# Exercício 1 — As cinco etapas do processo KDD (explicação simples)

1. **Seleção (Selection)**
   Escolher quais dados usar — quais tabelas, colunas, período, clientes, etc. É pegar o "conjunto bruto" relevante para o problema.

2. **Pré-processamento / Limpeza (Preprocessing / Cleaning)**
   Tratar valores faltantes, correções, remoção de duplicatas, padronizar formatos (datas, textos), tratar outliers. É a etapa onde a maioria do trabalho sujo acontece.

3. **Transformação (Transformation)**
   Normalizar, agregar, criar variáveis derivadas, pivotar (ex.: transação → matriz binária). Preparar os dados para o algoritmo que vai rodar.

4. **Mineração de Dados (Data Mining)**
   Aplicar algoritmos (classificação, clustering, regras de associação, regressão, etc.) para extrair padrões ou modelos.

5. **Interpretação / Avaliação (Interpretation / Knowledge Evaluation)**
   Avaliar a utilidade dos padrões (ex.: validação, métricas, revisão com especialista) e transformar isso em conhecimento acionável (relatório, regras de negócio, dashboard).

# Exercício 2 — V/F com explicação

1. "A mineração de dados é o processo completo de descoberta de conhecimento." — **F**
   **Por quê:** Mineração de dados é *uma* etapa chave do processo KDD; KDD engloba seleção, pré-processamento, transformação, mineração e interpretação.

2. "O KDD inclui a etapa de pré-processamento, enquanto a mineração de dados foca em extrair padrões." — **V**
   **Por quê:** Correto — KDD é o processo completo e inclui pré-processamento; mineração é extração de padrões.

3. "O termo 'Data Mining' surgiu antes do KDD." — **F** (dependendo da referência histórica pode confundir, mas a afirmação para fins didáticos: **F**)
   **Por quê:** Historicamente, termos surgiram em paralelo; para a disciplina, costuma-se apresentar KDD como framework abrangente e "Data Mining" como termo técnico de extração. (Se a prova exigir precisão histórica, cite fontes; mas a

ideia é que KDD e Data Mining são conceitos relacionados, não idênticos.)

4. “A mineração de dados pode ser usada dentro de um Data Warehouse.” — **V**
   **Por quê:** Um DW é repositório consolidado — ótimo para aplicar mineração sobre dados integrados.

5. “KDD e ETL são processos idênticos.” — **F**
   **Por quê:** ETL (Extract, Transform, Load) é parte do trabalho de alimentação e transformação (foca em mover e transformar dados para um DW). KDD é mais amplo — inclui seleção, limpeza, mineração e interpretação. ETL é uma etapa técnica/operacional; KDD é um processo analítico.

# Exercício 3 —

Para atender às solicitações, foram feitas três alterações no código. Primeiro, o suporte mínimo do algoritmo Apriori foi reduzido de 0.05 para 0.03, permitindo que itens menos frequentes fossem considerados durante a mineração de regras. Em seguida, as regras geradas foram filtradas para manter apenas aquelas com confiança superior a 0.7, garantindo que somente associações fortes fossem analisadas. Por fim, as regras resultantes foram exportadas para um arquivo chamado regras_mercado.csv utilizando o comando `regras.to_csv("regras_mercado.csv", index=False)`.

Após aplicar esses filtros, apenas uma regra permaneceu no conjunto final de resultados: a regra “Monitor → Mouse”. Essa regra apresentou o maior valor de lift dentre as regras filtradas, aproximadamente 1.43. A confiança dessa regra foi igual a 1.0, o que significa que, em todas as transações que continham o produto Monitor, o produto Mouse também estava presente. O lift maior que 1 indica que essa associação não ocorre por acaso; portanto, há uma relação significativa entre esses itens.

No contexto da base simulada, essa regra significa que clientes que compram um Monitor consistentemente também compram um Mouse na mesma transação. Isso sugere um padrão de compra que poderia ser explorado para promoções conjuntas, recomendações ou estratégias de cross-selling.

# Exercício 7 — Comparando regras (resposta)

- Regra A: (Acessou Vídeo → Fez Exercício)
  suporte = 0,6 ; confiança = 0,75 ; lift = 0,93

- Regra B: (Participou Fórum → Fez Exercício)
  suporte = 0,6 ; confiança = 0,93 ; lift = 1,25

a) **Qual regra é mais relevante?** → Regra B. Tem confiança maior (0,93 vs 0,75) e lift > 1 (1,25) indicando associação positiva (participar do fórum aumenta a probabilidade de fazer exercício acima do esperado). Regra A tem lift < 1 (0,93) — embora a confiança seja 0,75, o lift < 1 indica que a ocorrência de "Acessou Vídeo" está associada a uma probabilidade *menor* de "Fez Exercício" do que a média (ou seja, acessou vídeo não traz maior probabilidade que o comportamento médio).

b) **O que significa um lift < 1?** → Significa associação negativa: a presença do antecedente reduz a probabilidade do consequente comparado à probabilidade geral do consequente. Ou seja, antecedente e consequente tendem a ocorrer **menos juntos** do que por acaso.

c) **Exemplo de regra inútil (associação fraca)** → Em um serviço de streaming: (Assistiu Documentário Sobre Plantas → Assistiu Série de Ficção Científica) com suporte 0,02 e confiança 0,4 e lift ≈ 0,9 — baixa confiança e lift ≤1, sem utilidade para recomendação

# Exercício 8 — Jaccard (cálculo manual)

Conjuntos:

- A = {maçã, banana, pão, leite, queijo}
- B = {banana, leite, queijo, arroz, feijão}
- C = {pão, café, leite, queijo, bolacha}

Calculemos interseção e união de cada par:

1. **A ∩ B** = {banana, leite, queijo} → tamanho = 3
   **A** ∪ **B** = {maçã, banana, pão, leite, queijo, arroz, feijão} → tamanho = 7
   J(A,B) = |A∩B| / |A∪B| = 3 / 7 ≈ 0,4286

2. **A ∩ C** = {pão, leite, queijo} → tamanho = 3
   **A** ∪ **C** = {maçã, banana, pão, leite, queijo, café, bolacha} → tamanho = 7
   J(A,C) = 3 / 7 ≈ 0,4286

3. **B ∩ C** = {leite, queijo} → tamanho = 2
   **B** ∪ **C** = {banana, leite, queijo, arroz, feijão, pão, café, bolacha} → tamanho = 8
   J(B,C) = 2 / 8 = 0,25

Tabela resumida:

| Par | Interseção | União | Jaccard |
|---|---|---|---|
| A,B | 3 | 7 | 3/7 ≈ 0,4286 |
| A,C | 3 | 7 | 3/7 ≈ 0,4286 |
| B,C | 2 | 8 | 2/8 = 0,25 |

3. **Par mais parecido:** A e B ou A e C (empatados em ≈0,4286), ambos mais parecidos que B e C.

4. **Significado de 1 e 0:**

- J = 1 → conjuntos idênticos (todas as características/itens iguais).
- J = 0 → conjuntos disjuntos (nenhum item em comum).

.