CADERNO DE EXERCÍCIOS – PARTE 01

1 CAPÍTULO 01: INTRODUÇÃO

1.1 EXERCÍCIOS CONCEITUAIS

1.1.1 Qual a diferença entre as terminologias banco de dados e base de dados?

Resposta: Banco de dados é o repositório onde os dados são armazenados e gerenciados por sistemas de gerenciamento de banco de dados (SGDB) como Oracle, MS SQL Server, etc. Base de dados são coleções organizadas de dados para permitir uma recuperação eficiente dos dados, gerenciados pelos SGDB para serem lidos e manípulados.

1.1.2 A Fig. 1.4 ilustra a multidisciplinaridade da mineração de dados. Discuta como cada uma das áreas citadas contribui para a mineração de dados.

*Resposta:*A **matemática** é importante para entender conceitos utilizados na manipulação dos dados através de algorítimos, como manipulação de vetores e matrizes, conceitos de grafos, algebra línear e operações com conjuntos.
A **estatística** é utilizado como base de diversos modelos utilizados na mineração de dados e algumas técnicas como análise de dispersão (Quartiles, Variância) e de medida central (média, mediana, moda e faixa de valores) combinadas com gráficos (Histogramas, Frequência, Barra, BoxPlot, Dispersão) são usadas para a exploração dos dados. Esses gráficos fazem parte da **visualização** onde sua importância é dar uma forma de análise para entendimento do estudo. A visualização pode ser feita tanto como forma gráfica como por tabelas estruturadas.
Os **bancos de dados** são muito importantes para a mineração de dados pois são responsáveis por armazenarem as bases de dados que serão trabalhadas no processo.
A **inteligência artificial** é responsável pelas técnicas de análises que serão utilizados para extrair valores das bases de dados, podem ser através de algorítimos de *machine learning* ou *deep learning* por exemplo.
Toda parte de coleta, armazenamento e distribuição dos dados são de responsabilidades da **engenharia de dados**. São etapas muito importântes para a mineração de dados para preparar a base de dados para as etapas seguintes serem executadas.

1.1.3 Discuta as principais similaridades e diferenças entre a análise de grupos e a classificação de dados.

Resposta: A análise de grupo e classificação de dados são utilizados para identificar os objetos em uma base de dados através de suas similaridades. A diferença é que na classificação os rótulos dos dados de treinamentos são conhecidos, enquanto que na análise de grupos, no processo para agrupar objetos similares, os rótulos não são conhecidos.

1.1.4 Sabendo que a tarefa de mineração de regras de associação busca encontrar relações de coocorrência entre os atributos da base, cite cinco exemplos de problemas práticos para os quais essa tarefa é útil.

Resposta: Análise de carrinho de supermercados onde são buscados relações entre os produtos comprados em conjunto. Muitas vezes utilizados para melhorar a disposição de itens em gondolas próximo aos produtos comprados em conjunto. Exemplo: colocar salgadinhos próximo as cervejas.
Recomendações por associação que examinam os padrões de consumo de clientes em *e-commerce* para determinar produtos que são adquiridos em conjuntos e ofertados a novos clientes. Exemplo. *75% dos clientes que compraram esse livro costumam comprar este outro livro também.*
Outro tipo de recomendação é com base no conteúdo de determinados itens como autor, compositor, editor, gênero, etc. Como exemplo da associação de conteúdo são os aplicativos de músicas que fazem recomendação pelo gênero musical que você mais ouve. Quem gosta dessa banda de rock, pode gostar desta outra por também ser de rock.
Essa prática também é utilizada para recomendações de filmes no Netflix, onde são analisados filmes que você já assistiu para lhe recomendar filmes parecidos ou pelo Linkedin que oferece vagas de acordo com seu perfil ou

buscas realizadas.
Sequência de ações também é uma forma de associação. Nesse caso, a sequência de um usuário é mapeado e analisado para fazer propagandas por exemplo. Isso acontece com frequência se você fizer uma busca no google sobre determinado assunto e na sequência aparece um banner de algum produto ou serviço associado ao tema buscado. Mesmo que sua busca não tenha sido para alguma tipo de consumo.

