PUC-SP

Mineração de Dados

Prof. Dr. Daniel Rodrigues da Silva

PUC-SP

Introdução a Mineração de Dados

Bibliografia básica:

Introdução à mineração de dados : conceitos básicos, algoritmos e aplicações. Leandro Nunes de Castro e Daniel Gomes Ferrari. – São Paulo : Saraiva, 2016.

Inteligência Artificial - Uma Abordagem de Aprendizado de Máquina . André Carlos Ponce de Leon Ferreira et al. 2 Ed. LTC, 2024.

PUC-SP

Ementa

- Conceitos de Mineração de Dados
- Análise exploratória
- Análise preditiva,
- Agrupamentos
- Associações.

Principais Ferramentas Computacionais usadas na disciplina

PUC-SP

pandas

NumPy

seaborn

Critérios de Avaliação

PUC-SP

Pelo menos 75% de presença

Média final maior ou igual a 5,0

$$MF = \frac{N_1 + N_2}{2} \quad ; \quad N_i = \frac{P_i + A_i}{2} \ , \quad i = 1\,,2$$

Onde:

$P_i = nota\ do\ projeto\ do\ bimestre$

$A_i = nota\ da\ atividade/prova\ do\ bimestre$

INTRODUÇÃO

PUC-SP

Estima-se que mais de **400 milhões de terabytes** de dados são criados a cada dia

Estima-se que por volta de **181 zettabytes** de dados serão gerados em 2025

Os vídeos são responsáveis por **mais da metade** do tráfego de dados da Internet

Os EUA têm **mais de 2.700** centros de dados

Informações extraídas em julho/2025 de: explodingtopics.com

Em perspectiva, esta é a quantidade de dados gerados por dia em algumas unidades de medida:

PUC-SP

Unidade de Medida	Dados gerados
Zetabytes	0,4
Exabytes	402,74
Petabytes	402.740
Terabytes	402,74 milhões
Gigabytes	402,74 bilhões
Megabytes	402,74 trilhões
Quilobytes	402,74 quatrilhões
Bytes	402,74 quintilhões

PUC-SP

Projeções para o crescimento da criação de dados

A quantidade de dados gerados por ano cresce ano a ano desde 2010.

Estima-se que 90% dos dados do mundo foram gerados somente nos últimos dois anos.

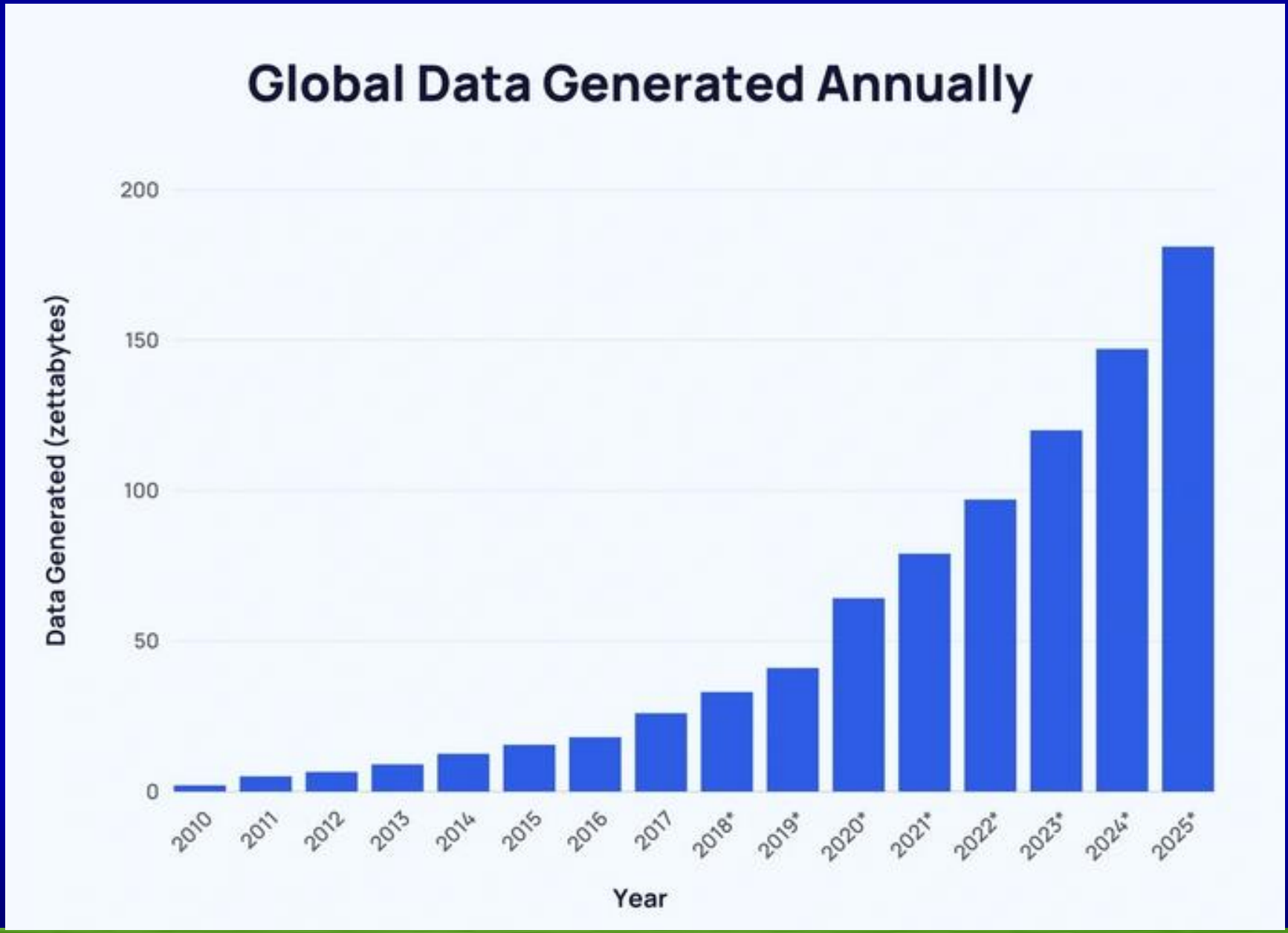
Em apenas 13 anos, esse número aumentou em cerca de **74 vezes** , de apenas **2 zettabytes** em 2010.

Estima-se que os **120 zettabytes** gerados em 2023 aumentem em **mais de 50%** em 2025, atingindo **181 zettabytes** .

Ano	Dados gerados	Mudança em relação ao ano anterior	Mudança em relação ao ano anterior (%)
2010	2 zetabytes	-	-
2011	5 zetabytes	↑ 3 zetabytes	↑ 150%
2012	6,5 zetabytes	↑ 1,5 zetabytes	↑ 30%
2013	9 zetabytes	↑ 2,5 zetabytes	↑ 38,46%
2014	12,5 zetabytes	↑ 3,5 zetabytes	↑ 38,89%
2015	15,5 zetabytes	↑ 3 zetabytes	↑ 24%
2016	18 zetabytes	↑ 2,5 zetabytes	↑ 16,13%
2017	26 zetabytes	↑ 8 zetabytes	↑ 44,44%
2018*	33 zetabytes	↑ 7 zetabytes	↑ 26,92%
2019*	41 zetabytes	↑ 8 zetabytes	↑ 24,24%
2020*	64,2 zetabytes	↑ 23,2 zetabytes	↑ 56,59%
2021*	79 zetabytes	↑ 14,8 zetabytes	↑ 23,05%
2022*	97 zetabytes	↑ 18 zetabytes	↑ 22,78%
2023*	120 zetabytes	↑ 23 zetabytes	↑ 23,71%
2024*	147 zetabytes	↑ 27 zetabytes	↑ 22,5%
2025*	181 zetabytes	↑ 34 zetabytes	↑ 23,13%

PUC-SP

Geração de dados de 2010 a 2025



PUC-SP

Criação de dados por categoria

PUC-SP

Os vídeos são responsáveis por mais de **53%** de todo o tráfego global de dados.

