INSTITUTO FEDERAL DE
EDUCAÇÃO, CIÊNCIA E TECNOLOGIA
SÃO PAULO
Campus Birigui

Mineração de Dados

Apresentação da Disciplina

Professor Dr. Murilo Varges da Silva
murilo.varges@ifsp.edu.br

REDE FEDERAL DE EDUCAÇÃO PROFISSIONAL E TECNOLÓGICA

Ementa

- A disciplina visa apresentar conceitos, técnicas e ferramentas relevantes na área de descoberta de conhecimento em bases de dados e mineração de dados, sob a perspectiva da área de bases de dados, explorando com profundidade os aspectos associados a esta área.

Objetivos

- Apresentar os conceitos do processo de descoberta de conhecimento em grandes bases de dados.
- Este processo busca padrões intrínsecos a estes dados, de forma a descobrir informações não óbvias que possam ser relevantes e dar suporte à tomada de decisão.
- Entender a aplicação e saber quando e como utilizar técnicas de preparação de dados e os principais métodos de mineração.
- Apresentar conceitos e técnicas de visualização de dados, além de exemplos de ferramentas para mineração de dados.

Conteúdo Programático

- Introdução aos conceitos do processo de descoberta de conhecimento.
- Técnicas de preparação de dados. Técnicas de redução de dados.
- Tarefas e técnicas de mineração de dados: classificação, regressão, detecção de agrupamentos, descoberta de regras de associação, sumarização, modelagem de dependências, detecção de tendências e exceções.

Conteúdo Programático

- Conceitos e técnicas de visualização de dados. Mineração de dados visual.
- Exemplos de mineração de dados em domínios específicos, tais como: bioinformática, sistemas de informação geográfica, bases de dados de imagens, bases de dados de documentos textuais, bases de dados multimídia, sistemas Web, bases de dados espaço-temporais, sistemas baseados em data streams.
- Ferramentas de mineração de dados.

Critérios Avaliação

A Nota do Componente Curricular, NC, será calculada a partir das avaliações teóricas e práticas, apresentação de listas de exercício e trabalhos produzidos individual e em grupo a serem aplicadas ao longo do semestre.

$MD_{avaliações} = (A_1, A_2, \ldots A_n)$ → **Peso 5**

$MD_{projeto} = (P_1, P_2, \ldots P_N)$ → **Peso 5**

$$\mathbf{NC = MD_{avaliações} + MD_{projeto}}$$

Se **NC** >= 6 → **Aprovado**

Se **NC** entre 4 e 6 → **Instrumento Final de Avaliação (IFA)**

Se **NC** < 4 → **Retido**

Se **FCC*** < 75% → **Retido**

* Frequência

REDE FEDERAL DE EDUCAÇÃO PROFISSIONAL E TECNOLÓGICA

Mineração de Dados

“Porque a diferença entre dado, informação e conhecimento é maior que nossa capacidade de transformar um no outro de maneira rápida e precisa, principalmente para massas de dados com grandes volumes, variedades e velocidades”

Leandro Nunes de Castro – Livro Introdução à Mineração de Dados

Mineração de Dados

- A mineração de dados surgiu como área de pesquisa e aplicação independente em meados da década de 1990;
- Origens na matemática, estatística e computação são muito anteriores a esse período;
- A área também ganhou evidência nos últimos anos depois de ser cunhado o termo Big Data.

Mineração de Dados

- A mineração de dados surgiu como área de pesquisa e aplicação independente em meados da década de 1990;
- Origens na matemática, estatística e computação são muito anteriores a esse período;
- A área também ganhou evidência nos últimos anos depois de ser cunhado o termo Big Data.

Mineração de Dados

As coisas estão mudando [...] Na última década, a maior parte do trabalho que fazemos [...] migrou para o computador, agora ligado a uma rede. Estamos ligados a um colega equipado com uma memória fenomenal, um senso estranho de tempo e nenhuma lealdade [...] corremos o risco de nos tornarmos servos e escravos da informação que produzimos.

BAKER, S. The Numeratti, 2009, p. 18.

Mineração de Dados

- Usuários da internet passou de 16 milhões de pessoas em 1995 para aproximadamente ***2,8 bilhões*** em 2013;
- Quantidade de artigos publicados apenas em inglês na Wikipédia passou de 500 mil em 2005 para quase ***4,4 milhões*** em 2013;
- O rádio atingiu uma audiência de 50 milhões de pessoas em 38 anos, a TV 13 anos e a ***internet apenas quatro*** anos para alcançar esse mesmo número de pessoas;
- Buscas diárias no Google ultrapassa ***cinco bilhões***, são escritos 500 milhões de tuítes por dia e vistas ***200 milhões*** de horas de vídeos no YouTube diariamente.
- Ainda no YouTube, foram enviadas ***13 milhões de horas*** de vídeo apenas no ano 2010, o que corresponde a aproximadamente oito anos de conteúdo enviados todos os dias.