1.1.5 Discuta como o número de objetos e o número e tipos de atributos podem influenciar o processo de mineração de dados.

Resposta: O número de objetos pode influenciar diretamente o resultado esperado se a quantidade de dados for muito pequena. A análise dos algorítimos e amostragem podem dar resultados imprecisos se a quantidade examinada não forem o suficiente para conseguir um resultado confiável.
É muito importante ter uma análise preliminar da base para conhecer bem os atributos que à compõe. Saber com que tipo de dados a base esta composta, irá determinar as técnicas que a mineração de dados irá utilizar. Para isso, técnicas de análise descritivas como medidas de tendência central, análise de componentes principais e outros métodos estatísticos podem ser aplicados para conhece-la melhor antes de iniciar o processo de mineração de dados.

1.1.6 Considerando as muitas nomenclaturas já existentes na literatura, tais como Inteligência Artificial, Aprendizagem de Máquina e Inteligência Computacional, como você justificaria a necessidade de criação da nomenclatura Mineração de Dados?

Resposta: Ter uma nomenclatura como mineração de dados é importante para destacar que existe um processo com etapas definidas para exploração de grandes quantidades de dados com o objetivo de obter informações. Desta forma, o termo fica bem definido para destacar a sua área de atuação dentre as outras existentes.

1.1.7 Há um tipo de aprendizagem intermediário entre a supervisionada e a não supervisionada, intitulada semi-supervisionada. Explique, em linhas gerais, qual a diferença entre elas.

Resposta: A aprendizagem supervisionada é baseado em um conjunto de objetos para os quais as saídas desejadas são conhecidas, enquanto que a não supervisionada é baseado apenas nos objetos da base, cujos rótulos são desconhecidos.
Já o semi-supervisionada é uma junção da duas anteriores, onde são usados dados rotulados e não rotulados.

1.1.8 Explique a diferença entre, apresente e explique um exemplo de base de dados (problema) em que ambas as análises de cada alínea (a e b) abaixo podem ser feitas:

a Agrupamento e classificação.
Resposta: Agrupamento é o nome dado ao processo de separar, particionar ou segmentar um conjunto de objetos em grupos de objetos similares. A classificação é o agrupamento de dados que considera dados de entrada não rotulados. Como exemplo, 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.

b Classificação e estimação.
Resposta: A classificação é usada para predizer valores discretos, ao passo que a estimação é usada para predizer valores contínuos.
Por exemplo: Em um exemplo de atribuição de crédito, saber se o crédito será oferecido ou não faz parte da classificação. Qual o valor do crédito é a estimação.

2 CAPÍTULO 02: PRÉ-PROCESSAMENTO DE DADOS

2.1 EXERCÍCIOS CONCEITUAIS

2.1.1 Classifique os dados abaixo em estruturados, semiestruturados e não estruturados:

a Tabela com os dados de cadastro dos funcionários de uma empresa.
Resposta: Estruturados.

b Arquivos de Som.
Resposta: Não estruturados.

c Apresentações em PowerPoint.
Resposta: Não estruturados.

d Textos com palavras-chave identificadas.
Resposta: Semi estruturados.

e Imagens com tags.
Resposta: Semi estruturados

2.1.2 Os três principais tipos de problemas com os dados são incompletude, inconsistência e ruído, e as principais tarefas de pré-processamento são limpeza, integração, redução, transformação e discretização. Explique quais tarefas estão associadas a quais problemas e como se dá tal associação.

Resposta:

- **Limpeza:** Essa etapa trata incompletude fazendo imputação de valores ausentes, ruídos fazendo suas remoções e corrigindo inconsistências.
- **Integração:** É utilizado para unir dados de múltiplas fontes, facilitando a análise dos dados buscando as informações em um único local.
- **Redução:** Utilizado para reduzir a dimensão da base de dados, agrupando ou eliminando atributos redundantes, ou reduzindo a quantidade de objetos da base.
- **Transformação:** Faz a padronização dos dados para facilitar a execução das técnicas de mineração de dados.
- **Discretização:** Normaliza os atributos para poderem ser trabalhados em conjuntos maiores de problemas e reduz a quantidade de valores de atributos contínuos.