As mídias sociais estão cheia de vídeos: **O TikTok** é totalmente baseado em vídeos e continua a aumentar sua base de usuários ano após ano.

Enquanto o **Facebook** evoluiu a ponto de **51%** do conteúdo compartilhado na plataforma ser baseado em vídeo.

Criação de dados por categoria

PUC-SP

Embora os dados públicos relacionados ao Snapchat sejam limitados, estima-se que cada snap enviado exija **1 MB** , muitos dos quais são vídeos

Juntamente com as redes sociais **(12,69%)** e os jogos **(9,86%)** , essas três categorias representam mais de 3/4 de todo o tráfego de dados da Internet.

Estima-se que **quase 250 milhões de** e-mails são enviados a cada minuto.

Categoria	Proporção do tráfego de dados da Internet
Vídeo	53,72%
Social	12,69%
Jogos	9,86%
Navegação na web	5,67%
Mensagens	5,35%
Mercado	4,54%
Compartilhamento de arquivos	3,74%
Nuvem	2,73%
VPN	1,39%
Áudio	0,31%

PUC-SP

PUC-SP

Tipo de mídia	Quantidade por minuto	Valor por dia
E-mails enviados	231,4 milhões	333,22 bilhões
Criptomoeda comprada	90,2 milhões	129,89 bilhões
Textos enviados	16 milhões	24,04 bilhões
Pesquisas do Google	5,9 milhões	8,5 bilhões
Snaps compartilhados no Snapchat	2,43 milhões	3,5 bilhões
Pedaços de conteúdo compartilhados no Facebook	1,7 milhões	2,45 bilhões
Deslizes no Tinder	1,1 milhão	1,58 bilhão
Horas transmitidas	1 milhão	1,44 bilhão
USD gasto na Amazon	443.000	637,92 milhões
USD enviado no Venmo	437.600	630,14 milhões
Tweets compartilhados no Twitter	347.200	499,97 milhões
Horas gastas em reuniões do Zoom	104.600	150,62 milhões
USD gasto no DoorDash	76.400	110,02 milhões

YouTube

PUC-SP

Assistir a vídeos do YouTube em resolução 480p consome **mais de 500 MB** de dados por hora.

Enquanto vídeos 4k do YouTube usam **cerca de 30 vezes mais** .

Qualidade de vídeo do YouTube	Dados usados por hora	Dados usados por dia
480p	562,5 MB	13,5 GB
720p	1,86 GB	44,64 GB
1080p	3,04 GB	72,96 GB
4k	15,98 GB	383,52 GB

Spotify: As configurações padrão do Spotify usam 2 MB por música de 3 minutos.

Netflix: Cada transmissão de definição padrão da Netflix usa 1 GB de dados por hora (24 GB por dia).

Os streamings de alta definição da Netflix podem usar até 3 GB de dados por hora (72 GB por dia).

Ultra HD usa 7 GB por hora (168 GB por dia).

Fontes:Sandvine,Domo,TechJury, iNews

PUC-SP

Cinco maiores data centers do mundo

Os EUA têm **mais de 10 vezes mais** data centers **(5.426)** do que qualquer outro país.

Alemanha **(529)**

Reino Unido **(523)**

China **(449)**

Canadá**(337)**

A seguir estão as 15 principais nações por data centers no mundo:

PUC-SP

Classificação	País	Região	Número de Data Centers
1	EUA	América do Norte	5.426
2	Alemanha	Europa	529
3	Reino Unido	Europa	523
4	China	Ásia	449
5	Canadá	América do Norte	337
6	França	Europa	322
7	Austrália	Oceânia	314
8	Holanda	Europa	298
9	Rússia	Europa/Ásia	251
10	Japão	Ásia	222
11	Brasil	América do Sul	197
12	México	América do Norte	173
13	Itália	Europa	168
14	Índia	Ásia	153
15	Polônia	Europa	144

PUC-SP

Em 1965, Gordon Moore, um dos fundadores da Intel, publicou um artigo no qual observou que a quantidade de componentes em um circuito integrado (CI) estava dobrando aproximadamente a cada ano desde a sua invenção, em 1958, e essa taxa permaneceria por pelo menos mais dez anos. Em 1975, Moore atualizou sua estimativa para períodos de dois anos, em vez de um ano. Essa elevada taxa de crescimento na quantidade de componentes do CI está diretamente relacionada à velocidade de processamento e capacidade de memória dos computadores e também tem servido de meta para a indústria de hardware computacional

PUC-SP

Paradoxalmente, esses avanços da tecnologia – tanto de hardware quanto de comunicação – têm produzido um problema de superabundância de dados, pois a capacidade de coletar e armazenar dados tem superado a habilidade de analisar e extrair conhecimento destes.

Nesse contexto, é necessária a aplicação de técnicas e ferramentas que transformem, de maneira inteligente e automática, os dados disponíveis em informações úteis, que representem conhecimento para uma tomada de decisão estratégica nos negócios e até no dia a dia de cada um de nós.

Uma das mais emblemáticas representantes dessa superabundância de dados é a *computação em nuvem* (*cloud computing*), que se refere ao fornecimento de recursos computacionais como serviço em vez de produto.

Na nuvem, recursos (hardware e software) são compartilhados por meio de uma rede e os usuários podem acessar as aplicações hospedadas em servidores remotos utilizando browsers, desktops e até aplicativos móveis.

Essa possibilidade tem atraído muitas empresas e usuários comuns, principalmente em virtude dos elevados custos e da complexidade de manutenção de servidores de dados e aplicação.

É exatamente nesse contexto de superabundância de dados que surgiu **a *mineração de dados***, como um processo sistemático, interativo e iterativo, de preparação e extração de conhecimentos a partir de grandes bases de dados.

Neste curso temos como objetivo apresentar, de maneira didática, o processo de mineração de dados.

O QUE É MINERAÇÃO DE DADOS?

PUC-SP

O processo de *mineração* corresponde à extração de *minerais valiosos*, como ouro e pedras preciosas de uma *mina*.

Uma característica importante desses materiais é que, embora não possam ser cultivados ou produzidos artificialmente, existem de maneira oculta e muitas vezes desconhecida em alguma fonte, podendo ser extraídos.

Esse processo requer acesso à mina, o uso de ferramentas adequadas de mineração, a extração dos minérios propriamente dita e o seu posterior preparo para comercialização

PUC-SP

O termo ***Mineração de Dados*** (MD) foi cunhado como alusão ao processo de mineração descrito anteriormente, uma vez que se explora uma *base de dados* (mina) usando *algoritmos* (ferramentas) adequados para obter *conhecimento* (minerais preciosos).

Os *dados* são símbolos ou signos não estruturados, sem significado, como valores em uma tabela, e a *informação* está contida nas descrições, agregando significado e utilidade aos dados.

Por fim, o *conhecimento* é algo que permite uma tomada de decisão para a agregação de valor, então, por exemplo, saber se vai chover no fim de semana pode influenciar sua decisão de viajar.

PUC-SP

Dados

- 1000 milibares
- 5,1 m/s; 95°
- 30 °C
- poucas
- 1000 mts

Informação

- Pressão atmosférica = 1000 milibares
- Velocidade e direção do vento = 5,1 m/s; 95°
- Temperatura do ar = 30 °C
- Nuvens = poucas
- Visibilidade = 1000 mts

Conhecimento

- A probabilidade de chuva é baixa, portanto, posso ir à praia.

Diferença entre:

- Dados
- Informação e
- Conhecimento

O QUE É MINERAÇÃO DE DADOS?

PUC-SP

A mineração de dados é parte integrante de um processo mais amplo, conhecido como ***descoberta de conhecimento em bases de dados*** (*knowledge discovery in databases,* ou ***KDD***).

