(FCC / AL-AP – 2020)** 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.

92. **(FCC / TRF4 – 2019)** Um Tribunal pretende analisar fatos (fatores ambientais e perfis profissionais, entre outros) que esclareçam por que alguns colaboradores se destacam profissionalmente enquanto outros não se desenvolvem e acabam por se desligar do órgão. 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 prazos. 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 descrevam tendências futuras, denominado:

a) snowflake.
b) drill over.
c) star schema.
d) slice accross.
e) data mining.

93. **(FCC / SEFAZ/BA – 2019)** *Além dos indicadores reativos que, uma vez implantados, automaticamente detectam as ocorrências com base nos indicadores mapeados, existem também os controles proativos, que requerem que os gestores os promovam periodicamente. Uma das técnicas que os gestores podem usar requer que sejam selecionadas, exploradas e modeladas grandes quantidades de dados para revelar padrões, tendências e relações que podem ajudar a identificar casos de fraude e corrupção. Relações ocultas entre pessoas, entidades e eventos são identificadas e as relações suspeitas podem ser encaminhadas para apuração específica. As anomalias apontadas por esse tipo de técnica não necessariamente indicam a ocorrência de fraude*

e corrupção, mas eventos singulares que merecem avaliação individualizada para a exclusão da possibilidade de fraude e corrupção e, no caso da não exclusão, uma investigação.

(Adaptado de: TCU - Tribunal de Contas da União)

O texto se refere à técnica de:

a) data mart.
b) data warehousing.
c) big data.
d) OLAP.
e) data mining.

94. **(FCC / SANASA – 2019)** Considere que a SANASA busca realizar a gestão de recursos hídricos subterrâneos com base em parâmetros conhecidos que determinam a 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.

95. **(FCC / SANASA Campinas – 2019)** Considere que a SANASA busca realizar a gestão de recursos hídricos subterrâneos com base em parâmetros conhecidos que determinam a 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.

96. **(FCC / SABESP – 2018)** O conceito de Data Mining descreve:

a) o uso de teorias, métodos, processos e tecnologias para organizar uma grande quantidade de dados brutos para identificar padrões de comportamentos em determinados públicos.

b) o conjunto de métodos, tecnologias e estratégias para atração voluntária de visitantes, buscando a conversão consistente de leads em clientes (realização de compra).

c) as atividades coordenadas de modo sistemático por uma determinada organização para relacionamento com os seus distintos públicos, bem como com outras organizações, sejam públicas, privadas ou não governamentais.

d) o conjunto de tarefas e processos, organizados e sistematizados, normalmente como uso de uma plataforma tecnológica (hardware e software, ou até mesmo em cloud computing) para a gestão do relacionamento com clientes.

e) o trabalho de produzir levantamento sobre os hábitos de consumo de mídia de um determinado público, identificando horários, tempo gasto etc., associando ao perfil socioeconômico, potencial de consumo, persuasão etc.

97. **(FCC / SEFAZ-SC – 2018)** Para responder à questão, considere o seguinte caso hipotético:

Um Auditor da Receita Estadual pretende descobrir, após denúncia, elementos que possam caracterizar e fundamentar a possível existência de fraudes, tipificadas como sonegação tributária, que vêm ocorrendo sistematicamente na arrecadação do ICMS.

A denúncia é que, frequentemente, caminhões das empresas Org1, Org2 e Org3 não são adequadamente fiscalizados nos postos de fronteiras. Inobservâncias de procedimentos podem ser avaliadas pelo curto período de permanência dos caminhões dessas empresas na operação de pesagem, em relação ao período médio registrado para demais caminhões.

Para caracterizar e fundamentar a existência de possíveis fraudes, o Auditor deverá coletar os registros diários dos postos por, pelo menos, 1 ano e elaborar demonstrativos para análises mensais, trimestrais e anuais.

A aplicação de técnicas de mineração de dados (data mining) pode ser de grande valia para o Auditor. No caso das pesagens, por exemplo, uma ação típica de mining, que é passível de ser tomada com o auxílio de instrumentos preditivos, é:

a) quantificar as ocorrências de possíveis pesagens fraudulentas ocorridas durante todo o trimestre que antecede a data da análise, em alguns postos selecionados, mediante parâmetros comparativos preestabelecidos.

b) analisar o percentual de ocorrências das menores permanências de caminhões nos postos, no último ano, em relação ao movimento total.

c) relacionar os postos onde ocorreram, nos últimos seis meses, as menores permanências das empresas suspeitas e informar o escalão superior para a tomada de decisão.

d) realizar uma abordagem surpresa em determinado posto, com probabilidade significativa de constatar ocorrência fraudulenta.

e) reportar ao escalão superior as características gerais das pesagens e permanências de todos os caminhões, nos cinco maiores postos do Estado, no mês que antecede a data de análise.

98. **(FCC / DPE-RS - 2017)** Uma das técnicas bastante utilizadas em sistemas de apoio à decisão é o Data Mining, que se constitui em uma técnica:

a) para a exploração e análise de dados, visando descobrir padrões e regras, a princípio ocultos, importantes à aplicação.

b) para se realizar a criptografia inteligente de dados, objetivando a proteção da informação.

c) que visa sua distribuição e replicação em um cluster de servidores, visando aprimorar a disponibilidade de dados.

d) de compactação de dados, normalmente bastante eficiente, permitindo grande desempenho no armazenamento de dados.

e) de transmissão e recepção de dados que permite a comunicação entre servidores, em tempo real.

99. **(FCC / AL-MS – 2016)** Um famoso site de vendas sempre envia ao cliente que acabou de comprar um item X, ou o está analisando, a seguinte frase: Pessoas que compraram o item X também compraram o Y. Para isso, o site deve estar aplicando a técnica de Data Mining denominada:

a) profiling.
b) coocorrência.
c) regressão múltipla.
d) regressão logística.
e) classificação.