O que é Mineração de Dados?

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

Mina => Bases de dados

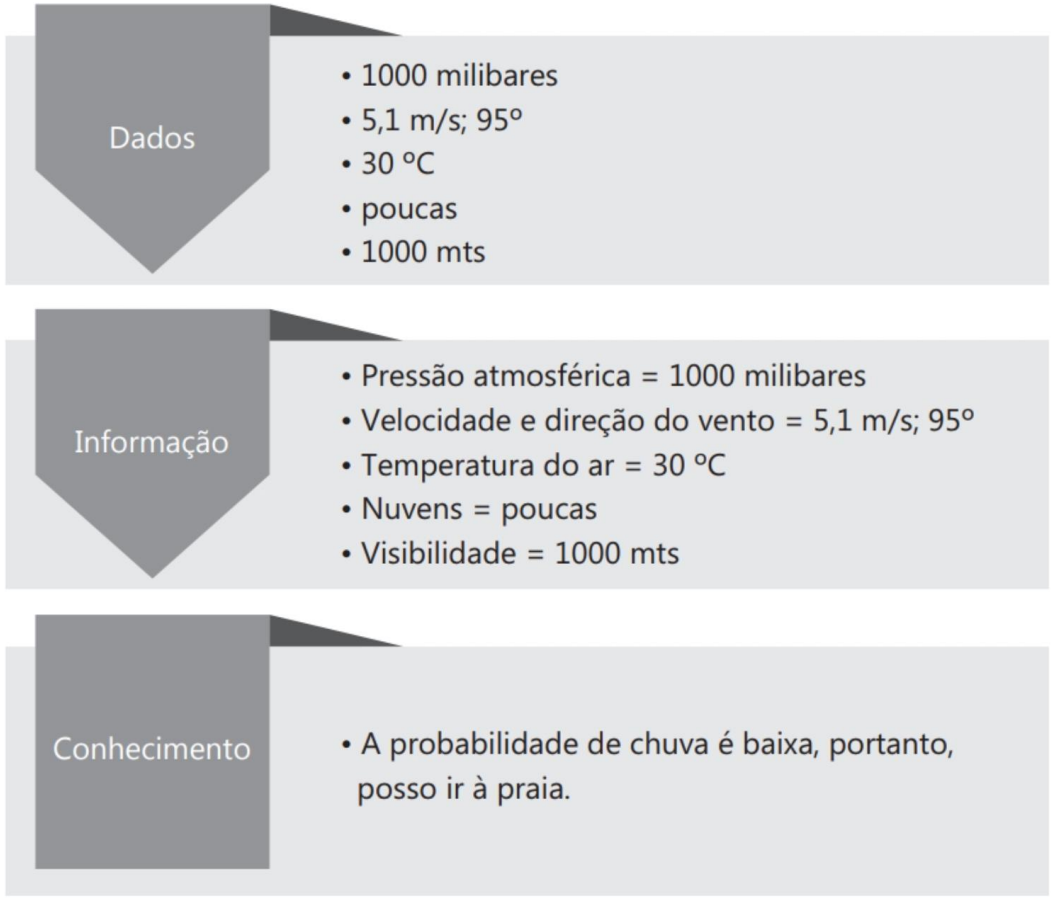
Ferramentas => Algoritmos e hardware

Minerais preciosos => Conhecimento

O que é Mineração de Dados?

- **Dados** são símbolos ou signos não estruturados, sem significado, como valores em uma tabela;
- **Informação** está contida nas descrições, agregando significado e utilidade aos dados, como o valor da temperatura do ar;
- **Conhecimento** é algo que permite uma tomada de decisão para a agregação de valor, então, por exemplo, saber, que vai chover no fim de semana pode influenciar sua decisão de viajar ou não para a praia.

O que é Mineração de Dados?



Fonte: FERRARI, DANIEL GOMES, and LEANDRO NUNES DE CASTRO SILVA. *Introdução a mineração de dados*. Saraiva Educação SA, 2017.

O que é Mineração de Dados?

- 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);

O que é Mineração de Dados?