PUC-SP

Embora muitos usem mineração de dados como sinônimo de KDD, na primeira conferência internacional sobre KDD, realizada na cidade de Montreal, Canadá, em 1995, foi proposto que a terminologia *descoberta de conhecimentos em bases de dados* se referisse a todo o processo de extração de conhecimentos a partir de dados. Foi proposto também que a terminologia **mineração de dados** fosse empregada exclusivamente para a etapa de descoberta do processo de KDD, que inclui:

- **a *seleção e integração das bases de dados***
- **a *limpeza da base***
- **a *seleção e transformação dos dados***
- **a *mineração* e a *avaliação dos dados***.

Sintetizamos o processo de KDD em quatro partes principais:

Base de dados: coleção organizada de dados, ou seja, valores quantitativos ou qualitativos referentes a um conjunto de itens que permite uma recuperação eficiente dos dados. Conceitualmente, os dados podem ser entendidos como o nível mais básico de abstração a partir do qual a informação e, depois, os conhecimentos podem ser extraídos;

Preparação ou Pré-Processamento de Dados: são etapas anteriores à mineração que visam preparar os dados para uma análise eficiente e eficaz. Essa etapa inclui a *limpeza* (remoção de ruídos e dados inconsistentes), a *integração* (combinação de dados obtidos a partir de múltiplas fontes), a *seleção* ou *redução* (escolha dos dados relevantes à análise) e a *transformação* (transformação ou consolidação dos dados em formatos apropriados para a mineração);

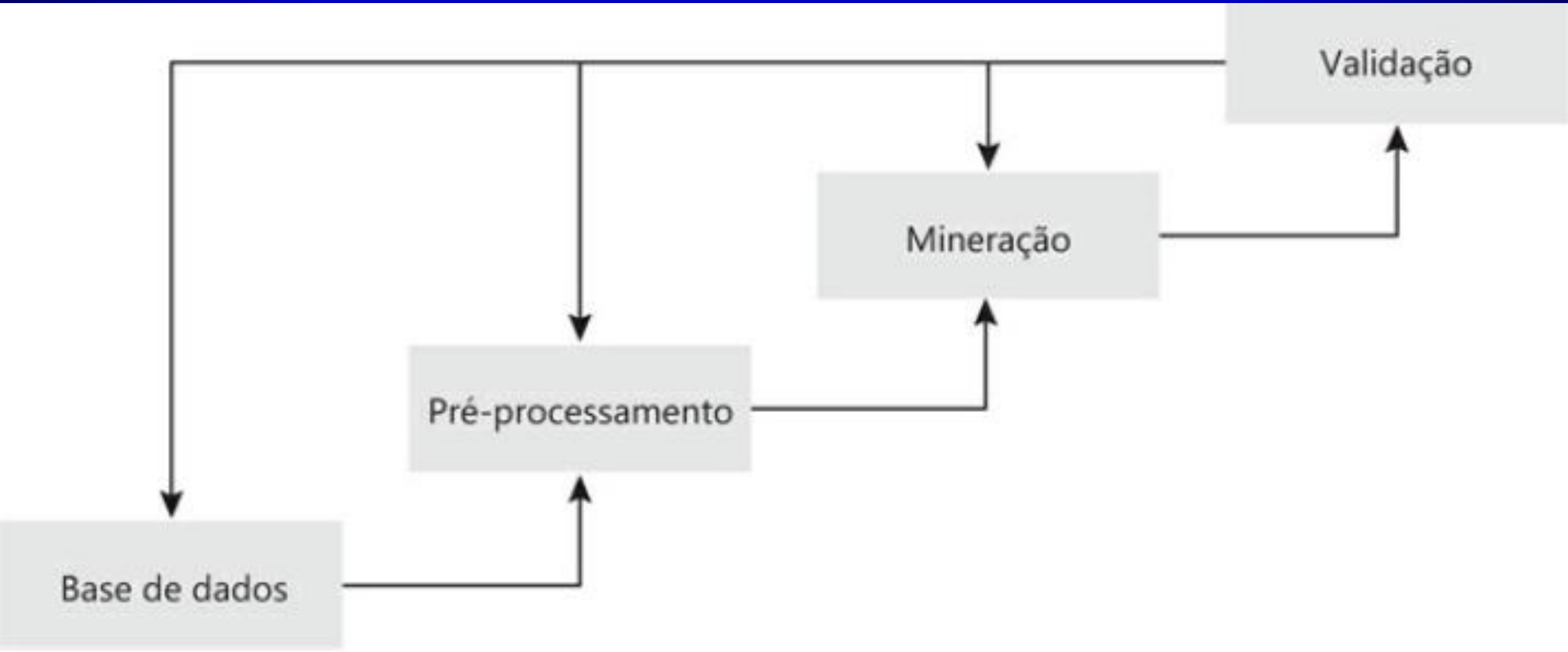
PUC-SP

Mineração de dados: essa etapa do processo corresponde à aplicação de algoritmos capazes de extrair conhecimentos a partir dos dados pré-processados. Serão discutidas técnicas de *análise descritiva* (medidas de distribuição, tendência central e variância, e métodos de visualização), *agrupamento* (segmentação de bases de dados) e *detecção de anomalias e outliers*; e

Avaliação ou validação do conhecimento: avaliação dos resultados da mineração objetivando identificar conhecimentos verdadeiramente úteis e não triviais.

Processo de descoberta de conhecimento em bases de dados

PUC-SP



Essas quatro etapas são correlacionadas e interdependentes de tal forma que a abordagem ideal para extrair informações relevantes em bancos de dados consiste em considerar as inter-relações entre cada uma dessas etapas e sua influência no resultado final

Aqui, vale a pena ressaltar que, sob uma perspectiva de armazenamento de dados (*data ware house*), o processo de mineração de dados pode ser visto como um estágio avançado do processamento analítico *online* (*on-line analytical processing – OLAP*).

Entretanto, a mineração de dados vai muito além do escopo restrito típico de um OLAP, baseado em métodos de resumo ou sumarização de dados, incorporando técnicas mais avançadas para a compreensão e a extração de conhecimentos dos dados.

A Mineração de Dados é uma disciplina interdisciplinar e multidisciplinar que envolve conhecimento de áreas como:

- Banco de dados
- Estatística
- Matemática
- Aprendizagem de Máquina
- Computação de alto desempenho

PUC-SP

PUC-SP

- Reconhecimento de padrões
- Visualização de dados
- Recuperação de informação
- Processamento de imagens e de sinais
- Análise espacial de dados
- Inteligência artificial , entre outras.

PUC-SP

Multidisciplinaridade da mineração de dados

Principais tarefas da Mineração de Dados

PUC-SP

As funcionalidades da mineração de dados são usadas para especificar os tipos de informações a serem obtidas nas tarefas de mineração. Em geral, essas tarefas podem ser classificadas em duas categorias:

- ***Descritivas*:** caracterizam as propriedades gerais dos dados; e
- ***Preditivas*:** fazem inferência a partir dos dados objetivando predições.

Em muitos casos, o usuário não tem ideia do tipo de conhecimento contido nos dados ou como usá-lo para gerar modelos preditivos, tornando importante a capacidade das ferramentas de mineração em encontrar diferentes tipos de conhecimento. As principais tarefas de mineração de dados são descritas sucintamente aqui.

PUC-SP

Análise Descritiva de Dados: Os algoritmos de aprendizagem de máquina são ferramentas poderosas para a descoberta de conhecimentos em bases de dados. Entretanto, uma etapa inicial do processo de mineração que não requer elevado nível de sofisticação é a *análise descritiva dos dados*.

Especificamente, essas análises permitem investigar a *distribuição de frequência*, as *medidas de centro e variação*, e as *medidas de posição relativa e associação* dos dados.

Além disso, técnicas elementares de *visualização* também são empregadas para um melhor entendimento da natureza e distribuição dos dados.