100. **(FCC / CNMP – 2015)** Em relação às ferramentas de Data Discovery e os fundamentos de Data Mining, é correto afirmar:

a) As ferramentas de Data Mining permitem ao usuário avaliar tendências e padrões não conhecidos entre os dados. Esses tipos de ferramentas podem utilizar técnicas avançadas de computação como redes neurais, algoritmos genéticos e lógica nebulosa, dentre outras.

b) Data Mining é o processo de descobrir conhecimento em banco de dados, que envolve várias etapas. O KDD – Knowledge Discovery in Database é uma destas etapas, portanto, a mineração de dados é um conceito que abrange o KDD.

c) A etapa de KDD do Data Mining consiste em aplicar técnicas que auxiliem na busca de relações entre os dados. De forma geral, existem três tipos de técnicas: Estatísticas, Exploratórias e Intuitivas. Todas são devidamente experimentadas e validadas para o processo de mineração.

d) Os dados podem ser não estruturados (bancos de dados, CRM, ERP), estruturados (texto, documentos, arquivos, mídias sociais, cloud) ou uma mistura de ambos (emails, SOA/web services, RSS). As ferramentas de Data Discovery mais completas possuem conectividade para todas essas origens de dados de forma segura e controlada.

e) Estima-se que, atualmente, em média, 80% de todos os dados disponíveis são do tipo estruturado. Existem diversas ferramentas open source e comerciais de Data Discovery. Dentre as open source está a InfoSphere Data Explorer e entre as comerciais está a Vivisimo da IBM.

101. (FCC / TRF-3R – 2014) Mineração de dados é a investigação de relações e padrões globais que existem em grandes bancos de dados, mas que estão ocultos no grande volume de dados. Com base nas funções que executam, há diferentes técnicas para a mineração de dados, dentre as quais estão:

I. identificar afinidades existentes entre um conjunto de itens em um dado grupo de registros. Por exemplo: 75% dos envolvidos em processos judiciais ligados a ataques maliciosos a servidores de dados também estão envolvidos em processos ligados a roubo de dados sigilosos.

II. identificar sequências que ocorrem em determinados registros. Por exemplo: 32% de pessoas do sexo feminino após ajuizarem uma causa contra o INSS solicitando nova perícia médica ajuízam uma causa contra o INSS solicitando ressarcimento monetário.

III. as categorias são definidas antes da análise dos dados. Pode ser utilizada para identificar os atributos de um determinado grupo que fazem a discriminação entre 3 tipos diferentes, por exemplo, os tipos de processos judiciais podem ser categorizados como infrequentes, ocasionais e frequentes.

Os tipos de técnicas referenciados em I, II e III, respectivamente, são:

a) I - Padrões sequenciais
II - Redes Neurais
III - Árvore de decisão

b) I - Redes Neurais
II - Árvore de decisão
III - Padrões sequenciais

c) I - Associação
II - Padrões sequenciais
III - Classificação

d) I - Classificação
II - Associação
III - Previsão

e) I - Árvore de decisão
II - Classificação
III - Associação

102. (FCC / TCE-RS – 2014) A revista da CGU – Controladoria Geral da União, em sua 8ª edição, publicou um artigo que relata que foram aplicadas técnicas de exploração de dados, visando a descoberta de conhecimento útil para auditoria, em uma base de licitações extraída do sistema ComprasNet, em que são realizados os pregões eletrônicos do Governo Federal. Dentre as técnicas preditivas e descritivas utilizadas, estão a classificação, clusterização e regras de associação. Como resultado, grupos de empresas foram detectados em que a média de participações juntas e as vitórias em licitações levavam a indícios de conluio. As técnicas aplicadas referem-se a:

a) On-Line Analytical Processing.
b) Data Mining.
c) Business Process Management.
d) Extraction, Transformation and Load.
e) Customer Churn Trend Analysis.

103. (FCC / BANESE – 2012) 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.

104. (FCC / TRT/14ª Região – 2011) No contexto de DW, é uma categoria de ferramentas de análise denominada open-end e que permite ao usuário avaliar tendências e padrões não conhecidos entre os dados. Trata-se de:

a) slice.
b) star schema.
c) ODS.
d) ETL.
e) data mining.

105. (FCC / INFRAERO – 2011) No âmbito da descoberta do conhecimento (KDD), a visão geral das etapas que constituem o processo KDD (Fayyad) e que são executadas de forma interativa e iterativa apresenta a seguinte sequência de etapas:

a) seleção, pré-processamento, transformação, data mining e interpretação/avaliação.
b) seleção, transformação, pré-processamento, interpretação/avaliação e data mining.
c) data warehousing, star modeling, ETL, OLAP e data mining.
d) ETL, data warehousing, pré-processamento, transformação e star modeling.
e) OLAP, ETL, star modeling, data mining e interpretação/avaliação.

106. (FCC / TRT/4ª Região – 2010) Sobre data mining, é correto afirmar:

a) É o processo de descoberta de novas correlações, padrões e tendências entre as informações de uma empresa, por meio da análise de grandes quantidades de dados armazenados em bancos de dados usando técnicas de reconhecimento de padrões, estatísticas e matemáticas.

b) Não requer interação com analistas humanos, pois os algoritmos utilizados conseguem determinar de forma completa e eficiente o valor dos padrões encontrados.

c) Na mineração de dados, encontrar padrões requer que os dados brutos sejam sistematicamente "simplificados", de forma a desconsiderar aquilo que é genérico e privilegiar aquilo que é específico.

d) É um grande banco de dados voltado para dar suporte necessário nas decisões de usuários finais, geralmente gerentes e analistas de negócios.

e) O processo de descobrimento realizado pelo data mining só pode ser utilizado a partir de um data warehouse, onde os dados já estão sem erros, sem duplicidade, são consistentes e habilitam descobertas abrangentes e precisas.

107. (FCC / TCE-SP – 2010) NÃO é um objetivo da mineração de dados (mining), na visão dos diversos autores,

a) garantir a não redundância nos bancos transacionais.
b) conhecer o comportamento de certos atributos no futuro.
c) possibilitar a análise de determinados padrões de eventos.
d) categorizar perfis individuais ou coletivos de interesse comercial.

e) apoiar a otimização do uso de recursos limitados e/ou maximizar variáveis de resultado para a empresa.