Etapas KDD: 1. Seleção de bases de dados; 2. Pré-processamento; 3. Transformação de dados; 4. Mineração de dados; 5. Interpretação/Avaliação.

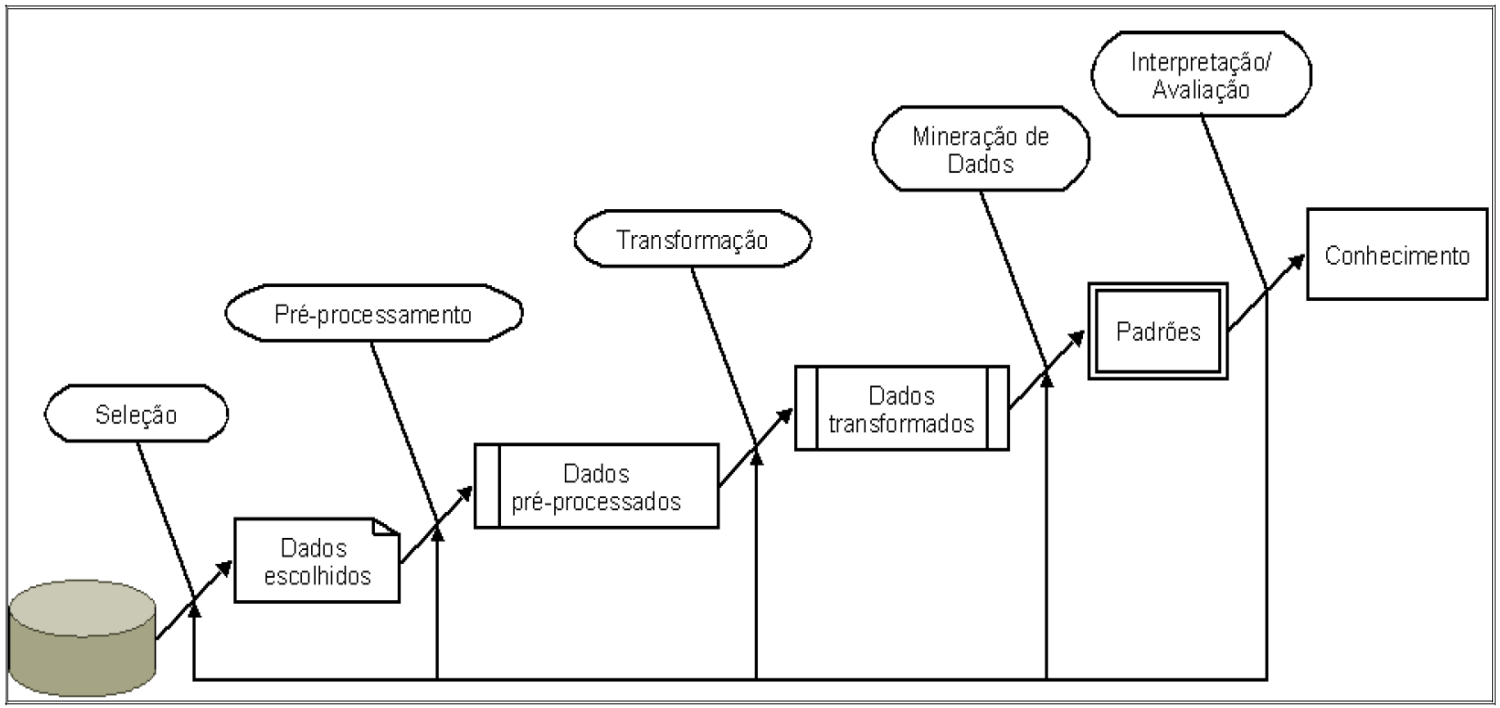
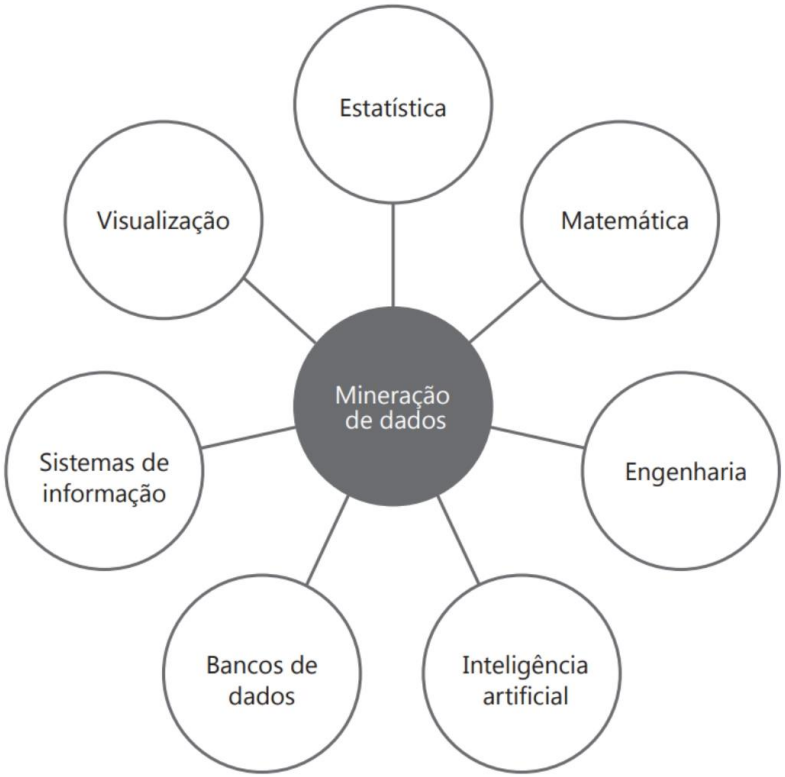


Figura 1: Processo de descoberta de conhecimento em bases de dados. (Fonte: Fayyad et al., 1996)

O que é Mineração de Dados?

Figura 1.4 Multidisciplinaridade da mineração de dados



Fonte: FERRARI, DANIEL GOMES, and LEANDRO NUNES DE CASTRO SILVA. *Introdução a mineração de dados*. Saraiva Educação SA, 2017.

Principais Tarefas Mineração de Dados

- Em geral, essas tarefas podem ser classificadas em duas categorias:

1. **Descritivas:** caracterizam as propriedades gerais dos dados;

2. **Preditivas:** fazem inferência a partir dos dados objetivando predições.

Análise descritiva

Análise Descritiva de Dados:

- Não requer elevado nível de sofisticação;
- Descreve características intrínsecas aos dados;
- Permite investigar a distribuição de frequência, as medidas de centro e variação, e as medidas de posição relativa e associação dos dados;
- Técnicas elementares de visualização também são empregadas para um melhor entendimento da ***natureza*** e ***distribuição*** dos dados.

Análise descritiva

Análise Descritiva de Dados:

- As análises descritivas permitem uma ***sumarização*** e ***compreensão*** dos objetos da base e seus atributos;
- Como qual o ***salário médio*** dos professores universitários brasileiros ou qual a ***distribuição*** salarial desses professores;
- É possível saber, qual a posição relativa de um salário quando comparada à distribuição de salários disponível, o que permite identificar, por sua vez, se um salário está ***abaixo ou acima*** da média.

Predição

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.

Predição

Predição: Classificação e 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***.

Predição

Predição: Classificação e estimação

Considere o problema de ***atribuição de crédito***, no qual um cliente se dirige a um banco com o objetivo de conseguir um financiamento para trocar seu veículo.

1. A primeira pergunta a ser respondida corresponde a uma tarefa de ***classificação***: o crédito será oferecido ou não?
2. Em seguida, há outra pergunta que pode ser relevante responder: qual o valor do crédito a ser oferecido?

Predição

Predição: Classificação e estimação

Qual é o valor do crédito?

É uma estimação (*regressão*) 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)

Predição

Predição: Classificação e estimação

- 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).

Predição

Predição: Classificação e estimação

- **Tarefas de classificação:** identificação de spams, classificação de objetos, atribuição de crédito e detecção de fraudes.
- **Tarefas de estimação:** 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.

Predição

<u>Exercício 1:</u> Responda para cada pergunta se é uma tarefa de classificação ou estimação (regressão):

- Definir o ***valor de um imóvel*** baseado na localização e tamanho?
- Definir se um paciente ***possui determinada doença*** baseado em exame?

Análise de grupos

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.

Análise de grupos

Agrupamento: Clustering

- 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

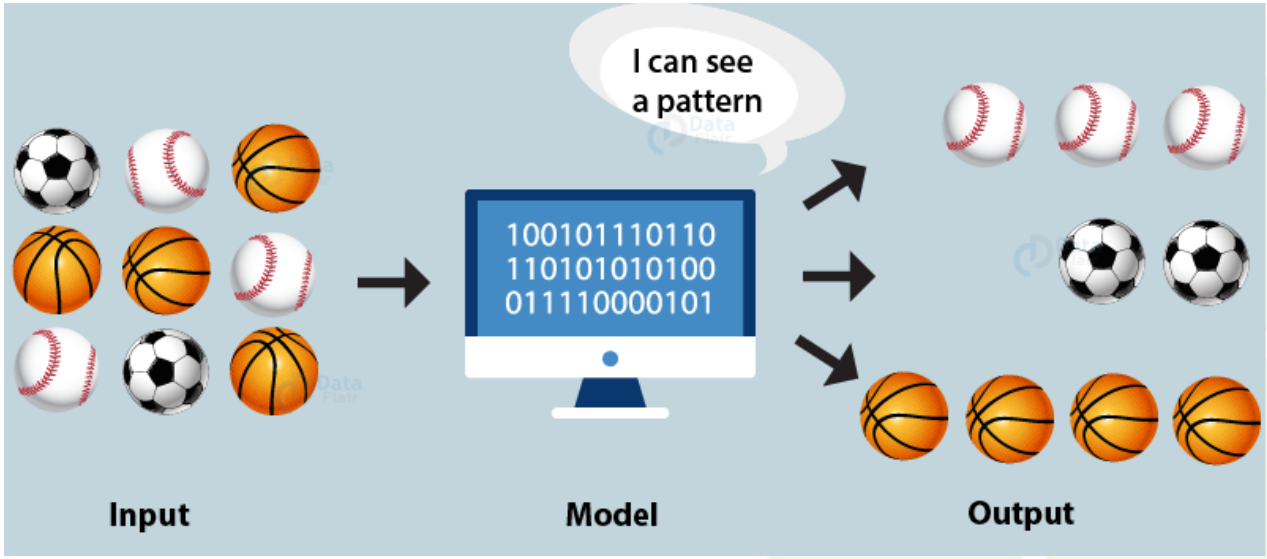
Agrupamento: Clustering

- Considere o problema de segmentar uma base de dados descrevendo ***bolas***, na qual cada bola está descrita por um conjunto de atributos, como ***tamanho***, ***cor*** e ***textura***.
- Suponha que haja bolas de ***futebol, basquete e beisebol*** nessa base de dados e que o algoritmo precisa segmentá-las sem ter conhecimento algum sobre a classe da bola, recebendo apenas informações dos atributos.

Análise de grupos

Agrupamento: Clustering

- Como a ***tamanho, cor e textura*** das bolas de futebol, basquete e beisebol substancialmente diferentes, durante o agrupamento o algoritmo deverá, naturalmente, colocar cada bola em um grupo.



Fonte: https://data-flair.training/blogs/clustering-in-machine-learning/

Associação

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

Fonte: https://data-flair.training/blogs/clustering-in-machine-learning/

Associação

- 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.

Associação

- ***Pense no seguinte:*** os gerentes de marketing gostam muito de frases como “90% dos clientes que compram um ***smartphone*** assinam um ***plano de dados*** para seu aparelho”;

- A ***regra encontrada*** pela ferramenta de análise de dados e que está refletida nessa afirmação é aquela que associa smartphone ao plano de dados.
- Regras dessa natureza são chamadas de ***regras de associação***.

Associação

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.

Mais formalmente, regras de associação possuem a forma $X \longrightarrow Y$:

$$A_1 \text{ e } A_2 \text{ e } \ldots \text{ e } A_m \longrightarrow B_1 \text{ e } B_2 \text{ e } \ldots \text{ e } B_n,$$

onde A_i, $i = 1, \ldots, m$, e B_j, $j = 1, \ldots, n$, são pares de valores de atributos.

As regras de associação X ⟶ Y são interpretadas da seguinte forma: registros da base de dados que satisfazem à condição em X também satisfazem à condição em Y. No caso do exemplo apresentado no início desta seção, ***X = smartphone*** e ***Y = plano de dados***.

REDE FEDERAL DE EDUCAÇÃO PROFISSIONAL E TECNOLÓGICA

Associação

- Outro aspecto importante mencionado no exemplo é que "***smartphone → plano de dados***" em ***90% dos casos***,
- Confiança de 90% de que um cliente que comprar 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.
- A confiança é uma medida de significância da regra.

Detecção de Anomalias

- 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.

Detecção de 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.

Próxima aula

- **Pré-processamento de dados**
  - Introdução
  - Limpeza de dados
  - Integração de dados
  - Redução de dados
  - Transformação de dados
  - Discretização
  - Exemplos

Até a próxima aula!

REDE FEDERAL DE EDUCAÇÃO PROFISSIONAL E TECNOLÓGICA