PUC-SP

Predição: Classificação e Estimação

Predição é uma terminologia usada para se referir à construção e ao uso de um modelo para avaliar a classe de um objeto não rotulado ou para estimar o valor de um ou mais atributos de dado objeto. No primeiro caso, denominamos a tarefa de *classificação* e, no segundo, denominamos de *regressão* (em estatística) ou simplesmente *estimação*.

Sob essa perspectiva, *classificação e estimação* constituem os dois principais tipos de problemas de predição, sendo que a classificação é usada para predizer *valores discretos*, ao passo que a estimação é usada para predizer *valores contínuos*.

Para exemplificar, considere o problema de atribuição de crédito, no qual um cliente se dirige a uma instituição financeira com o objetivo de conseguir um financiamento para trocar seu veículo.

A primeira pergunta a ser respondida corresponde a uma tarefa de classificação: o crédito será oferecido ou não?

Em seguida, há outra pergunta que pode ser relevante responder: qual o valor do crédito a ser oferecido?

PUC-SP

Essa última é uma estimação e ela faz sentido na medida em que o sistema de predição percebe que o cliente possui uma capacidade de pagamento superior ao que está sendo solicitado ou que o valor solicitado é muito alto, mas pode ser ajustado à sua capacidade financeira.

Nesse caso, uma ferramenta capaz de estimar a capacidade de pagamento do cliente pode gerar maior lucro ou segurança para a empresa financiadora.

Como os rótulos das classes dos dados de treinamento são conhecidos *a priori* e usados para ajustar o modelo de predição, esse processo é denominado *treinamento supervisionado* (ou *aprendizagem supervisionada*).

Exemplos de tarefas de classificação incluem identificação de *spams*, classificação de objetos, atribuição de crédito e detecção de fraudes.

Exemplos de tarefas de estimação incluem predição de produtividade de grãos, estimativa de desempenho de atletas, estimativa de crédito, estimativa de valores futuros em bolsas de valores e previsão do clima, ...

Análise de Grupos

PUC-SP

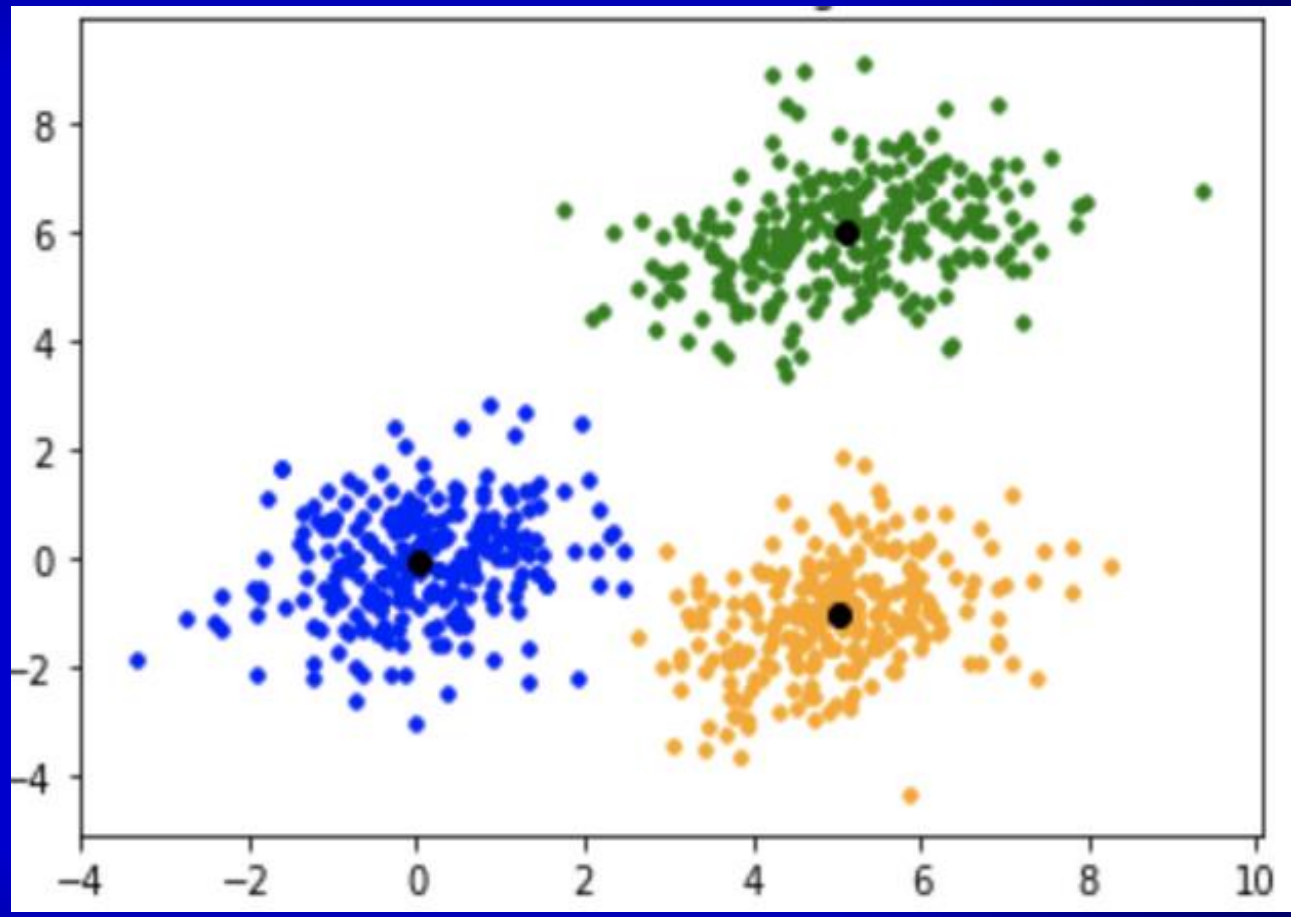
***Agrupamento* (*clustering*)** é o nome dado ao processo de separar (particionar ou segmentar) um conjunto de objetos em *grupos* (do inglês *clusters*) de objetos similares.

Diferentemente da tarefa de classificação, o agrupamento de dados considera dados de entrada não rotulados, ou seja, o grupo (classe) ao qual cada dado de entrada (objeto) pertence não é conhecido *a priori*. O processo de agrupamento (ou *clusterização*) normalmente é utilizado para identificar tais grupos e, portanto, cada grupo formado pode ser visto como uma classe de objetos. Como os rótulos das classes dos dados de treinamento não são conhecidos *a priori*, esse processo é denominado *treinamento não supervisionado* (ou ***aprendizagem não supervisionada***).

Análise de Grupos

Em um processo de agrupamento, os objetos são agrupados com o objetivo de maximizar a distância interclasse e minimizar a distância intraclasse, ou, dito de outra forma, maximizar a similaridade intraclasse e minimizar a similaridade interclasse.

Portanto, um *cluster* pode ser definido como uma coleção de objetos similares uns aos outros e dissimilares aos objetos pertencentes a outros *clusters*.



PUC-SP

Para ilustrar uma tarefa de agrupamento, considere o problema de segmentar uma base de dados descrevendo frutas, na qual cada fruta está descrita por um conjunto de atributos, como forma, cor e textura.

Suponha que haja maçãs e bananas nessa base de dados e que o algoritmo precisa segmentá-los sem ter conhecimento algum sobre a classe da fruta, recebendo apenas informações dos atributos.

Como a forma, cor e textura das bananas são substancialmente diferentes da forma, cor e textura das maçãs, durante o agrupamento o algoritmo deverá, naturalmente, colocar bananas em um grupo e maçãs em outro.

Associação

PUC-SP

Nas análises de grupos e preditivas, o objetivo em geral é encontrar relações (grupos, classes ou estimativas) entre os objetos da base. Entretanto, há diversas aplicações práticas nas quais o objetivo é encontrar relações entre os atributos (ou variáveis), e não entre os objetos.