2.1.3 Discuta as possíveis implicações de objetos e atributos duplicados em uma base de dados no processo de mineração de dados.

Resposta: A duplicidade de objetos ou atributos podem trazer problemas para análises porque podem causar distorções ou anomalias. Ela pode ser causada pela integração de bases de dados ou por inserções de usuários ou sistemas. O processo de normalização pode prevenir esse problema. Em casos como backup dos dados ou para promover consistências, a duplicidade pode ser positiva.

2.2 EXERCÍCIOS NUMÉRICOS

ID	BI-RADS	Idade	Forma	Contorno	Densidade	Severidade
1	5	67	Lobular	Especulada	Baixa	Maligno
2	4	43	Redonda	Circunscrita	?	Maligno
3	5	58	Irregular	Especulada	Baixa	Maligno
4	4	28	Redonda	Circuscrita	Baixa	Benigno
5	5	74	Redonda	Especulada	?	Maligno
6	4	65	Redonda	?	Baixa	Benigno
7	4	70	?	?	Baixa	Benigno
8	5	42	Redonda	?	Baixa	Benigno
9	5	57	Redonda	Especulada	Baixa	Maligno
10	5	60	?	Especulada	Alta	Maligno

Tabela 2.4

2.2.1 Para a amostra da base de dados Mamo apresentada na Tabela 2.4 considere o atributo severidade como a classe alvo e faça:

a Impute os valores ausentes do atributo forma usando a moda por classe;
Resposta: A moda é o valor que aparece com mais frequência em um conjunto de dados. Levando o atributo severidade como a classe alvo, temos:
$M = \{Maligno, Maligno, Maligno, Benigno, Maligno, Benigno, Benigno, Benigno, Maligno, Maligno\}$

ID	Contorno	Severidade
4	Circunscrita	Benigno
6	?	Benigno
7	?	Benigno
8	?	Benigno
1	Especulada	Maligno
2	Circunscrita	Maligno
3	Especulada	Maligno
5	Especulada	Maligno
9	Especulada	Maligno
10	Especulada	Maligno

$Maligno = 6$
$Benigno = 4$

Observando os atributos Forma e Severidade

Forma	Severidade
Lobular	Maligno
Redonda	**Maligno**
Irregular	Maligno
Redonda	Benigno
Redonda	**Maligno**
Redonda	Benigno
?	Benigno
Redonda	Benigno
Redonda	**Maligno**
?	Maligno

A forma com maior frequência com severidade *maligno* é *Redondo*.

Tabela atualizada:

ID	BI-RADS	Idade	Forma	Contorno	Densidade	Severidade
1	5	67	Lobular	Especulada	Baixa	Maligno
2	4	43	Redonda	Circunscrita	?	Maligno
3	5	58	Irregular	Especulada	Baixa	Maligno
4	4	28	Redonda	Circunscrita	Baixa	Benigno
5	5	74	Redonda	Especulada	?	Maligno
6	4	65	Redonda	?	Baixa	Benigno
7	4	70	Redonda	?	Baixa	Benigno
8	5	42	Redonda	?	Baixa	Benigno
9	5	57	Redonda	Especulada	Baixa	Maligno
10	5	60	Redonda	Especulada	Alta	Maligno

b Impute os valores ausentes do atributo contorno usando o método hot-deck;
Resposta: Utilizando o método de imputar de acordo com a última observação (last observation carried forward), que é um tipo de hot-deck, o valor imputado será o da cédula imediatamente anterior ao valor ausente. A tabela ordenada ficou da seguinte forma.

Nesse caso, o valor será *Circunscrita* para os três valores ausentes.