108. (FCC / TCE-SP – 2010) Considere uma dada população de eventos ou novos itens que podem ser particionados (segmentados) em conjuntos de elementos similares, tal como, por exemplo, uma população de dados sobre uma doença que pode ser dividida em grupos baseados na similaridade dos efeitos colaterias produzidos. Como um dos modos de descrever o conhecimento descoberto durante a data mining este é chamado de:

a) associação.
b) otimização.
c) classificação.
d) clustering.
e) temporização.

109. (FCC / TCM-PA – 2010) Especificamente, um data mining onde as tendências são modeladas conforme o tempo, usando dados conhecidos, e as tendências futuras são obtidas com base no modelo possui a forma de mining:

a) textual.
b) flocos de neve.
c) espacial.
d) estrela.
e) preditivo.

QUESTÕES COMENTADAS – FGV

110. (FGV / Câmara dos Deputados – 2023) CRISP-DM (Cross Industry Standard Process for Data Mining) é uma metodologia utilizada em projetos de Ciência dos Dados. De acordo com esta metodologia, a definição do problema que será investigado por meio de técnicas de mineração de dados ocorre na etapa:

a) *modeling*.
b) *evaluation*.
c) *data preparation*.
d) *data understanding*.
e) *business understanding.*

111. (FGV / Câmara dos Deputados – 2023) O Coeficiente Silhouette é utilizado na análise de agrupamentos, principalmente para examinar:

a) a separação e a coesão dos agrupamentos.
b) a preservação de pequenos agrupamentos.
c) a completude e a interseção dos agrupamentos.
d) a heterogeneidade dos agrupamentos.
e) a forma convexa dos agrupamentos.

112. (FGV / Câmara dos Deputados – 2023) Uma escola está planejando um sistema de acompanhamento temporal de seus alunos, de modo a classificá-los em relação ao desempenho em português e em matemática ao longo de cada ano.

Na escola há uma base de dados históricos que anualmente armazena, para cada aluno, em cada série, a nota final de cada uma dessas duas disciplinas. Essa nota é um valor decimal, entre 0 e 10. Note-se que essa escola, como em outras, há professores que aplicam diferentes graus de exigência nas suas avaliações, uns sendo mais "generosos" e outros, mais "rigorosos".

Três estratégias de transformação de dados foram discutidas, à luz das ideias da Ciência de Dados, como descritas a seguir.

I. Agrupar os alunos a partir de intervalos de notas finais, do tipo "0 até 2,0", "2,1 até 4,0", ..., "8,1 até 10".

II. Rotular grupos de desempenho, "Aprovado" e "Reprovado" e agrupar os alunos de acordo com os critérios de aprovação vigentes em cada situação.

III. Rotular grupos de desempenho, do tipo "Grupo A", "Grupo B", ..., "Grupo E", e agrupar separadamente os alunos de cada conjunto ano/série/disciplina/professor de acordo com a distribuição relativa das notas em cada conjunto.

À luz da ciência de dados e do exposto acima, assinale a afirmativa correta.

a) A primeira estratégia é a melhor para a escola, pois manipula unicamente números que produzem conclusões irrefutáveis.

b) As estratégias II e III complementam-se, pois uma classifica os alunos a partir de parâmetro importante e, a outra, permite uma análise que tenta isolar o grau de exigência de cada professor, e as nuances didáticas de cada disciplina.

c) A segunda estratégia é a melhor para a escola, pois, no fundo, a nota de aprovação adotada em uma escola é a verdadeira medida que reflete o aproveitamento nas disciplinas referidas, independentemente dos critérios do professor.

d) Embora as notas sejam todas numéricas, não existem algoritmos que criem os agrupamentos da estratégia III que sejam diferentes dos agrupamentos que seriam obtidos na estratégia I.

e) As estratégias I e III lidam diretamente com as notas e é impossível gerar novos conhecimentos que alterem a interpretação preconizada pelas notas.

113. **(FGV / SMF-RJ – 2023)** O fiscal de rendas Renan está explorando a base de dados sobre a situação fiscal de empresas que atuam no Rio de Janeiro, e encontrou os seguintes padrões:

• TIPO_EMPRESA = "MEI", RENDA _ANO = "NIVEL A", -> QUANTIDADE_SOCIOS = 1, SITUACAO_FISCAL = "INADIMPLENTE" (suporte = 50%, confiança = 70%)
• TIPO_EMPRESA = "Simples", RENDA_ANO = "NIVEL B" -> QUANTIDADE_SOCIOS = 2, SITUACAO_FISCAL = "REGULAR" (suporte 30%, confiança = 80%)

A técnica de Mineração de dados que Renan aplicou para descobrir elementos que ocorrem em comum dentro de um determinado conjunto de dados foi:

a) análise de cluster;
b) modelos preditivos;
c) árvores de decisão;
d) regras de associação;
e) técnicas de amostragem.

114. **(FGV / SMF-RJ – 2023)** Observe a seguinte estrutura do conjunto de dados PESSOA que contém dados sobre pessoas e a sua renda anual.

Coluna	Tipo	Descrição
Idade	Contínua	Idade em anos
Ganho_capital	Contínua	Ganho de capital
Anos_estudo	Contínua	Anos de estudo
Horas_trabalhadas	Contínua	Horas trabalhadas
Sexo	Categórica	Sexo
Raça / Etnia	Categórica	Raça / Etnia
Educação	Categórica	Educação
Ocupação	Categórica	Ocupação
Classe_trabalho	Categórica	Classe de trabalho
Classe	Categórica	Renda (> 50 mil, <= 50 mil)

O conjunto de dados PESSOA será usado para a tarefa de aprendizagem supervisionada de classificação com a finalidade de prever se a renda (Classe) de uma pessoa excede 50 mil por ano. Para isso, a operação de pré-processamento de dados que deve ser executada no conjunto de dados PESSOA é:

a) exclusão da coluna do tipo categórica "Classe" que possui outlier;

b) discretização das colunas do tipo categórica "Sexo, Raça / Etnia e Educação";

c) normalização por padronização das colunas do tipo categórica "Ocupação e Classe_trabalho";

d) normalização das colunas do tipo continua "Idade, Ganho_capital, Anos_estudo e Horas_trabalhadas";

e) imputação de valores com base na média dos valores existentes na coluna do tipo categórica "Sexo" que possui valores faltantes.