Para ilustrar esse caso, vamos considerar uma aplicação típica em marketing: a análise de carrinho de supermercado.

Nesse tipo de análise, há um conjunto de transações (pedidos ou compras), e o objetivo é encontrar itens (produtos) que são comprados em conjunto; nesse sentido, as transações correspondem aos objetos da base e os itens, aos atributos.

PUC-SP

A *análise por associação*, também conhecida por ***mineração de regras de associação***, corresponde à descoberta de regras de associação que apresentam valores de atributos que ocorrem concomitantemente em uma base de dados.

Esse tipo de análise costuma ser usado em ações de marketing e para o estudo de bases de dados transacionais.

Há dois aspectos centrais na *mineração de regras de associação*: a proposição ou *construção* eficiente das regras de associação e a quantificação da *significância* das regras propostas. Ou seja, um bom algoritmo de mineração de regras de associação precisa ser capaz de propor associações entre itens que sejam estatisticamente relevantes para o universo representado pela base de dados.

Formalmente, as regras de associação possuem a forma: $X \rightarrow Y$

$$A_1 , A_2 , \cdots , A_m \rightarrow B_1 , B_2 , \cdots , B_m$$

Onde: A_i ***e*** B_i **são pares de valores de atributos** $(i = 1 , 2 , \ldots , m)$

A relação $X \rightarrow Y$ é interpretada assim: registros da base de dados que satisfazem à condição em X também satisfazem à condição em Y.

Um exemplo com smartphones: $X =$ *smartphone* e $Y =$ plano de dados. Sabe-se que há uma *confiança* de 90% de que um cliente que compra um *smartphone* também assinará um plano de dados

Essa informação é estratégica para o negócio, pois pode induzir, por exemplo, promoções conjuntas de *smartphones* e planos de dados.

Detecção de anomalias

PUC-SP

Uma base de dados pode conter objetos que não seguem o comportamento ou não possuem a característica comum dos dados ou de um modelo que os represente. Esses dados são conhecidos como *anomalias* ou *valores discrepantes* (*outliers*).

A maioria das ferramentas de mineração descarta as anomalias – por exemplo, ruídos ou exceções –, entretanto, em algumas aplicações, como na detecção de fraudes, os eventos raros podem ser mais informativos do que aqueles que ocorrem regularmente.

As anomalias podem ser detectadas de diversas formas, incluindo métodos estatísticos que assumem uma distribuição ou modelo de probabilidade dos dados, ou medidas de distância por meio das quais objetos substancialmente distantes dos demais são considerados anomalias.

Por exemplo, no caso de fraudes em cartões de crédito, valores muito acima dos usuais para um dado cliente, assim como o tipo, o local e a frequência de uma dada compra, são indicativos de uma possível anomalia.

PUC-SP

Uma característica marcante das anomalias é que elas compõem uma classe que ocorre com uma frequência bem inferior às das classes normais.

Isso faz com que os algoritmos de classificação e suas respectivas medidas de avaliação sejam fortemente impactados, forçando o uso de algoritmos e medidas de desempenho desenvolvidos especificamente para tratar tais problemas.

Ao mesmo tempo, a vasta amplitude de problemas nessa área e sua relevância prática motivam a discussão em separado do tema.

Dicas para uma análise eficiente e eficaz

A mineração pode levar a uma capacidade preditiva e analítica poderosa dos dados. Mesmo quando aplicada de maneira correta, a capacidade de trabalhar com múltiplas variáveis e suas relações pode tornar os processos de mineração e interpretação dos resultados substancialmente complexos.

Considerando essa complexidade, é preciso que o *analista de dados*, também conhecido como *cientista de dados*, esteja atento aos fundamentos conceituais necessários para o uso e o entendimento de cada técnica.

A seguir, apresentamos uma lista de considerações (inevitavelmente incompleta) que podem servir como guia para uma mineração eficiente e eficaz:

Estabelecer a significância da mineração: tanto a significância estatística quanto a prática da mineração devem ser consideradas.

A *significância estatística* está relacionada à confiabilidade dos resultados obtidos, ou seja, se a base de dados foi preparada adequadamente para a análise, se os resultados apresentados são coerentes e se os algoritmos propostos tem o desempenho desejado, pois uma amostragem ou normalização inadequada da base pode gerar resultados que não tenham nenhuma significância estatística e que, portanto, são inúteis.

A *significância prática*, por sua vez, questiona sobre a aplicabilidade prática das análises realizadas, ou seja, se essas análises podem ser usadas em algum processo de tomada de decisão.

As características da base de dados influenciam todos os resultados: O processo de mineração opera, quase em sua totalidade, sobre uma base de dados pré-processada. Assim, é importante reconhecer que a quantidade de objetos na base, a dimensão (número de atributos) desses objetos, o tipo de atributos e seus domínios, a ausência de valores na base, as inter-relações entre os atributos e muitas outras características dos dados afetarão fortemente o resultado da análise.

Necessidade de conhecer os dados: A discussão apresentada implica que análises preliminares dos dados – mais especificamente as técnicas de análise descritiva, como medidas de tendência central (por variável), análise de componentes principais e muitos outros métodos (estatísticos) simples devem ser aplicados à base com o objetivo de entendê-la melhor antes de se iniciar a mineração propriamente dita.

Buscar pela parcimônia: Boa parte dos algoritmos de mineração resulta em uma espécie de modelo dos dados que poderá ser utilizado posteriormente para fazer alguma inferência ou predição.

É possível que a escolha de diferentes amostras dos dados, ou mesmo diferentes execuções dos algoritmos, resultem em modelos com características distintas.

Nesses casos, a escolha por um ou outro modelo deve considerar, entre outros aspectos, a parcimônia da solução, ou seja, a complexidade do modelo resultante. Muitas vezes, a complexidade de criação do modelo é um aspecto crucial na escolha de uma ferramenta dentro de um conjunto de possibilidades

Verificar os erros: todos os algoritmos de mineração podem ter seu desempenho avaliado.

No caso dos algoritmos de agrupamento, há medidas que permitem avaliar a qualidade dos agrupamentos propostos:

- Nas tarefas de predição, é possível avaliar o erro de predição;
- Na mineração de regras de associação, avalia-se a significância das regras e
- Para os algoritmos de detecção de anomalias, verifica-se o seu desempenho por meio de medidas específicas para esse tipo de problema.

Em todos os casos, é preciso fazer um diagnóstico de desempenho do algoritmo, identificando os erros, o porquê de sua ocorrência, e empregar esse conhecimento para realimentar o processo de análise.

Validar seus resultados: Os resultados de uma análise precisam ser validados de diversas formas, por exemplo:

- Comparando com o resultado de outras técnicas
- Analisando a capacidade de generalização dos métodos
- Combinando com outras técnicas e
- Até utilizando um especialista de domínio capaz de validar se os resultados apresentados fazem sentido e se são de boa qualidade.

Assim como no caso anterior, a validação é central para realimentar o processo de análise de dados.

PUC-SP

As diferentes nomenclaturas

A literatura está cheia de diferentes nomenclaturas para as muitas técnicas de solução de problemas e algoritmos computacionais que surgiram nas últimas décadas.

Como há profissionais de diversas área envolvidos com essas técnicas, naturalmente surgem nomenclaturas distintas para contextos muitas vezes comuns, causando confusão e dificuldade de entendimento.

Dentre as muitas nomenclaturas disponíveis na literatura técnico-científica merecem destaque as seguintes:

- ***Inteligência artificial***
- ***Inteligência computacional***
- ***Aprendizagem de máquina* e**
- ***Computação natural*.**

Além dessas, uma nova terminologia chamada de ***big data***, vem sendo amplamente usada, sobretudo no mundo empresarial.

Faremos aqui uma breve descrição do significado de cada uma dessas nomenclaturas, justificando, em alguns casos, o porquê de sua proposição.