Tabela atualizada:

ID	BI-RADS	Idade	Forma	Contorno	Densidade	Severidade
1	5	67	Lobular	Especulada	Baixa	Maligno
2	4	43	Redonda	Circunscrita	?	Maligno
3	5	58	Irregular	Especulada	Baixa	Maligno
4	4	28	Redonda	Circunscrita	Baixa	Benigno
5	5	74	Redonda	Especulada	?	Maligno
6	4	65	Redonda	Circunscrita	Baixa	Benigno
7	4	70	Redonda	Circunscrita	Baixa	Benigno
8	5	42	Redonda	Circunscrita	Baixa	Benigno
9	5	57	Redonda	Especulada	Baixa	Maligno
10	5	60	Redonda	Especulada	Alta	Maligno

c Impute os valores ausentes do atributo densidade usando uma constante global.
Resposta: Substituição por constante global constitui em substituir o valor ausente por um valor possível, de acordo com o tipo de atríbuto, em todos os ausentes. Nesse caso foi escolhido a densidade *Baixa* para substituir os valores 2 e 5.

Tabela atualizada:

ID	BI-RADS	Idade	Forma	Contorno	Densidade	Severidade
1	5	67	Lobular	Especulada	Baixa	Maligno
2	4	43	Redonda	Circunscrita	Baixa	Maligno
3	5	58	Irregular	Especulada	Baixa	Maligno
4	4	28	Redonda	Circunscrita	Baixa	Benigno
5	5	74	Redonda	Especulada	Baixa	Maligno
6	4	65	Redonda	Circunscrita	Baixa	Benigno
7	4	70	Redonda	Circunscrita	Baixa	Benigno
8	5	42	Redonda	Circunscrita	Baixa	Benigno
9	5	57	Redonda	Especulada	Baixa	Maligno
10	5	60	Redonda	Especulada	Alta	Maligno

2.2.2 Ainda para a base de dados Mamo, faça a suavização (encaixotamento) em caixas de mesma frequência do atributo idade usando duas caixas. Faça a suavização pela média das caixas e também pelos extremos.

Resposta: Atributo idade ordenado: $I = \{28, 42, 43, 57, 58, 60, 65, 67, 70, 74\}$
CAIXA 1$[28, 73] : 28, 42, 43, 57, 58$
CAIXA 2$[60, 74] : \{60, 65, 67, 70, 74\}$

Suavização pela média:
CAIXA $1 = 71, 71, 71, 71, 71$
CAIXA $2 = 68, 68, 68, 68, 68$

Suavização pelos extremos:
CAIXA $1 = 28, 28, 28, 73, 73$
CAIXA $2 = 60, 60, 60, 60, 74, 74$

2.2.3 Normalize o atributo dia da base de dados da Tabela 2.3 utilizando os seguintes métodos: max-min no intervalo [0,1]; escore-z; escalonamento decimal (j=2); e range inter-quartil. Utilizando a normalização pelo escore-z ou range interquartil é possível identificar algum valor que se distancia excessivamente dos demais? Discuta.

Resposta: $DIA = \{19, 11, 16, 3, 23, 14\}$

1. **MAX-MIN:** $DIA = \{0.8, 0.4, 0.65, 0, 1, 0.55\}$
2. **SCORE-Z:**
   (a) **Média:** $M_A = \frac{19+11+16+3+23+14}{5} = 17.2$
   (b) **Desvio Padrão:** $DP = \sqrt{\frac{(19-17.2)^2+(11-17.2)^2+(16-17.2)^2+(3-17.2)^2+(23-17.2)^2+(14-17.2)^2}{5}}$

$$DP = \sqrt{\frac{288.64}{5}}$$
$$DP = 7.60$$

(c) **Escore-Z:** $a' = \frac{(a-\overline{a})}{\alpha_a}$
$a' = \{0.23, -0.81, -0.16, -1.87, 0.76, -0.42\}$

3. **Escalonamento Decimal ($j = 2$):** $a' = \frac{a}{10^j}$
$a' = \{0.19, 0.11, 0.16, 0.03, 0.23, 0.14\}$