115. **(FGV / EPPGG - 2023)** A mineração de dados ou data mining é uma disciplina interdisciplinar e multidisciplinar que envolve diversas áreas de conhecimento. Assinale a alternativa que enumera corretamente dois tipos de modelagem para análise de dados:

a) Preditivas e Descritivas.
b) Baseadas em Dados e Baseadas em Informação.
c) Matemáticas e Não-Numéricas.
d) Estatísticas e Visualização.
e) Extração e Processamento.

116. **(FGV / Receita Federal - 2023)** A Análise de Componentes Principais (PCA) é uma técnica de transformação de dados que tem como objetivo encontrar as direções de maior variação nos dados, geralmente representadas pelos chamados componentes principais, e gerar novas representações dos dados.

Assinale o objetivo principal dessa técnica.

a) Discretização dos dados.
b) Redução da dimensionalidade dos dados.
c) Normalização dos dados.
d) Padronização dos dados.
e) Cálculo de distâncias entre os dados.

117. (FGV / TCE-TO – 2022) Ao analisar um grande volume de dados, João encontrou algumas anomalias, por exemplo: pessoas com mais de 200 anos de idade e salário de engenheiro menor que salário de pedreiro.

A operação de limpeza da fase de preparação de dados para tratar os pontos extremos existentes em uma série temporal a ser executada por João é:

a) Normalização;
b) Discretização;
c) Classificação;
d) Tratamento de outlier;
e) Redução de dimensionalidade.

118. (FGV / TJDFT – 2022) Maria está explorando a seguinte tabela da base de dados de vendas do mercado HortVega:

IDvenda	*ItensComprados*
1	Cacau, castanha, cogumelo, chia
2	Cacau, chia
3	Cacau, aveia
4	Castanha, cogumelo, tâmara

Utilizando técnicas de Mineração de Dados, Maria encontrou a seguinte informação:

Se um cliente compra Cacau, a probabilidade de ele comprar chia é de 50%. Cacau => Chia, suporte = 50% e confiança = 66,7%.

Para explorar a base de dados do HortVega, Maria utilizou a técnica de Mineração de Dados:

a) normalização;
b) classificação;
c) regra de associação;
d) clusterização;
e) redução de dimensionalidade.

119. (FGV / SEFAZ-AM – 2022) Leia o fragmento a seguir. "CRISP-DM é um modelo de referência não proprietário, neutro, documentado e disponível na Internet, sendo amplamente utilizado para descrever o ciclo de vida de projetos de Ciência de Dados. O modelo é composto por seis fases:

1. entendimento do negócio;
2. _____;
3. _____;
4. Modelagem;
5. _____ ; e
6. implantação".

Assinale a opção cujos itens completam corretamente as lacunas do fragmento acima, na ordem apresentada.

a) modelagem do negócio – limpeza de dados – testagem.
b) modelagem de requisitos – raspagem de dados – execução.
c) modelagem do negócio – mineração de dados – reexecução.
d) compreensão dos dados – preparação dos dados – avaliação.
e) mapeamento de metadados – mineração de dados – testagem.

120. (FGV / SEFAZ-AM – 2022) O tipo de aprendizado máquina, que consiste em treinar um sistema a partir de dados que não estão rotulados e/ou classificados e utilizar algoritmos que buscam descobrir padrões ocultos que agrupam as informações de acordo com semelhanças ou diferenças, é denominado:

a) dinâmico.
b) sistêmico.
c) por reforço.
d) supervisionado.
e) não supervisionado.

121. (FGV / SEFAZ-AM – 2022) Leia o fragmento a seguir.

"A tarefa de detecção de anomalias é um caso particular de problema de _____, onde a quantidade de objetos da classe alvo (anomalia) é muito inferior à quantidade de objetos da classe normal e, adicionalmente, o custo da não detecção de uma anomalia (_____) é normalmente muito maior do que identificar um objeto normal como uma anomalia (_____)".

Assinale a opção cujos itens completam corretamente as lacunas do fragmento acima, na ordem apresentada.

a) aumento de dimensionalidade – redundância – conflito.
b) redução de dimensionalidade – ruído – desvio padrão.

c) análise associativa – discretização – inconsistência.
d) classificação binária – falso negativo – falso positivo.
e) análise probabilística – conflito – ruído.

122. (FGV / SEFAZ-ES – 2021) Maria está preparando um relatório sobrea as empresas de serviços de um município, de modo a identificar e estudar o porte dessas empresas com vistas ao estabelecimento de políticas públicas de previsões de arrecadações. Maria pretende criar nove grupos empresas, de acordo com os valores de faturamento, e recorreu às técnicas usualmente empregadas em procedimentos de data mining para estabelecer as faixas de valores de cada grupo. Assinale a opção que apresenta a técnica diretamente aplicável a esse tipo de classificação:

a) Algoritmos de associação.
b) Algoritmos de clusterização.
c) Árvores de decisão.
d) Modelagem de dados.
e) Regressão linear.

123. (FGV / DETRAN-RN – 2010) Sobre Data Mining, pode-se afirmar que:

a) Refere-se à implementação de banco de dados paralelos.
b) Consiste em armazenar o banco de dados em diversos computadores.
d) Relaciona-se à capacidade de processar grande volume de tarefas em um mesmo intervalo de tempo.
e) Permite-se distinguir várias entidades de um conjunto.
e) Refere-se à busca de informações relevantes a partir de um grande volume de dados.

124. (FGV / Senado Federal – 2008 – Letra A) Em Regras de Associação, confiança refere-se a quantas vezes uma regra de associação se verifica no conjunto de dados analisado.

QUESTÕES COMENTADAS – DIVERSAS BANCAS

125. (FEPESE / ISS-Criciúma – 2022) Quais tipos de conhecimento podem ser descobertos empregando técnicas clássicas de mineração de dados?

1. Regras de associação
2. Hierarquias de classificação
3. Padrões sequenciais ou de série temporal
4. Conhecimento implícito, emergente e não estruturado
5. Agrupamentos e segmentações.

Assinale a alternativa que indica todas as afirmativas **corretas.**

a) São corretas apenas as afirmativas 3 e 5.
b) São corretas apenas as afirmativas 1, 2, 3 e 4.
c) São corretas apenas as afirmativas 1, 2, 3 e 5.
d) São corretas apenas as afirmativas 2, 3, 4 e 5.
e) São corretas as afirmativas 1, 2, 3, 4 e 5.

126. (FEPESE / ISS-Criciúma – 2022) São técnicas de Inteligência Artificial de Data Mining:

1. Estatística.
2. Reconhecimento de Padrões.
3. Representação do Conhecimento.
4. Regras de Associação.

Assinale a alternativa que indica todas as afirmativas **corretas.**

a) São corretas apenas as afirmativas 2 e 3.
b) São corretas apenas as afirmativas 1, 2 e 3.
c) São corretas apenas as afirmativas 1, 2 e 4.
d) São corretas apenas as afirmativas 1, 3 e 4.
e) São corretas as afirmativas 1, 2, 3 e 4.

127. (CESGRANRIO / BB – 2021) Um banco decidiu realizar uma ação de marketing de um novo produto. Buscando apoiar essa ação, a área de TI decidiu estabelecer um mecanismo para identificar quais clientes estariam mais inclinados a adquirir esse produto. Esse mecanismo partia de uma base histórica de clientes que receberam a oferta do produto, e tinha várias colunas com dados sobre os clientes e a oferta, além de uma coluna registrando se eles haviam efetuado ou não a compra do tal produto. Para isso, decidiram ser mais adequado usar um processo de mineração de dados baseado na noção de:

a) agrupamento
b) aprendizado não supervisionado
c) classificação
d) regressão linear
e) suavização

128. (AOCP / MJSP – 2020) Dentre os métodos de mineração de dados, existem aqueles que são supervisionados e os não supervisionados. Assinale a alternativa que apresenta corretamente um dos métodos supervisionados mais comuns para a aplicação da mineração de dados que é voltado às tarefas frequentes do dia a dia:

a) Regras de associação.
b) Bubble sort.
c) Clusterização.
d) Classificação.
e) Formulação.

129. (IBADE / Prefeitura de Vila Velha – 2020) O processo de explorar grandes quantidades de dados a procura de padrões consistentes, para detectar relacionamentos sistemáticos entre variáveis, detectando assim novos subconjuntos de dados é chamado de:

a) Data Lake.
b) Big Data.
c) Data Query.
d) Data Warehouse.
e) Data Mining.

130. (NC-UFPR / Itaipu – 2019) Os algoritmos de Mineração de Dados podem ser classificados quanto a seus objetivos, sendo alguns a classificação, o agrupamento e a identificação de regras de associação. A respeito dessas classificações e seus algoritmos, assinale a alternativa correta.

a) Algoritmos de agrupamento podem ser utilizados para classificação não supervisionada.
b) Algoritmos de agrupamento são também chamados de algoritmos supervisionados.
c) Algoritmos de classificação têm como resultado um modelo descritivo dos dados de entrada.
d) Algoritmos de identificação de regras são também conhecidos como algoritmos preditivos.
e) Algoritmos de agrupamento são equivalentes a algoritmos de identificação de anomalias.

131. (CESGRANRIO / BANCO DA AMAZÔNIA – 2018) As ferramentas e técnicas de mineração de dados (data mining) têm por objetivo:

a) preparar dados para serem utilizados em um "data warehouse" (DW).
b) permitir a navegação multidimensional em um DW.
c) projetar, de forma eficiente, o registro de dados transacionais.

d) buscar a classificação e o agrupamento (clusterização) de dados, bem como identificar padrões.
e) otimizar o desempenho de um gerenciador de banco de dados.

132. (COPESE / UFT – 2018 – Item III) Diversos modelos de Redes Neurais Artificiais podem ser utilizados na implementação de métodos de Mineração de Dados.

133. (FAURGS / FAURGS – 2018) Uma nuvem de palavras é um recurso gráfico (usado principalmente na internet) para descrever os termos mais frequentes de um determinado texto. O tamanho da fonte em que a palavra é apresentada é uma função da frequência da palavra no texto: palavras mais frequentes são desenhadas em fontes de tamanho maior, palavras menos frequentes são desenhadas em fontes de tamanho menor. Qual é a técnica de análise de dados descrita pelo texto acima?

a) Processamento de Linguagem Natural.
b) Agrupamento.
c) Classificação.
d) Redes Neurais.
e) Regressão Linear.

134. (CESGRANRIO / Petrobrás – 2018) Dois funcionários de uma empresa de crédito discutiam sobre quais algoritmos deveriam usar para ajudar a classificar seus clientes como bons ou maus pagadores. A empresa possui, para todos os empréstimos feitos no passado, um registro formado pelo conjunto de informações pessoais sobre o cliente e de como era composta a dívida inicial. Todos esses registros tinham classificações de bons ou maus pagadores, de acordo com o perfil de pagamento dos clientes. A partir desses dados, os funcionários querem construir um modelo, por meio de aprendizado de máquina, que classifique os novos clientes, que serão descritos por registros com o mesmo formato. A melhor opção, nesse caso, é usar um algoritmo:

a) supervisionado, como SVM.
b) supervisionado, como K-means.
c) não supervisionado, como regressão linear.
d) não supervisionado, como árvores de decisão.
e) semi-supervisionado, como redes bayesianas.

135. (ESAF / STN – 2018) Uma técnica de classificação em Mineração de Dados é uma abordagem sistemática para:

a) construção de controles de ordenação a partir de um conjunto de acessos.
b) construção de modelos de classificação a partir de um conjunto de dados de entrada.
c) construção de modelos de dados a partir de um conjunto de algoritmos.
d) construção de controles de ordenação independentes dos dados de entrada.
e) construção de modelos de sistemas de acesso a partir de um conjunto de algoritmos.

136. **(CESGRANRIO / TRANSPETRO – 2018)** Um desenvolvedor recebeu um conjunto de dados representando o perfil de um grupo de clientes, sem nenhuma informação do tipo de cada cliente, onde cada um era representado por um conjunto fixo de atributos, alguns contínuos, outros discretos. Exemplos desses atributos são: idade, salário e estado civil. Foi pedido a esse desenvolvedor que, segundo a similaridade entre os clientes, dividisse os clientes em grupos, sendo que clientes parecidos deviam ficar no mesmo grupo. Não havia nenhuma informação que pudesse ajudar a verificar se esses grupos estariam corretos ou não nos dados disponíveis para o desenvolvedor. Esse é um problema de data mining conhecido, cuja solução mais adequada é um algoritmo:

a) de regressão
b) não supervisionado
c) por reforço
d) semissupervisionado
e) supervisionado

137. **(FEPESE / CIASC – 2017)** Assinale a alternativa que contém as principais fases do processo de Data Minning CRISP-DM.

a) Amostragem; Exploração; Modificação; Modelagem; Execução; Avaliação.
b) Amostragem; Exploração; Modelagem; Modificação; Avaliação; Implementação.
c) Compreensão do negócio; Compreensão dos dados; Preparação dos dados; Modelagem; Avaliação; implementação.
d) Compreensão dos dados; Amostragem; Preparação dos dados; Implementação; Avaliação.
e) Compreensão do negócio; Exploração dos dados; Modificação dos dados; Implementação; Avaliação.

138. **(NC-UFPR / Itaipu Binacional – 2015)** Qual é a funcionalidade do Oracle Data Mining que encontra aglomerados de objetos de dados semelhantes em algum sentido entre si?

a) Aprior
b) Associação
c) Classificação
d) Clustering
e) Regressão

139. **(AOCP / TCE-PA – 2014)** O processo de explorar grandes quantidades de dados à procura de padrões consistentes com o intuito de detectar relacionamentos sistemáticos entre variáveis e novos subconjuntos de dados, é conhecido como:

a) Data Mart.
b) Data Exploring.
c) Objeto Relacional.

d) Relacionamento.
e) Data Mining.

140. (VUNESP / TJ-PA – 2014) Uma das tarefas implementadas por uma ferramenta de Data Mining consiste em realizar a determinação de um valor futuro de determinada característica ou atributo de um registro ou conjunto de registros. Tal tarefa corresponde à:

a) normalização.
b) indexação.
c) análise de afinidade.
d) predição.
e) análise de equivalência

141. (FUNDEP / IFN/MG – 2014) Ao se utilizar a técnica de data mining (mineração de dados), como é conhecido o resultado dessa mineração, em que, por exemplo, se um cliente compra equipamento de vídeo, ele pode também comprar outros equipamentos eletrônicos?

a) Regras de associação
b) Padrões sequenciais
c) Árvores de classificação
d) Padrões de aquisição

142. (FAURGS / TJ-RS – 2014) O resultado da mineração de dados pode ser a descoberta de tipos de informação "nova". Supondo-se que um cliente compre uma máquina fotográfica e que, dentro de três meses, compre materiais fotográficos, há probabilidade de que, dentro dos próximos seis meses, ele comprará um acessório. Um cliente que compra mais que duas vezes, em um período de baixa, deverá estar propenso a comprar, pelo menos uma vez, no período do Natal. Esse tipo de informação pode ser verificado através de:

a) predição de links.
b) regras de associação.
c) árvores de classificação.
d) árvores de decisão.
e) padrões sequenciais.

143. (CESGRANRIO / LIQUIGÁS – 2014) As empresas possuem grandes quantidades de dados. Em geral, a maioria delas é incapaz de aproveitar plenamente o valor que eles têm. Com o intuito de melhorar essa situação, surgiu o data mining, que se caracteriza por:

a) desenhar padrões já conhecidos
b) extrair padrões ocultos nos dados.
c) tomar decisões para os gestores.
d) não trabalhar com tendências.
e) não trabalhar com associações.

144. CCV-UFC / UFC / 2013) Sobre Mineração de Dados, assinale a alternativa correta.

a) É uma técnica de organização de grandes volumes de dados.

b) É um conjunto de técnicas avançadas para busca de dados complexos.

c) É o processo de explorar grande quantidade de dados para extração não-trivial de informação implícita desconhecida.

d) É um processo automatizado para a recuperação de informações caracterizadas por registros com grande quantidade de atributos.

e) É um processo de geração de conhecimento que acontece durante o projeto de banco de dados. Os requisitos dos usuários são analisados e minerados para gerar as abstrações que finalmente são representadas em um modelo de dados.

145. (FMP CONCURSOS / MPE-AC – 2013) Processo de explorar grandes quantidades de dados à procura de padrões consistentes, como regras de associação ou sequências temporais, para detectar relacionamentos sistemáticos entre variáveis, detectando assim novos subconjuntos de dados é conhecido como:

a) datawarehouse.
b) SGBD.
c) mineração de dados (data mining).
d) modelagem relacional de dados.
e) mineração de textos (text mining).

146. (IBFC / EBSERH – 2013) Processo de explorar grandes quantidades de dados à procura de padrões consistentes, como regras de associação ou sequências temporais:

a) Data Warehouse
b) Data Mining
c) Tunning
d) APS (Application Platform Suite)

147. (FUNRIO / MPOG – 2013) Qual o tipo de descoberta de conhecimento através de mineração de dados (do inglês "data mining"), em que se relaciona a presença de conjuntos de itens diversos, como por exemplo: "Quando uma mulher compra uma bolsa em uma loja, ela está propensa a comprar sapatos"?

a) Hierarquias de classificação.
b) Padrões sequenciais.

c) Regras de associação.
d) Séries temporais.
e) Agrupamentos por similaridade.

148. **(ESPP / MPE-PR – 2013)** Data Mining refere-se à busca de informações relevantes, ou "à descoberta de conhecimento", a partir de um grande volume de dados. Assim como a descoberta de conhecimento no ramo da inteligência artificial, a extração de dados tenta descobrir automaticamente modelos estatísticos a partir dos dados. O conhecimento obtido a partir de um banco de dados pode ser representado em regras. Duas importantes classes de problemas de extração de dados são as:

a) regras de indexação e regras de população.
b) regras de validação e regras de otimização.
c) regras de interpolação e regras de valoração.
d) regras de maximização e regras de generalização.
e) regras de classificação e regras de associação.

149. **(ESAF / MF – 2013)** A Mineração de Dados requer uma adequação prévia dos dados através de técnicas de pré-processamento. Entre elas estão as seguintes técnicas:

a) Agrupamento. Amostragem. Redução de dimensionalidade. Seleção de subconjuntos de recursos. Recursos pontuais. Polarização. Redução de variáveis.

b) Agregação. Classificação. Redução de faixas de valores. Seleção de subconjuntos de recursos. Redução de recursos. Terceirização e discretização. Transformação de variáveis.

c) Agrupamento. Classificação. Redução de dimensionalidade. Seleção de subconjuntos de usuários. Criação de recursos. Binarização e discretização. Transformação de conjuntos.

d) Agregação. Amostragem. Redução de dimensionalidade. Seleção de subconjuntos de usuários. Criação de recursos. Polarização. Transformação de conjuntos.

e) Agregação. Amostragem. Redução de dimensionalidade. Seleção de subconjuntos de recursos. Criação de recursos. Binarização e discretização. Transformação de variáveis.

150. **(IADES / EBSERH – 2012)** Existem algumas técnicas utilizadas em Data mining, para fins de estatísticas. A técnica que permite lidar com a previsão de um valor, em vez de uma classe, é denominada:

a) associação.
b) exploração.
c) classificação.
d) regressão.
e) árvore de decisão.

151. **(CESGRANRIO / EPE – 2012)** As técnicas de mineração de dados podem ser categorizadas em supervisionadas e não supervisionadas. As técnicas de árvores de decisão, agrupamento e regras de associação são categorizadas, respectivamente, como:

a) não supervisionada, não supervisionada, não supervisionada
b) não supervisionada, supervisionada e não supervisionada
c) supervisionada, não supervisionada e não supervisionada
d) supervisionada, não supervisionada e supervisionada
e) supervisionada, supervisionada e supervisionada

152. **(FMP CONCURSOS / TCE-RS – 2011)** Mineração de dados consiste em:

a) explorar um conjunto de dados visando a extrair ou a ajudar a evidenciar padrões, como regras de associação ou sequências temporais, para detectar relacionamentos entre estes.

b) acessar um banco de dados para realizar consultas de forma genérica, buscando recuperar informações (registros) que atendam um mesmo critério de pesquisa.

c) recuperar informações de um banco de dados específico, voltado a representar e armazenar dados relacionados com companhias de exploração petrolífera e de recursos mineralógicos.

d) um banco de dados específico voltado à gestão de negócios usando tecnologia de informação (TI) como, por exemplo, a área de BI (Business Inteligence).

e) representar informações de um banco de dados mediante vários modelos hierárquicos como, por exemplo, o de entidade-relacionamento (ER).

153. **(FUMARC / PRODEMGE – 2011)** Analise as afirmativas abaixo em relação às técnicas de mineração de dados.

I. Regras de associação podem ser usadas, por exemplo, para determinar, quando um cliente compra um produto X, ele provavelmente também irá comprar um produto Y.

II. Classificação é uma técnica de aprendizado supervisionado, no qual se usa um conjunto de dados de treinamento para aprender um modelo e classificar novos dados.

III. Agrupamento é uma técnica de aprendizado supervisionado que particiona um conjunto de dados em grupos.

Assinale a alternativa VERDADEIRA:

a) Apenas as afirmativas I e II estão corretas.
b) Apenas as afirmativas I e III estão corretas.

c) Apenas as afirmativas II e III estão corretas.
d) Todas as afirmativas estão corretas.

154. (FMP CONCURSOS / TCE/RS – 2011 – Letra B) Mineração de Dados é parte de um processo maior de pesquisa chamado de Busca de Conhecimento em Banco de Dados (KDD).

155. (ESAF / CVM – 2010) Mineração de Dados é:

a) o processo de atualizar de maneira semi-automática grandes bancos de dados para encontrar versões úteis.

b) o processo de analisar de maneira semi-automática grandes bancos de dados para encontrar padrões úteis.

c) o processo de segmentar de maneira semi-automática bancos de dados qualitativos e corrigir padrões de especificação.

d) o programa que depura de maneira automática bancos de dados corporativos para mostrar padrões de análise.

e) o processo de automatizar a definição de bancos de dados de médio porte de maior utilidade para os usuários externos de rotinas de mineração.

156. (ESAF / MPOG – 2010) Mineração de Dados:

a) é uma forma de busca sequencial de dados em arquivos.
b) é o processo de programação de todos os relacionamentos e algoritmos existentes nas bases de dados.
c) por ser feita com métodos compiladores, método das redes neurais e método dos algoritmos gerativos.
d) engloba as tarefas de mapeamento, inicialização e clusterização.
e) engloba as tarefas de classificação, regressão e clusterização.

157. (CESGRANRIO / ELETROBRÁS – 2010) Em uma reunião sobre prospecção de novos pontos de venda, um analista de TI afirmou que técnicas OLAP de análise de dados são orientadas a oferecer informações, assinalando detalhes intrínsecos e facilitando a agregação de valores, ao passo que técnicas de data mining tem como objetivo:

a) captar, organizar e armazenar dados colecionados a partir de bases transacionais, mantidas por sistemas OLTP.

b) facilitar a construção de ambientes de dados multidimensionais, através de tabelas fato e dimensionais.

c) melhorar a recuperação de dados organizados de forma não normalizada em uma base relacional conhecida como data warehouse.

d) extrair do data warehouse indicadores de controle (BSC) para apoio à tomada de decisão por parte da diretoria da empresa.

e) identificar padrões e recorrência de dados, oferecendo conhecimento sobre o comportamento dos dados analisados.

158. **(UFF / UFF – 2009)** O conjunto de técnicas que, envolvendo métodos matemáticos e estatísticos, algoritmos e princípios de inteligência artificial, tem o objetivo de descobrir relacionamentos significativos entre dados armazenados em repositórios de grandes volumes e concluir sobre padrões de comportamento de clientes de uma organização é conhecido como:

a) Datawarehouse;
b) Metadados;
c) Data Mart;
d) Data Mining;
e) Sistemas Transacionais.

159. **(COSEAC / DATAPREV – 2009)** "*Mining é parte de um processo maior de conhecimento, que o processo consiste, fundamentalmente, na estruturação do banco de dados; na seleção, preparação e pré-processamento dos dados; na transformação, adequação e redução da dimensionalidade dos dados; e nas análises, assimilações, interpretações e uso do conhecimento extraído do banco de dados*". O processo maior citado no início do texto é denominado:

a) Data mining
b) Data mart
c) Data warehouse
d) KDD
e) Segmentação de dados

GABARITO

1. CORRETO
2. ERRADO
3. CORRETO
4. ERRADO
5. CORRETO
6. LETRA A
7. CORRETO
8. CORRETO
9. ERRADO
10. CORRETO
11. LETRA D
12. LETRA B
13. LETRA A
14. LETRA E
15. CORRETO
16. ERRADO
17. LETRA B
18. CORRETO
19. ERRADO
20. CORRETO
21. CORRETO
22. ERRADO
23. ERRADO
24. CORRETO
25. CORRETO
26. CORRETO
27. CORRETO
28. CORRETO
29. CORRETO
30. ERRADO
31. ERRADO
32. ERRADO
33. CORRETO
34. CORRETO
35. CORRETO
36. CORRETO
37. CORRETO
38. CORRETO
39. CORRETO
40. ERRADO
41. CORRETO
42. CORRETO
43. CORRETO
44. CORRETO
45. ERRADO
46. CORRETO
47. ERRADO
48. LETRA E
49. LETRA B
50. CORRETO
51. CORRETO
52. CORRETO
53. CORRETO
54. LETRA C
55. CORRETO
56. ERRADO
57. CORRETO
58. CORRETO
59. CORRETO
60. CORRETO
61. ERRADO
62. CORRETO
63. ERRADO
64. CORRETO
65. CORRETO
66. CORRETO
67. CORRETO
68. ERRADO
69. ERRADO
70. ERRADO
71. ERRADO
72. LETRA C
73. ERRADO
74. ERRADO
75. CORRETO
76. ERRADO
77. CORRETO
78. CORRETO
79. CORRETO
80. ERRADO
81. ERRADO
82. ERRADO
83. ERRADO
84. CORRETO
85. ERRADO
86. ERRADO
87. ERRADO
88. CORRETO
89. LETRA A
90. LETRA E
91. LETRA C
92. LETRA E
93. LETRA E
94. LETRA B
95. LETRA B
96. LETRA A
97. LETRA D
98. LETRA A
99. LETRA B
100. LETRA A
101. LETRA C
102. LETRA B
103. LETRA C
104. LETRA E
105. LETRA A
106. LETRA A
107. LETRA A
108. LETRA D
109. LETRA E
110. LETRA E
111. LETRA A
112. LETRA B
113. LETRA D
114. LETRA D
115. LETRA A
116. LETRA B
117. LETRA D
118. LETRA C
119. LETRA D
120. LETRA E
121. LETRA D
122. LETRA B
123. LETRA E

124. ERRADO
125. LETRA C
126. LETRA A
127. LETRA C
128. LETRA D
129. LETRA E
130. LETRA A
131. LETRA D
132. CORRETO
133. LETRA A
134. LETRA A
135. LETRA B
136. LETRA B
137. LETRA C
138. LETRA D
139. LETRA E
140. LETRA D
141. LETRA A
142. LETRA E
143. LETRA B
144. LETRA C
145. LETRA C
146. LETRA B
147. LETRA C
148. LETRA E
149. LETRA E
150. LETRA D
151. LETRA C
152. LETRA A
153. LETRA A
154. CORRETO
155. LETRA B
156. LETRA E
157. LETRA E
158. LETRA D
159. LETRA D

ESSA LEI TODO MUNDO CONHECE:
PIRATARIA É CRIME.

Mas é sempre bom revisar o porquê e como você pode ser prejudicado com essa prática.

1. 

Professor investe seu tempo para elaborar os cursos e o site os coloca à venda.

2. 

Pirata divulga ilicitamente (grupos de rateio), utilizando-se do anonimato, nomes falsos ou laranjas (geralmente o pirata se anuncia como formador de "grupos solidários" de rateio que não visam lucro).

3. 

Pirata cria alunos fake praticando falsidade ideológica, comprando cursos do site em nome de pessoas aleatórias (usando nome, CPF, endereço e telefone de terceiros sem autorização).

4. 

Pirata compra, muitas vezes, clonando cartões de crédito (por vezes o sistema anti-fraude não consegue identificar o golpe a tempo).

5. 

Pirata fere os Termos de Uso, adultera as aulas e retira a identificação dos arquivos PDF (justamente porque a atividade é ilegal e ele não quer que seus fakes sejam identificados).

6. 

Pirata revende as aulas protegidas por direitos autorais, praticando concorrência desleal e em flagrante desrespeito à Lei de Direitos Autorais (Lei 9.610/98).

7. 

Concurseiro(a) desinformado participa de rateio, achando que nada disso está acontecendo e esperando se tornar servidor público para exigir o cumprimento das leis.

8. 

O professor que elaborou o curso não ganha nada, o site não recebe nada, e a pessoa que praticou todos os ilícitos anteriores (pirata) fica com o lucro.

Deixando de lado esse mar de sujeira, aproveitamos para agradecer a todos que adquirem os cursos honestamente e permitem que o site continue existindo.