PUC-SP

Inteligência artificial clássica:

J. McCarthy, um dos pioneiros da *inteligência artificial* (IA), define a área como a ciência e engenharia de máquinas inteligentes, especialmente programas inteligentes de computador. Ela está relacionada à tarefa de usar computadores para entender a inteligência humana, mas se restringindo necessariamente aos métodos inspirados na biologia.

Outra definição muito usada para a IA é aquela apresentada no livro de S. Russel e P. Norvig: eles definem a IA como uma tentativa de entender e construir entidades inteligentes, e uma razão para estudá-la é aprender mais sobre nós mesmos.

Nota-se que o foco dessas definições e a fonte básica de inspiração para o desenvolvimento da IA era a inteligência humana, nossa capacidade de percepção, resolução de problemas, comunicação, aprendizagem, adaptação e muitas outras

PUC-SP

As técnicas mais tradicionais de inteligência artificial, que surgiram na década de 1950 e prevaleceram até a década de 1980, ficaram conhecidas como *IA clássica*.

Elas eram essencialmente *simbólicas*, ou seja, propunham que uma manipulação algorítmica de estruturas simbólicas – por exemplo, palavras – seria necessária e suficiente para o desenvolvimento de sistemas inteligentes.

Essa tradição simbólica engloba também as abordagens baseadas em *lógica*, nas quais os símbolos são utilizados para representar objetos e relações entre objetos, e estruturas simbólicas são utilizadas para representar fatos conhecidos.

Uma característica marcante da IA clássica era a forma utilizada para construir o sistema inteligente.

Existia uma visão procedural sugerindo que sistemas inteligentes poderiam ser projetados codificando-se conhecimentos especialistas em algoritmos específicos.

Esses sistemas foram denominados genericamente *sistemas baseados em conhecimento* (*knowledge-based systems*) ou *sistemas especialistas* (*expert systems*).

Um exemplo clássico de sistema especialista é para diagnóstico médico, em que a ideia central é que se faça uma representação simbólica do conhecimento do médico acerca de um estudo específico e, a partir de então, o diagnóstico é feito com base na relação das regras representadas nesse modelo.

Atualmente, a IA clássica envolve basicamente os sistemas especialistas, diversos métodos de busca – como busca em profundidade e busca em largura –, alguns sistemas baseados em agentes e sistemas de raciocínio ou inferência baseados em lógica.

Aprendizagem de máquina

Em seu livro pioneiro, T. Mitchell define *aprendizagem de máquina* (AM) como a área de pesquisa que visa desenvolver programas computacionais capazes de automaticamente melhorar seu desempenho por meio da experiência.

A área de AM está baseada em conceitos e resultados de muitas outras áreas, como estatística, inteligência artificial, filosofia, teoria da informação, biologia, ciências cognitivas, complexidade computacional e teoria de controle.

Seguindo uma linha similar, Alpaydin define a aprendizagem de máquina como a programação de computadores para otimizar um critério de desempenho usando experiências passadas, chamadas de exemplos ou simplesmente dados de entrada. A ideia é que as técnicas envolvidas na AM sejam capazes, de alguma forma, de *aprender a resolver os problemas*.

A aprendizagem de máquina tem como foco extrair informação a partir de dados de maneira automática.

Portanto, ela está intimamente relacionada à mineração de dados, à estatística, à inteligência artificial e à teoria da computação, além de outras áreas como computação natural, sistemas complexos adaptativos e computação flexível, como veremos a seguir.

Os principais métodos investigados em aprendizagem de máquina são aqueles que trabalham com dados nominais, como as *árvores de decisão*, as *regras de associação* e *classificação*, tabelas de decisão e outros. Além desses, destacam-se os algoritmos baseados na *Teoria de Bayes*, alguns *métodos estatísticos* e *métodos de agrupamento de dados*

Paradigmas de aprendizagem

No contexto de mineração de dados, *aprendizagem* ou *treinamento* corresponde ao processo de ajuste e/ou construção do modelo usando um mecanismo de apresentação ou uso dos objetos da base de dados.

Por exemplo, em uma árvore de decisão, o treinamento consiste em escolher atributos da base de dados que comporão cada nível de nós da árvore e construir os ramos de forma que otimize algum critério de qualidade; no algoritmo de agrupamento das *k*-médias, o treinamento consiste em apresentar os objetos da base de dados e ajustar a posição de um conjunto de vetores, chamados protótipos, que representam os grupos de objetos da base.

Há casos em que a aprendizagem só ocorre no momento do uso do sistema e, portanto, não é realizado um ajuste ou construção prévia do modelo.

Normalmente, o algoritmo armazena toda a base de dados e a usa para inferir algo a respeito dos novos objetos dos quais se deseja obter alguma informação, como classe a que pertencem – esse tipo de aprendizagem é denominado *aprendizagem preguiçosa* (*lazy learning*).

Um exemplo de algoritmo que opera dessa forma é o algoritmo dos *k*- vizinhos mais próximos (*k*-NN, do inglês *k Nearest Neighbors*). O *k*-NN opera da seguinte maneira: dado um objeto cuja classe se deseja conhecer, esse objeto é comparado com todos os objetos da base de dados e sua classe é tomada como aquela dos *k* objetos mais próximos (similares) a ele.

Um procedimento bem definido para treinar uma técnica de aprendizagem de máquina é denominado *algoritmo de aprendizagem* ou *algoritmo de treinamento*, e a maneira pela qual o ambiente influencia a técnica em seu aprendizado define o *paradigma de aprendizagem*. Os dois paradigmas de aprendizagem mais comuns, são:

Aprendizado supervisionado: é baseado em um conjunto de objetos para os quais as saídas desejadas são conhecidas, ou em algum outro tipo de informação que represente o comportamento que deve ser apresentado pelo sistema;

Aprendizado não supervisionado (é o que estudaremos nesta disciplina): é baseado apenas nos objetos da base, cujos rótulos são desconhecidos. Basicamente, o algoritmo deve aprender a "categorizar" ou rotular os objetos.

Exemplos de Aplicação: Há uma vasta literatura sobre aplicações de técnicas de mineração de dados em problemas nas mais variadas áreas, tais como:

- **Análise e predição de crédito**
- **Detecção de fraudes**
- **Predição do mercado financeiro**
- **Relacionamento com clientes**
- **Predição de falência corporativa e muitas outras.**

PUC-SP

Exemplos de segmentos de aplicação, incluem:

- **Setor financeiro;**
- **Planejamento estratégico empresarial;**
- **Planejamento do setor portuário;**
- **Setores de energia (petróleo, gás, energia elétrica, biocombustíveis etc.);**
- **Educação;**
- **Logística;**
- **Planejamento das cadeias de produção, distribuição e suprimentos;**
- **Meio ambiente;**
- **Internet (portais, redes sociais, comércio eletrônico etc.).**

PUC-SP

Aplicações típicas incluem:

- **Identificação ou segmentação de clientes, parceiros, colaboradores;**
- **Detecção de fraudes e anomalias em sistemas e processos;**
- **Ações estratégicas de marketing,**
- **RH;**
- **Jogos e atividades educacionais;**
- **Gestão do conhecimento;**
- **Análise de padrões de consumo;**
- **Compreensão de bases de dados industriais, biológicas, empresariais e acadêmicas;**
- **Predição de retorno sobre investimento, despesas, receitas, investimentos etc. e**
- **Mineração de dados da web.**

Predição de produtividade de grãos: Com relação a valor econômico, o Brasil é o quarto maior exportador de produtos agropecuários do mundo, ficando atrás apenas da União Europeia (composta por 27 países), seguida dos Estados Unidos e Canadá.

Depois do Brasil estão a China, a Austrália, a Tailândia e a Argentina.

Nossa produção de grãos vem crescendo vertiginosamente ao longo da história, passando de 46.943 mil toneladas em 1976-1977 para 328,4 milhões de toneladas em 2025 (estimada).

Apesar do crescente aumento da produção de grãos no país, o setor ainda sofre com altas dívidas, baixo custo do grão, como matéria--prima, quando comparada a produtos industrializados, infraestrutura deficiente e pouco uso de tecnologia.

PUC-SP

Algoritmos de estimação podem ser utilizados para prever a produtividade de grãos em lavouras, o que é muito importante principalmente em um país que ainda possui boa parte de sua balança comercial equilibrada pela exportação de produtos agrícolas.

Estimar, ou seja, prever o resultado de uma colheita pode ajudar na indicação de técnicas para a correção do solo, adequação dos processos de irrigação e melhoria do controle de pragas, tudo isso feito antes da colheita e, portanto, evitando possíveis prejuízos financeiros e até ambientais.

PUC-SP

Uma maneira de empregar análise de dados para predição de colheita é utilizando amostras de folhas de plantas e amostras do solo da região para o treinamento dos algoritmos de predição.

Para cada amostra de solo e folha da lavoura, podem-se obter as composições químicas com relação à concentração de alumínio, chumbo, potássio, magnésio, manganês e enxofre.

Além dessas, também se pode considerar o pH do solo. O objetivo da predição é determinar a quantidade de calcário necessária para neutralizar o alumínio tóxico no solo, aumentar o cálcio, o magnésio e a base do solo.

Com essas informações torna-se possível fazer a correção do solo antes da colheita.

PUC-SP

Análise de sentimento em redes sociais

O poder da interação interpessoal em ambientes virtuais vem direcionando o mercado e promovendo a criação de novas empresas de internet.

Consequentemente, inúmeros projetos de pesquisa e desenvolvimento surgiram com o objetivo de dar suporte a essas redes sociais tanto sob o ponto de vista tecnológico, quanto de modelo de negócios.

Redes sociais como Instagram, Facebook, LinkedIn, Twitter e outras explodiram em popularidade nos últimos anos, também impulsionadas pelo aumento do poder aquisitivo da população mundial e pelas melhorias e redução de custos de toda a infraestrutura de comunicação.

Dentre as técnicas aplicáveis pode-se destacar a classificação de textos, que busca rotular um documento de acordo com suas características.

A *análise de sentimento*, também conhecida como *mineração de opinião*, é um tipo de classificação de textos que objetiva rotulá-los de acordo com o sentimento ou a opinião neles contidos.

Classificar um texto de acordo com o sentimento que o usuário desejou passar, por exemplo, *positivo*, *negativo* ou *neutro*, permite o dimensionamento do retorno sobre determinado produto, serviço, marca, empresa, etc.

PUC-SP

Para citar alguns poucos exemplos:

- **Consumidores podem usar a análise de sentimento para pesquisar sobre determinado produto ou serviço**
- **Empresas de marketing podem mensurar a opinião pública sobre uma campanha**
- **Empresas podem analisar críticas em uma nova versão de seu produto.**

É comum atualmente encontrarmos empresas especializadas no monitoramento e gerenciamento de redes sociais.

PUC-SP

Detecção de fraudes em cartões de crédito

Atualmente, vários tipos de transações comerciais ocorrem via cartões de crédito, como é o caso, por exemplo, de compras feitas pela internet e em lojas de comércio eletrônico.

O governo norte-americano estimou que, no final do século passado, aproximadamente 13 bilhões de dólares foram gastos em compras pela internet apenas usando cartões de crédito.

O medo de transportar dinheiro no bolso e a crescente quantidade de estabelecimentos que aceitam cartões de crédito contribuem para uma utilização massiva de cartões por empresas e cidadãos comuns.

PUC-SP

As fraudes em cartões de crédito podem ser divididas em duas grandes categorias: *fraudes comportamentais* e *fraudes de aplicação*.

As fraudes de aplicação ocorrem quando um indivíduo adquire um novo cartão de crédito utilizando informações pessoais falsas e, em seguida, gasta o máximo que pode em um curto intervalo de tempo.

As fraudes comportamentais, por outro lado, são aquelas que ocorrem quando os detalhes (dados) de um usuário legítimo são obtidos e usados de forma fraudulenta; ou seja, transações ilegítimas são autorizadas sem ser detectadas pelas administradoras. Essas fraudes podem ser resultado da interceptação de cartões de crédito enviados pelo correio, pela perda ou pelo roubo de um cartão, ou simplesmente pela aquisição e uso não autorizado de dados de um usuário legítimo.

PUC-SP

No combate às fraudes, as ações da empresa administradora também podem ser agrupadas em duas grandes categorias: ***prevenção* e *detecção***.

A prevenção consiste em medidas que visam impedir a ocorrência de fraudes, como a necessidade de uso de senhas pessoais e sistemas de segurança para transações via web.

A **detecção** de fraudes envolve a identificação rápida e eficiente de transações ilegítimas. É possível realizar a detecção de fraudes usando informações tanto sobre o padrão de comportamento normal de um usuário legítimo quanto usando dados sobre fraudes. Mesmo assim, a detecção de fraudes é um problema altamente complexo e desafiador em virtude de uma série de características:

PUC-SP

- a quantidade de transações que são feitas diariamente é muito alta;
- o padrão de comportamento de um usuário legítimo pode mudar repentinamente (por exemplo, em viagens);
- a quantidade de transações legítimas é muito superior à quantidade de transações ilegítimas;
- os fraudadores adaptam constantemente seus comportamentos de acordo com a sofisticação dos sistemas de detecção; e
- diferentes transações envolvem diferentes quantias e, portanto, representam variáveis perdas potenciais.

Combate a perdas não técnicas de energia elétrica

PUC-SP

A existência de perdas em um sistema de energia elétrica é consequência natural do consumo de energia.

As perdas podem ser categorizadas de acordo com o efeito, componente do sistema, ou causa e podem ser resumidas em:

Perdas técnicas: correspondem àquelas perdas intrínsecas ao sistema elétrico, o que inclui as perdas nos equipamentos, na transformação e na distribuição da energia.

Perdas comerciais: também chamadas de perdas não técnicas, são consequência, principalmente, de erros ou ausência de medição, medidores com defeito, consumidores clandestinos, desvio de consumo e furto de energia.

PUC-SP

Uma das formas de reduzir as perdas comerciais é realizar inspeções técnicas no local de consumo em busca de irregularidades, que vão desde a adulteração dos dispositivos de medição (fraude) até o furto ou desvio da energia propriamente dita.

Entretanto, além da impossibilidade de inspecionar todos os consumidores, o custo associado à inspeção é alto, uma vez que esse processo demanda tempo, requer o deslocamento de uma equipe em campo e muitos dos consumidores inspecionados não são fraudadores.

Com base nos dados de fiscalização obtidos a partir de medidas amostrais em campo, pode ser feita uma análise de dados para investigar inter-relações entre as amostras, segmentando os dados em grupos hierarquicamente vinculados, permitindo uma definição de pontos estratégicos de fiscalização.

Outra tarefa possível é a classificação automática dos cadastros disponíveis, a partir da qual se pode desenvolver um sistema de classificação que permita identificar de modo automático aqueles consumidores que provavelmente estejam causando perda de receita para a concessionária.

Trata-se, portanto, de uma etapa na qual é feita a prospecção de possíveis perdas comerciais. Essa informação pode ser empregada no direcionamento das equipes de fiscalização e auditoria, impactando diretamente na redução das perdas não técnicas.

Além dessas análises, dado o perfil de consumo dos usuários pode ser feito um levantamento das curvas típicas de hábito de consumo, permitindo uma identificação automática de novos clientes e de anomalias em clientes já existentes

Segmentação de curvas de carga em sistemas de energia elétrica

PUC-SP

Apesar do alto grau de desenvolvimento tecnológico da atualidade, só é possível armazenar energia elétrica em pequenas quantidades utilizando, para isso, baterias.

No caso da energia elétrica consumida pelas indústrias, empresas e residências, a capacidade produtiva das usinas deve ser aproximadamente igual à quantidade de energia consumida.

A pergunta que as usinas geradoras precisam responder, portanto, é qual será o consumo de energia elétrica a cada dia.

Nesse contexto, é necessário realizar a predição da demanda de energia elétrica para que uma quantidade suficiente seja produzida.

PUC-SP

A falta de planejamento e de investimentos no setor produtivo de energia elétrica pode causar apagões, cortes indesejáveis no fornecimento de energia, podendo até paralisar a produção industrial e deteriorar o desempenho de outros serviços.

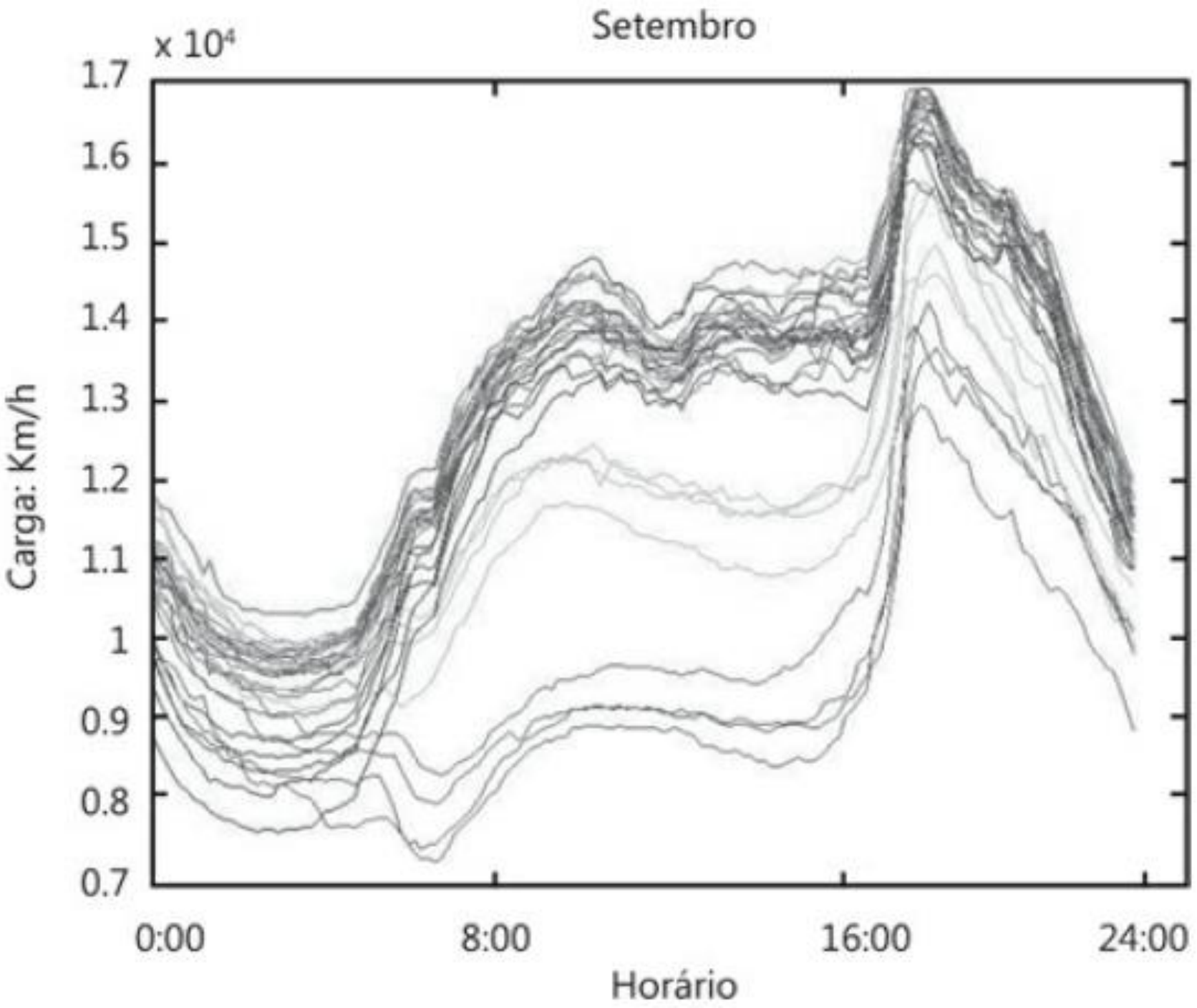
No Brasil, já tivemos alguns apagões ocorridos há alguns anos em decorrência da falta de planejamento ou outros problemas na geração ou distribuição da energia, o que levou o governo a estimular o racionamento voluntário, promovendo a economia e penalizando o desperdício de energia elétrica.

PUC-SP

Com o objetivo de melhorar o planejamento da produção de energia elétrica, é possível usar técnicas de análise de dados para a previsão de carga (consumo) em curto, médio e longo prazos de um sistema elétrico de potência.

No caso específico do curto prazo, para prever as cargas horárias de um dia, o padrão de carga horária e as cargas máxima e mínima devem ser determinados.

Suponha que o objetivo inicial seja identificar dias da semana com padrões de cargas horárias similares e, posteriormente, realizar a previsão de demanda do setor. A previsão de demanda de carga é um meio de fornecer informações para uma tomada de decisão criteriosa que proporciona economia e segurança no fornecimento de energia elétrica. Para isso, uma companhia elétrica precisa resolver vários problemas técnicos e econômicos no planejamento e controle da operação do sistema de energia elétrica.



PUC-SP

Curvas de carga (consumo) de energia elétrica ao longo de um mês

Para criar um modelo de segmentação de padrões de carga em sistemas de energia elétrica, pode-se utilizar uma base de dados referente ao consumo diário em determinados períodos do ano, como ilustrado na Figura .

Ao observarmos os perfis dessas curvas de carga, notamos a existência de padrões típicos de consumo.

Classificar os perfis de carga anteriormente à previsão permite uma previsão de demanda mais precisa e exemplifica o fato de que as técnicas a serem discutidas aqui podem ser usadas em conjunto para se atingir um objetivo final.

Nesse caso, um algoritmo de agrupamento é empregado antes de um algoritmo de predição.

Modelagem de processos siderúrgicos

PUC-SP

Boa parte das siderúrgicas está equiparada tecnologicamente, sendo o uso eficiente do conhecimento um diferencial importante.

O principal desafio é alcançar a excelência operacional pelo uso de tecnologias baseadas na experiência dos processos adquirida pelas pessoas e indústrias.

As indústrias siderúrgicas investem esforços no desenvolvimento de tecnologias e dispositivos capazes de aumentar a produtividade das usinas, como a *sub-lança a oxigênio*, uma ferramenta importante para o controle do processo de *conversores* (fornos basculantes que têm a função de transformar a matéria-prima em aço líquido)

PUC-SP

A sub-lança é basicamente utilizada para medir o teor de carbono e a temperatura do aço durante o sopro de oxigênio, além de permitir retirar uma amostra que é enviada ao laboratório para análise detalhada da composição química do aço.

A medição e a amostragem são realizadas antes do final do sopro de oxigênio e modelos matemáticos baseados nessa informação são utilizados para estimar a composição química que será obtida e, assim, redundar em ações corretivas do processo produtivo.

PUC-SP

Algoritmos de mineração de dados podem ser usados para prever os principais elementos químicos (carbono, manganês, fósforo e enxofre) da análise de final de sopro sem utilizar os resultados da amostra da sub-lança.

Essa solução permite uma redução do tempo de espera entre o recebimento do resultado da análise do laboratório e a execução do modelo de vazamento e pesagem das ferroligas.

Dessa forma, a solução antecipa o final do tratamento nos conversores, possibilitando uma padronização, continuidade e uniformidade da operação, reduzindo o tempo de tratamento no conversor e aumentando a produtividade.