4. **Range interquartil:**

(a) **IQR:** $IQR = Q_3 - Q_1$
$Q_1\{3, \mathbf{11}, 14\} = 11$
$Q_3\{16, \mathbf{19}, 23\} = 19$
$IQR = Q_3 - Q_1$
$IQR = 19 - 11$
$IQR = 8$

(b) **Mediana:** $m = \{3, 11, \mathbf{14, 16}, 19, 23\} = 15$

(c) **Range interquartil:** $a' = \frac{a-m}{IQR}$
$a' = \{-1.5, -0.5, -0.12, 0.12, 0.5, 1\}$

Os valores que mais se distanciam entre as normalizações de score-z e range interquartil, são 3 e 23, sendo o menor e maior valor respectivamente.

2.3 DESAFIO COMPUTACIONAL

2.3.1 O Capítulo 2 apresenta o exemplo do processo de preparação de base de dados aplicado à base Mamo. Para este desafio, realize o mesmo processo para a base Bancos descrita na Seção 2.1.2. Realize as etapas do processo de forma similar ao exemplo do livro.

Resposta:

3 CAPÍTULO 03: ANÁLISE DESCRITIVA DE DADOS

3.1 EXERCÍCIOS CONCEITUAIS

3.1.1 Qual é o objetivo da análise descritiva de dados?

Resposta: A análise descritivas de dados tem por objetivo descrever, simplificar ou sumarizar as principais características de uma base de dados.

3.1.2 Qual é o propósito na utilização de distribuições de frequência?

Resposta: As distribuições de frequências permitem a sumarização de grandes conjuntos de dados, ajudam a entender a natureza desses dados e fornecem uma base para a construção de gráficos.

3.1.3 Qual a diferença entre as medidas de tendência central, as medidas de dispersão e as medidas de forma?

Resposta: As medidas de tendência central, busca um valor central ou um valor típico de um atributo que tenta descrever um conjunto a partir deste ponto.
A medida de dispersão já tenta expressar quantitativamente a dispersão dos dados, medindo se a distribuição está compacta ou alongada.
E a medidas de forma trazem informações sobre o formato da distribuição, medindo a assimetria da função de distribuição de probabilidade de um atributo em torno de sua média.

3.2 EXERCÍCIOS NUMÉRICOS

3.2.1 Calcule a distribuição de frequência e desenhe o histograma simples do seguinte conjunto de valores, L = {6, 7, 1, 9, 8, 2, 6, 4, 6, 4, 5, 2, 3, 1, 10, 7, 10, 2, 10, 8, 6, 5, 3, 8, 3, 1, 8, 7, 8, 7}.

3.2.2 Para a base de dados abaixo, faça:

Atributo A	Atributo B	Atributo C	Atributo D
0,4190	0,1422	0,3474	0,5357
0,3908	0,0251	0,6606	0,0871
0,8161	0,4211	0,3839	0,8021
0,3174	0,1841	0,6273	0,9891
0,8145	0,7258	0,0216	0,0669
0,7891	0,3704	0,9106	0,9394
0,8523	0,8416	0,8006	0,0182
0,5056	0,7342	0,7458	0,6838
0,6357	0,5710	0,8131	0,7837
0,9509	0,1769	0,3833	0,5341
0,4440	0,9574	0,6173	0,8854
0,0600	0,2653	0,5755	0,8990
0,8667	0,9246	0,5301	0,6259
0,6312	0,2238	0,2751	0,1379
0,3551	0,3736	0,2486	0,2178
0,9970	0,0875	0,4516	0,1821
0,2242	0,6401	0,2277	0,0418
0,6525	0,1806	0,8044	0,1069
0,6050	0,0451	0,9861	0,6164
0,3872	0,7232	0,0300	0,9397

a Calcule o domínio, a média e a mediana de cada atributo.
Resposta:

a **Média:** Atributo A = 0,5857
Atributo B = 0,4307
Atributo C = 0,5220
Atributo D = 0,5047

b **Mediana:** Atributo A $= \frac{0.605 + 0.6312}{2} = 0.6181$
Atributo B $= \frac{0.3704 + 0.3736}{2} = 0.372$
Atributo C $= \frac{0.5301 + 0.5755}{2} = 0.5528$
Atributo D $= \frac{0.5357 + 0.6164}{2} = 0.5760$

b Calcule o primeiro, segundo e terceiro quartis para todos os atributos.
Resposta:

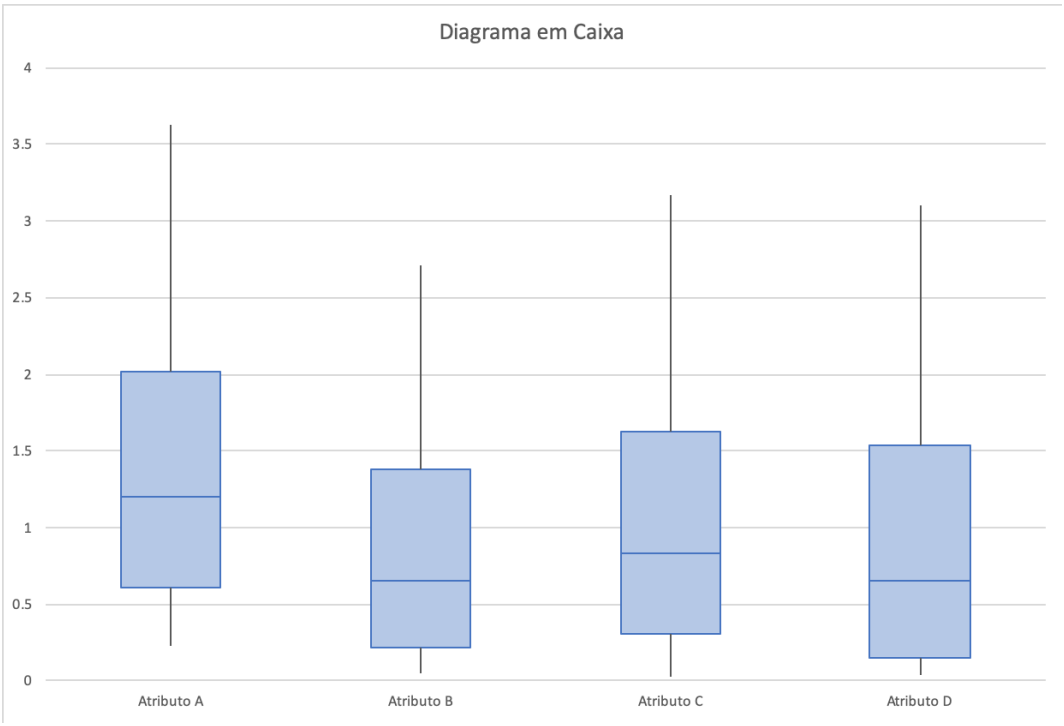
a **Q1**
Atributo A = 0.3872
Atributo B = 0.1769
Atributo C = 0.2751
Atributo D = 0.1069

b **Q2**
Atributo A $= \frac{0.605 + 0.6312}{2} = 0.6181$
Atributo B $= \frac{0.3704 + 0.3736}{2} = 0.372$
Atributo C $= \frac{0.5301 + 0.5755}{2} = 0.5528$
Atributo D $= \frac{0.5357 + 0.6164}{2} = 0.5760$

c **Q3**
Atributo A = 0.8161
Atributo B = 0.7258
Atributo C = 0.8006
Atributo D = 0.8854

c Desenhe o diagrama em caixa para todos os atributos.

Resposta:



3.3 DESAFIO COMPUTACIONAL

3.3.1 O Capítulo 3 traz o exemplo do processo de análise descritiva de dados aplicado aos atributos Mês, Dia, DMC e DC da base Fires. Para este desafio, realize o mesmo processo para os atributos X, Y, Temp e UR da base Fires. Realize as etapas do processo de forma similar ao exemplo exposto no livro.

Resposta: