



Digital
College

FORMAÇÃO EM

DATA ANALYTICS

MÓDULO 2:

**EXTRAÇÃO, TRANSFORMAÇÃO
E CARGA (ETL)**

UNIDADE 1:

**> PROCESSO DE DESCOBERTA
DE CONHECIMENTO >**



Sumário

1. Introdução	04
1.1. Dados, informação, conhecimento	05
1.2. Processo de Descoberta do Conhecimento	06
1.3. Etapas operacionais do processo de descoberta do conhecimento	08
1.3.1. Pré-processamento	09
1.3.2. Descoberta do Conhecimento	10
1.3.3. Pós-processamento	11
2. Fases do Pré-Processamento	12
2.1. Seleção dos dados	14
2.2. Limpeza dos dados	18
2.3. Codificação	21
2.4. Enriquecimento dos dados	24
2.5. Normalização dos dados	25
3. Descoberta do Conhecimento – Análise de Dados	26
3.1. 4 Tipos de análise de dados	27
4. Pós-Processamento	29
4.1. O que é visualização de dados?	30
4.2. O que faz uma boa visualização de dados?	31
4.3. Por que a visualização de dados é importante?	33
4.4. Alfabetização de dados	34
4.5. A história da visualização de dados	35
5. Gestão de Dados	36
5.1. Princípios	37
5.2. Principais Funções	38



Sumário

6. CRISP-DM – Conceitos	40
6.1. As 6 fases do CRISP-DM	41
6.1.1. Compreensão do Negócio	41
6.1.2. Compreensão dos Dados	42
6.1.3. Preparação dos Dados	43
6.1.4. Modelagem	43
6.1.5. Avaliação	43
6.1.6. Desenvolvimento (implantação)	44

1. Introdução

Os constantes avanços na área da Tecnologia da Informação têm viabilizado o armazenamento de grandes e múltiplas bases de dados. Tecnologias como internet, redes sociais, ambientes virtuais de aprendizagem, dispositivos móveis, aplicativos embarcados, leitores de códigos de barras, sensores para coleta de diferentes tipos de dados, memória secundária de maior capacidade de armazenamento de menor custo, sistemas de telecomunicações, mecanismos de autoatendimento e sistemas de informação em geral são alguns exemplos de recursos que têm tornado possíveis a criação e o crescimento de inúmeras bases de dados de naturezas administrativa, científica, comercial, educacional, governamental e social.

O valor dos dados armazenados está tipicamente ligado à capacidade de se extrair conhecimento de mais alto nível a partir deles, ou seja, informação útil que sirva para apoio à tomada de decisão, e/ou para exploração e melhor entendimento do fenômeno gerador dos dados. Os dados podem apresentar padrões ou tendências úteis e interessantes que, se descobertos, possuem potencial, por exemplo, para otimizar procedimentos em uma empresa, ajudar na compreensão dos resultados de um experimento científico, ou auxiliar médicos a interpretar os efeitos de um tratamento, para citar alguns casos.

Diante desse cenário, naturalmente surgem algumas questões do tipo: “O que fazer com todos os dados armazenados?”, “Como utilizar o patrimônio digital em benefício das instituições?”, “Como analisar e utilizar todo o volume de dados disponível?”, “De que maneira informações subjacentes aos dados armazenados podem ser úteis no contexto ao qual pertencem?”.

A análise de grandes quantidades de dados pelo homem é inviável sem o auxílio de ferramentas computacionais apropriadas. Portanto, torna-se imprescindível a disponibilização de ferramentas que auxiliem o homem na tarefa de analisar, interpretar, e relacionar esses dados, para que se possa elaborar e selecionar estratégias de ação em cada domínio de aplicação.

1.1. Dados, informação, conhecimento

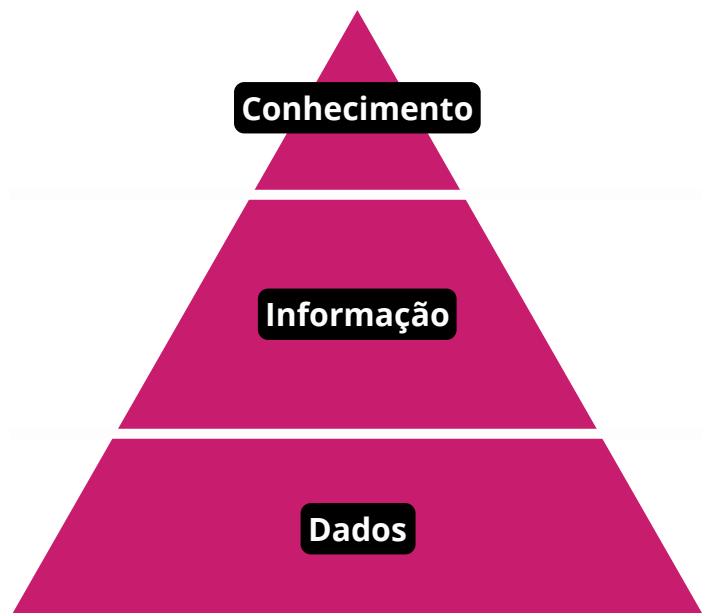


Figura 01

Os **dados**, na base da pirâmide, podem ser interpretados como itens elementares, captados e armazenados por recursos da Tecnologia da Informação. São **cadeias de símbolos** e não possuem **semântica**. Seu propósito é expressar fatos do mundo real de forma a serem tratados no contexto computacional.

As **informações** representam os dados processados, com significados e contextos bem definidos. Diversos recursos da Tecnologia da Informação são utilizados para processar dados e obter informações.

No topo da pirâmide está o conceito de **conhecimento**, que corresponde a um padrão ou conjunto de padrões cuja formulação pode envolver e relacionar dados e informações.

Outros exemplos de conhecimento são: tendências de vendas em uma determinada região, relacionamentos entre o sobe e desce de ações na Bolsa de Valores e certos parâmetros monetários, similaridades entre os comportamentos de compra de clientes de uma empresa etc.

Informação e conhecimento constituem, em geral, a base para se tomar decisões em diversos cenários. Sistemas que auxiliam seus usuários na tomada de decisão diante de situações a eles apresentadas são denominados **sistema de apoio à decisão**. São várias as áreas que podem se beneficiar de sistemas de apoio à decisão. Por exemplo: Medicina, Direito, Engenharia, Meio Ambiente, Educação, dentre muitas outras.

A criação de um sistema de apoio à decisão requer necessariamente a modelagem de informação e conhecimento proveniente do respectivo domínio de aplicação. O processo de aquisição e formalização do conhecimento a ser processado por um sistema de apoio à decisão pode ocorrer junto a diversas **fontes de conhecimento**. Especialistas na área do domínio da aplicação, assim como bases de dados históricas sobre o tema em questão são exemplos de fontes de conhecimento muito utilizadas na construção de sistemas de apoio à decisão.

1.2. Processo de Descoberta do Conhecimento

Para melhor compreensão do processo de descoberta do conhecimento, é necessária uma apresentação dos principais elementos envolvidos em aplicações nesta área. Basicamente, esse processo de descoberta possui três componentes: (a) o problema em que será aplicado; (b) os recursos disponíveis para a solução do problema; (c) os resultados obtidos a partir da aplicação dos recursos disponíveis para a solução do problema.

Definição do problema

O problema a ser submetido ao processo de descoberta pode ser caracterizado por três elementos: o conjunto de dados, o especialista do domínio da aplicação e os objetivos da aplicação.

- **Conjunto de dados**

Como o próprio nome diz, um conjunto de dados corresponde aos dados medidos acerca de determinadas entidades e que serão analisados do processo de descoberta. São exemplos de entidades: clientes, produtos, alunos, compras, eventos, viagens, documentos em um portal de notícias, imagens de eletroencefalograma de pacientes em um hospital, formulários de solicitação de crédito para clientes em uma instituição bancária etc. Todo conjunto de dados pode ser observado sob os aspectos intencional e extensional. O aspecto intencional se refere à estrutura ou ao esquema do conjunto de dados. Neste contexto encontram-se, portanto, as características ou atributos do conjunto de dados. Os casos, instâncias ou registros compõem o aspecto extensional do conjunto de dados.

Em geral, o processo de descoberta pressupõe que os dados sejam organizados em uma única estrutura tabular bidimensional contendo casos e características do problema a ser analisado. É importante destacar que o processo de descoberta de conhecimento não requer que os dados a serem analisados pertençam à Data Warehouses. No entanto, o tratamento e a consolidação dos dados necessários à estruturação e à carga neste tipo de ambiente são extremamente úteis e desejáveis ao processo de descoberta.

- **Especialista do domínio**

O *especialista do domínio da aplicação* representa a pessoa ou o grupo de pessoas que conhece o assunto e o ambiente em que deverá ser realizada a aplicação de descoberta. Em geral, pertencem a esta classe, analistas de negócios interessados em identificar novos conhecimentos que possam ser utilizados em sua área de atuação. Esses profissionais costumam deter o chamado **conhecimento prévio** sobre o problema ("background knowledge"). As informações prestadas pelas pessoas deste grupo são de fundamental importância no processo de descoberta, pois influenciam desde a definição dos objetivos do processo até a avaliação dos resultados.

- **Objetivos da aplicação**

Os *objetivos da aplicação* compreendem basicamente a(s) tarefa(s) de descoberta a ser(serem) realizada(s) e as características esperadas quanto ao modelo de conhecimento a ser produzido pelo processo de descoberta. Tais objetivos retratam, portanto, restrições e expectativas dos especialistas do domínio da aplicação acerca do modelo de conhecimento a ser gerado.

Recursos disponíveis

Entre os recursos disponíveis para solução do problema podem ser destacados o especialista da TI, as ferramentas de descoberta e a plataforma computacional utilizada. Abaixo, seguem comentários sobre cada um deles:

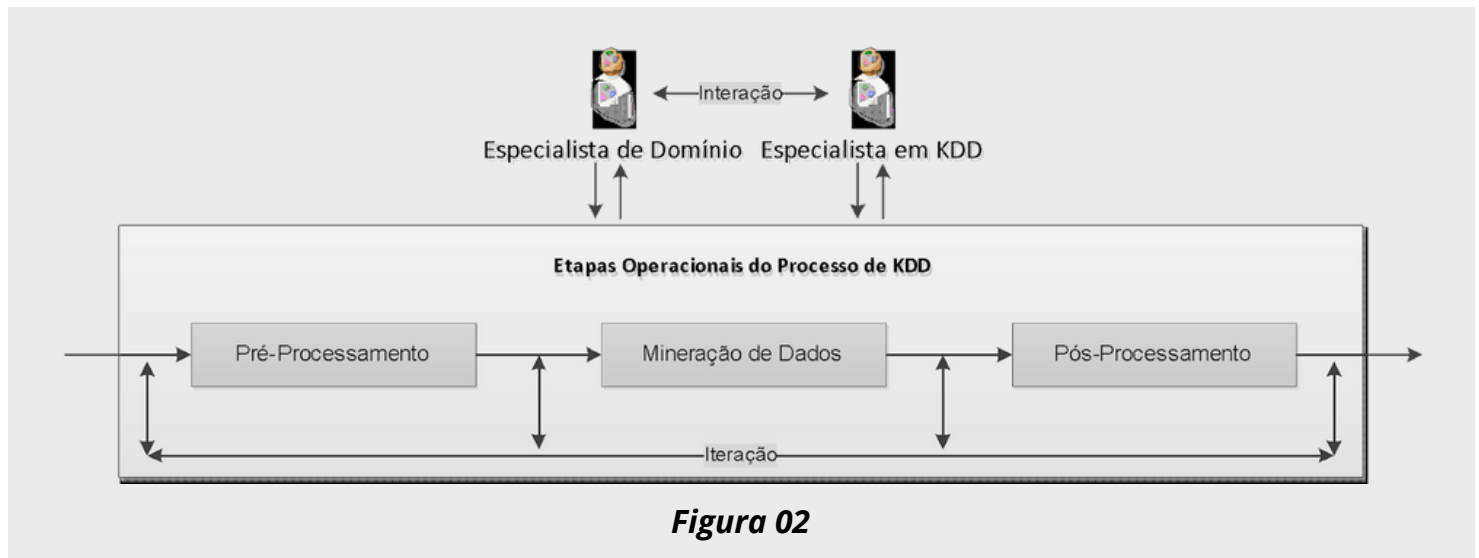
- **O especialista da TI** representa a pessoa ou o grupo de pessoas que possui experiência na execução de processos de descoberta do conhecimento. A pessoa ou grupo que desempenha este papel tem a atribuição de interagir com o especialista do domínio da aplicação e direcionar a condução do processo de descoberta, definindo o que, como e quando deve ser realizada cada ação do processo. Suas atribuições variam desde a identificação e a utilização do conhecimento prévio existente sobre o problema até o direcionamento das ações do processo, que englobam a seleção e a aplicação das ferramentas disponíveis, além da avaliação dos resultados obtidos;
- **A expressão ferramenta de descoberta** está relacionada a qualquer recurso computacional que possa ser utilizado no processo de análise de dados;
- **A plataforma computacional** indica a infraestrutura necessária para a execução de aplicações de sistemas de apoio à decisão.

Resultados obtidos

Os resultados obtidos a partir da aplicação dos recursos disponíveis para solução do problema compreendem, principalmente, os modelos de conhecimento descobertos ao longo do processo de descoberta e o histórico das ações realizadas.

1.3. Etapas operacionais do processo de descoberta do conhecimento

A descoberta do conhecimento é um processo de várias etapas, não trivial, interativo e iterativo, para identificação de padrões compreensíveis, válidos, novos e potencialmente úteis a partir de bases de dados.





1.3.1: Pré-processamento:

Esta etapa compreende todas as funções relacionadas à captação, à organização e ao tratamento dos dados. O objetivo dessa etapa é deixar os dados com o máximo de qualidade possível.

- **Seleção de Dados**

Esta função, também denominada Redução dos Dados, compreende, em essência, a identificação do subconjunto das bases de dados existentes, que deve ser efetivamente considerado durante o processo de descoberta do conhecimento.

A seleção dos dados pode ter dois fatores distintos:

- A seleção de atributos;
- A seleção de registros.

Na seleção de atributos, vamos escolher quais campos/colunas do nosso conjunto de dados será relevante para análise. Por exemplo, num conjunto de dados de alunos, para uma determinada análise, podemos escolher a data de nascimento e o sexo; já para outra análise, podemos escolher o endereço de residência.

Na seleção de registros, vamos escolher quais dados serão relevantes para uma determinada análise. Por exemplo: num conjunto de dados de alunos, só queremos alunos maiores de 18 anos ou alunos que moram no bairro X.

- **Limpeza de Dados**

Esta atividade abrange qualquer tratamento realizado sobre os dados selecionados de forma a assegurar a qualidade (completude, veracidade e integridade) dos fatos representados. Informações ausentes, errôneas ou inconsistentes nas bases de dados devem ser corrigidas de forma a não comprometer a qualidade dos modelos de conhecimento a serem extraídos ao final do processo de descoberta do conhecimento.

Como exemplo podemos ter algum atributo faltante no nosso conjunto de dados, nesse contexto temos dois caminhos a seguir: simplesmente retirar os registros faltantes ou, por meio de inúmeras técnicas disponíveis, completar esses atributos.

- **Codificação de Dados**

Nesta função, os dados devem ser codificados para que possam ser usados como entrada dos algoritmos para análise de dados. A codificação pode ser: numérica-categórica, que transforma valores reais em categorias, ou categórica-numérica, que representa numericamente valores de atributos categóricos.

- **Enriquecimento dos Dados**

A função de enriquecimento dos dados consiste em conseguir de alguma maneira, mais informações que possam ser agregadas aos registros existentes, tornando-os “ricos” para o processo de descoberta do conhecimento. Como exemplo podemos citar o cálculo da idade a partir de uma data de nascimento.



1.3.2. Descoberta do conhecimento

A etapa de descoberta do conhecimento compreende a busca efetiva por conhecimentos úteis no contexto da aplicação. É a principal etapa no processo com um todo. Esta etapa envolve a aplicação de técnicas sobre os dados em busca de conhecimento implícito e útil.

Nesta etapa são definidos as técnicas e os algoritmos a serem utilizados no problema em questão. A escolha da técnica depende muitas vezes do tipo de descoberta a ser realizada. Como exemplo de técnicas, podemos citar:

- Descoberta de Associações;
- Classificação;
- Regressão;
- Sumarização;
- Detecção de Desvios;
- Descobertas de Sequências;

1.3.3. Pós-processamento

Essa etapa abrange o tratamento do conhecimento obtido na Descoberta do Conhecimento. Tal tratamento tem como objetivo facilitar a interpretação e a avaliação por especialistas do domínio da aplicação, no que se refere à utilidade do conhecimento descoberto.

Entre as principais funções da etapa de pós-processamento estão:

- Elaboração e organização do conhecimento obtido;
- Elaboração de gráficos, diagramas ou relatórios.

2. Fases do pré-processamento

Nessa seção expomos os dados que serão trabalhados ao longo deste módulo. Na tabela 1, temos a descrição da estrutura de dados simplificada dos convênios do Estado do Ceará. Para mais detalhes, consultar o site do Ceará Transparente, de onde vieram os dados desta seção.

ATRIBUTO	TIPO DE DADO	DESCRIÇÃO DO DOMÍNIO
exercicio	integer	Ano do Convênio
municipio	varchar	Nome do Município
orgao	varchar	Nome do órgão
numero	varchar	Número identificador
situacao	varchar	Situação do Convênio
prestacao_contas	varchar	Descrição da Prestação de Contas
valor	numeric(18,2)	Valor do Convênio
dataassinatura	date	Data que foi assinado
data_inicio	date	Data de início
data_termino	date	Data de término

Tabela 1 – Estrutura de Dados de exemplo sobre convênios do Estado do Ceará

A tabela 1 acima é o que podemos chamar de dicionário de dados. Nela temos os nomes dos atributos, seu tipo de dados e uma descrição sobre o que seria o atributo. Uma das primeiras tarefas do Analista de Dados é a compreensão dos dados: deve-se descobrir a origem do dado e analisar a melhor forma de extraí-lo.

quanto à representação de seus valores (tipo de dado) e quanto à natureza da informação (tipo de variável). Os tipos de dados indicam a forma em que os dados estão armazenados. Os tipos de variáveis expressam a natureza com que a informação deve ser interpretada. Assim, as variáveis se classificam em:

Os atributos descritos acima podem ser classificados sobre dois aspectos:

a) Nominais ou Categóricas – são variáveis utilizadas para nomear ou atribuir rótulos a objetos. Podem assumir valores pertencentes a um conjunto finito e geralmente pequeno;

b) Discretas – assemelham-se às variáveis nominais, mas os valores que elas podem assumir possuem um ordenamento, e este possui algum significado;

c) Contínuas – são variáveis quantitativas cujos valores possuem uma relação de ordem entre si, assim como as variáveis discretas. O conjunto de valores de uma variável contínua pode ser finito ou infinito.

ATRIBUTO	TIPO DE VARIÁVEL	OBSERVAÇÕES
exercicio	Discreta	Envolve uma relação de ordem entre seus valores
municipio	Nominal	Não possui relação de ordem entre seus valores
orgao	varchar	Não possui relação de ordem entre seus valores
numero	Nominal	Não possui relação de ordem entre seus valores
situacao	Nominal	Não possui relação de ordem entre seus valores
prestacao_contas	Nominal	Não possui relação de ordem entre seus valores
valor	Contínua	Conjunto de valores teoricamente infinito
data_assinatura	Discreta	Envolve uma relação de ordem entre seus valores
data_inicio	Discreta	Envolve uma relação de ordem entre seus valores
data_termino	Discreta	Envolve uma relação de ordem entre seus valores

Tabela 2 – Classificação dos atributos.



2.1: Seleção dos dados

Esta função compreende, em essência, a identificação de quais informações, dentre as bases de dados existentes, devem ser efetivamente consideradas durante o processo de descoberta.

Em geral, os dados que serão objetos de análise se encontram organizados em bases de dados transacionais que sofrem constantes atualizações ao longo do tempo. Assim, recomenda-se que seja sempre feita uma cópia dos dados a fim de que o processo de descoberta não interfira nas rotinas operacionais eventualmente relacionadas com a base de dados. Nos casos em que já existe uma estrutura de Data Warehouse, deve-se verificar a possibilidade de que esta seja utilizada no processo de descoberta. Nos demais casos, é comum a congregação dos dados relevantes em um único conjunto. Tal fato justifica-se porque a maioria dos métodos de descoberta e análise de dados pressupõe que os dados estejam organizados em uma única, possivelmente muito grande, estrutura tabular bidimensional. Perceba, portanto, que o processo de descoberta pode ocorrer independente da disponibilidade ou não do Data Warehouse.

Por simplicidade, consideramos que o conjunto de dados a ser pré-processado esteja armazenado na forma de tabela, possivelmente em um Banco de dados relacional, viabilizando, assim, a execução dos comandos ANSI SQL apresentados em exemplos.

A junção de dados relevantes em um único conjunto de dados pode ocorrer de duas formas:

- Junção Direta;
- Junção Orientada.

Convém mencionar que há situações em que, no início do processo, o especialista em KDD já recebe o conjunto de dados a ser analisado. Em nosso exemplo, estamos partindo deste tipo de situação, em que o conjunto de clientes foi fornecido sem uma análise da origem dos dados.

Assim, considerando que os dados estejam reunidos em uma mesma estrutura tabular, a função de seleção dos dados pode ter dois enfoques distintos: a escolha de atributos ou a escolha de registros a serem considerados no processo de descoberta.

Redução de dados horizontal

A seleção por redução de dados horizontal é caracterizada pela escolha de casos. Entre as operações de redução de dados horizontal, temos:

- Segmentação do conjunto de dados – nesta operação, deve-se escolher um ou mais atributos para nortear o processo de segmentação. Como exemplo podemos citar a escolha de analisar apenas os convênios de 2021. Para tal operação podemos usar o comando abaixo.

SELECT * FROM CONVENIOS WHERE EXERCICIO = 2021

Exercício	Município	Órgão	Número	Situação	Valor
2021	MUNICIPIO DE IPU	SECRETARIA DA SAUDE	1196102	Execução Normal - Liberação de Recursos Suspensa	R\$ 139.576,67
2021	MUNICIPIO DE IPU	SECRETARIA DA EDUCACAO	1194787	Execução Normal - Liberação de Recursos Suspensa	R\$ 470.137,50
2021	MUNICIPIO DE IPU	SECRETARIA DA EDUCACAO	1157499	CONCLUÍDO	R\$ 639.509,63
2021	MUNICÍPIO DE UMIRIM	SECRETARIA DA EDUCACAO	1194785	Execução Normal - Liberação de Recursos Suspensa	R\$ 413.400,00
2021	MUNICÍPIO DE UMIRIM	SECRETARIA DA EDUCACAO	1157045	VENCIDO	R\$ 217.601,21

Figura 03

- Eliminação direta dos casos – esta operação pode ser interpretada como uma variação da anterior, em que são especificados os casos a eliminar, e não os casos que devem permanecer. Como exemplo podemos citar a escolha por analisar todos os exercícios, exceto o de 2021. Para tal operação podemos usar o comando abaixo.

SELECT * FROM CONVENIOS WHERE EXERCICIO = 2021

Exercício	Município	Órgão	Número	Situação	Valor
2022	MUNICIPIO DE IPU	SECRETARIA DA EDUCACAO	1199541	EM EXECUÇÃO - NORMAL	R\$ 778.602,98
2022	MUNICIPIO DE IPU	SUPERINTENDÊNCIA DE OBRAS PÚBLICAS	1211654	EM EXECUÇÃO - NORMAL	R\$ 3.724.699,61
2020	MUNICIPIO DE IPU	SECRETARIA DA INFRA ESTRUTURA	1131494	Execução Normal - Liberação de Recursos Suspensa	R\$ 1.238.097,20
2020	MUNICIPIO DE IPU	SECRETARIA DA SAUDE	1131393	Execução Normal - Liberação de Recursos Suspensa	R\$ 424.519,61
2020	MUNICIPIO DE IPU	SECRETARIA DAS CIDADES	1128375	Execução Normal - Liberação de Recursos Suspensa	R\$ 4.061.021,77

Figura 04

- Amostragem aleatória – consiste em sortear, do conjunto de dados, um número preestabelecido de registros de forma que o conjunto resultante possua menos registros do que o original, existem várias técnicas para realizar essa operação:
 - Amostragem aleatória simples sem reposição;
 - Amostragem aleatória simples com reposição;
 - Amostragem de clusters;
 - Amostragem estratificada.
- Agregação de informações – esta operação consiste em reunir (agregar) alguns registros de forma a produzir um conjunto de dados de tamanho menor que o original. Os dados são consolidados em um nível menor de detalhamento.

Redução de dados vertical

A seleção de dados por redução de dados vertical, também denominada redução de dimensionalidade, é uma operação de pré-processamento importante no processo de descoberta, posto que visa a escolher o subconjunto de atributos mais relevante para alcançar o objetivo traçado na definição do problema.

Entre as principais motivações para a aplicação da redução de dados vertical temos:

- Um conjunto de atributos bem selecionado pode conduzir a modelos de conhecimento mais concisos e com melhor desempenho;
- A eliminação de um atributo é muito mais significativa em termos de redução do tamanho de um conjunto de dados do que a exclusão de um registro.

A eliminação direta de atributos se refere à eliminação de atributos cujo conteúdo não seja relacionado ao processo de descoberta. Na imagem abaixo, retiramos o atributo “Número”, pois este serve como identificador no sistema de origem, contudo, para a análise de dados ele tem pouca relevância.

Exercício	Município	Órgão	Situação	Valor
2022	MUNICIPIO DE IPU	SECRETARIA DA EDUCACAO	EM EXECUÇÃO - NORMAL	R\$ 778.602,98
2022	MUNICIPIO DE IPU	SUPERINTENDÊNCIA DE OBRAS PÚBLICAS	EM EXECUÇÃO - NORMAL	R\$ 3.724.699,61
2020	MUNICIPIO DE IPU	SECRETARIA DA INFRA ESTRUTURA	Execução Normal - Liberação de Recursos Suspensa	R\$ 1.238.097,20
2020	MUNICIPIO DE IPU	SECRETARIA DA SAUDE	Execução Normal - Liberação de Recursos Suspensa	R\$ 424.519,61
2020	MUNICIPIO DE IPU	SECRETARIA DAS CIDADES	Execução Normal - Liberação de Recursos Suspensa	R\$ 4.061.021,77

Figura 05

Outra heurística que pode ser aplicada é a eliminação de todos os atributos que apresentam valores constantes em todos os registros do conjunto de dados.

Redução de valores

A operação de redução de valores é uma alternativa à redução de dados verticais. Esta operação consiste em reduzir o número de valores distintos em determinados atributos. Isso pode proporcionar um melhor desempenho a diversas consultas, sobretudo aquelas que envolvem manipulações e comparações de dados. Existem métodos de redução de valores nominais e métodos de redução de valores contínuos, conforme abaixo:

Redução de valores nominais

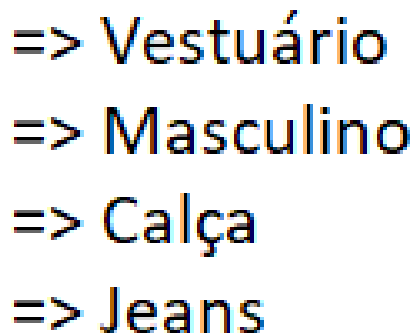
Esta operação é aplicável somente a variáveis nominais. Variáveis nominais possuem um número finito de valores distintos sem ordenação entre valores.

- Identificação de hierarquia de atributos – podemos citar como exemplo os atributos relativos à localização:

=> Estado
=> Cidade
=> Bairro
=> Logradouro

Figura 06

- Identificação de hierarquia de valores – podemos citar dados de produtos e seus grupos:



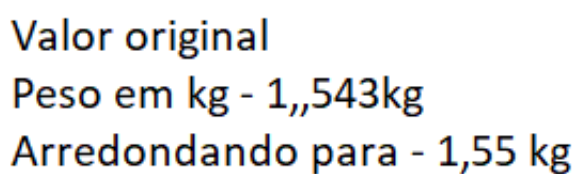
```
graph TD; A[=> Vestuário] --> B[=> Masculino]; B --> C[=> Calça]; C --> D[=> Jeans];
```

Figura 07

Redução de valores contínuos ou discretos

Esta operação é aplicável somente a variáveis contínuas ou discretas. Variáveis contínuas e discretas possuem uma relação de ordenação entre seus valores.

- Arredondamento de valores – também chamado de aproximação de valores:



```
graph TD; A[Valor original] --> B[Peso em kg - 1,,543kg]; B --> C[Arredondando para - 1,55 kg];
```

Figura 08

2.2. Limpeza dos dados

A limpeza de dados desempenha função importante, pois permite identificar se os dados em mãos são suficientes para orientar a tomada de decisão, entendendo quais tipos de conteúdo exigem recolhimento no futuro.

O que é limpeza de dados?

Data cleaning, ou limpeza de dados, é um processo de eliminação de dados inválidos, “sujos” ou pouco informativos. Assim, a empresa consegue focar naqueles dados que são de fato importantes para o processo de tomada de decisão.

A limpeza de dados representa um exercício para se chegar a dados de maior qualidade, partindo do princípio de que, munido-se deles, será possível atingir melhores resultados. É comum confundir o conceito, aliás, com o de data mining, mas vamos de uma vez por todas nos livrar desse comparativo.

Enquanto a limpeza de dados é a identificação dos que são mais valiosos para alimentar um algoritmo de maneira adequada, o data mining volta seu foco à descoberta de padrões. Para que aconteça, a limpeza deve ocorrer antes.

Como acontece a limpeza de dados?

Para compreender melhor como se dá a limpeza de dados, vale a pena recorrer a um exemplo. Imagine que uma empresa queira fazer a segmentação dos produtos que têm à disposição para atender a um público específico, como os donos de pets.

Ela, então, identifica o perfil dessas pessoas (por exemplo: mulher, solteira, entre 25 e 30 anos, moradora de São Paulo, capital).

Mas sua base de dados contém uma série de variações que não condizem com esse perfil.

Ao olhar para os dados, encontram-se informações como São Paulo (estado) ou São Paulo (cidade). Alimentar os algoritmos a partir daí gera um problema na segmentação, portanto é necessário limpar os dados.

Ao aplicar as características do perfil traçado nos dados brutos, é possível eliminar ruídos, como as variações de localização capazes de interferir na tomada de decisão. Isso permite que sejam realmente encontradas informações úteis na hora da escolha.

Por que a limpeza de dados é importante?

Agora que já se sabe como a limpeza de dados funciona, fica mais fácil entender os motivos pelos quais esse processo é importante. Veja, nos tópicos abaixo, as principais razões para fazê-lo.

- **Precisão**

O primeiro motivo consiste no aumento da precisão da análise. Embora o processo de limpeza já represente em si uma análise — em que se decide conteúdos válidos que fazem sentido para a análise de dados —, não é difícil identificar que se trata de uma avaliação preliminar.

A limpeza garante que os dados processados pelos sistemas farão sentido para atingir os objetivos do negócio. Nesse cenário, ainda que preliminar, o processo é fundamental para garantir o máximo de eficácia na análise.



- **Familiarização**

Outro ponto importante na limpeza dos dados, pois ajuda os profissionais a se familiarizarem com os próprios dados. Muitas vezes há dificuldade em realizar o processo nas empresas porque é complicado determinar quais informações estarão à disposição e a quais não se tem acesso.

Ao executar a análise de dados, o profissional ganha familiaridade com o contexto e com o que quer dizer para a empresa. Isso assegura que a transformação dos dados em informação acionável (insights) receba um acréscimo de eficiência.

- **Remoção de inconsistências**

Há, ainda, outra vantagem importante na limpeza dos dados: encontrar inconsistências nas informações armazenadas pela empresa.

Dados em duplicidade podem ser fatais na hora de traçar um perfil do seu consumidor, e o processo de limpeza ajuda a eliminá-los, bem como outras inconsistências — entre elas, erros de digitação — que aparecerem pelo caminho.

Como fazer a limpeza de dados?

Pronto para começar a fazer a limpeza de dados? Então confira o passo a passo, evitando erros ao longo do processo.

- **Passo 1: elimine as respostas em branco**

O primeiro estágio da limpeza de dados é o mais simples: eliminar as respostas em branco, que não trazem nenhuma informação sobre a sua pesquisa.

Respostas em branco podem surgir porque o respondente da pesquisa ou do formulário deixou de inserir algumas informações ou porque estas não foram captadas pelo seu time no momento do estudo.

Dissolver essas respostas não significa a impossibilidade de aprender com elas. Lembre-se de que respostas em branco acontecem por um motivo — pesquisa longa demais, pesquisa pouco envolvente ou pesquisa incompleta, por exemplo. Considere isso na hora de organizar formulários no futuro ou de treinar a sua equipe para preenchê-los.

- **Passo 2: equipare os critérios**

O segundo passo da limpeza de dados consiste em eliminar todos os respondentes que não atendam ao seu critério.

Seguindo o exemplo do tópico anterior, esses respondentes seriam homens, pessoas acima dos 30 anos, ou todos aqueles que não vivem em São Paulo (capital). Eliminar essas pessoas ajuda a chegar aos dados buscados.

- **Passo 3: elimine respostas fora da curva**

Além de riscar da lista os dados incompatíveis com aquilo que se procura, você também deve eliminar respostas fora da curva.

Se no campo “endereço”, em vez de submeter um endereço válido, o respondente inseriu qualquer outra informação, isso também exige remoção da pesquisa, mesmo que outros critérios estejam de acordo à segmentação traçada anteriormente.



- **Passo 4: não faça colunas desnecessárias**

É provável que a sua pesquisa inclua algumas colunas desnecessárias para o tipo de análise feita. Atendo-se ao exemplo anterior, podemos ter a seguinte informação: ensino superior completo.

Se durante a segmentação esse não foi um dos critérios adotados por você para traçar o perfil dos dados procurados, elimine a coluna de informação desnecessária.

- **Passo 5: reduza os *outliers***

Há, ainda, uma importante etapa na limpeza de dados: eliminar os *outliers*. São chamados de *outliers* os dados que fogem do padrão e dificultam o processo de generalização em um modelo de dados. Se estamos generalizando donos de pet como mulheres entre 25 e 30 anos, por exemplo, um *outlier* pode ser uma mulher de 31 anos que segue o mesmo perfil.

Incluir essa *outlier* na sua pesquisa, puxa a média de idade do seu grupo para cima, um resultado não desejado. Por isso, verifique se não há esse tipo de inconsistência na hora de limpar os dados.

O processo de limpeza de dados é de suma importância para o sucesso da análise de dados. Por isso, dê atenção especial a ele e leve em consideração as dicas que aprendeu aqui na hora de limpá-los.

2.3. Codificação

Tem a finalidade de transformar os domínios de valores de determinados atributos do conjunto de dados. A codificação pode ser:

Codificação Numérica – Categórica

• Mapeamento direto - **Figura 09**

SEXO

1 > M

0 > F

• Mapeamento em intervalos

INTERVALO	FREQUÊNCIA
1000 - 1600	3
1600 - 4400	5
4400 - 5400	2

Figura 10: Intervalos com comprimento definido pelo usuário

INTERVALO	FREQUÊNCIA
1000 - 2000	4
2000 - 3000	1
3000 - 4000	2
4000 - 5000	3

Figura 11: Intervalos divididos com igual comprimento



INTERVALO	FREQUÊNCIA
1000 - 1800	4
1800 - 2600	1
2600 - 3400	2
3400 - 4200	3
4200 - 5000	

Figura 12: Divisão em intervalos em função do tamanho da amostra

• Codificação Categórica – Numérica

Representação binária padrão

VALORES ORIGINAIS	REPRESENTAÇÃO BINÁRIA PADRÃO
Casado	001
Solteiro	010
Viúvo	100
Divorciado	011
Outro	110

Figura 13



Representação binária 1-de-N

VALORES ORIGINAIS	REPRESENTAÇÃO BINÁRIA 1-DE-N
Casado	00001
Solteiro	00010
Viúvo	00100
Divorciado	01000
Outro	10000

Figura 14



2.4. Enriquecimento dos dados

Consiste em agregar mais informações a cada registro do conjunto de dados para que estes forneçam mais elementos para o processo de descoberta de conhecimento. Exemplo:

NOME	DATA DE NASCIMENTO		NOME	DATA DE NASCIMENTO	IDADE
João	05/06/1978	→	João	05/06/1978	44
Maria	12/09/2001		Maria	12/09/2001	21
Pedro	26/07/1998		Pedro	26/07/1998	24

Tabela 3 – Enriquecimento dos Dados.



2.5. Normalização dos dados

Esta operação consiste em ajustar a escala dos valores de cada atributo de forma que estes sejam mapeados para valores restritos a pequenos intervalos, tais como -1 a 1, ou de 0 a 1. Tal ajuste faz-se necessário para evitar que alguns atributos, por apresentarem uma escala de valores maiores que outras, influenciem de forma tendenciosa em determinados métodos de análises.

Abaixo, seguem os tipos de normalização existentes:

- Normalização linear;
- Normalização por desvio-padrão;
- Normalização pela soma dos elementos;
- Normalização pelo valor máximo dos elementos;
- Normalização por escala decimal.

3. Descoberta do Conhecimento – Análise de Dados

Uma vez que os dados em si não dizem nada, precisamos analisá-los para decifrar seus significados e relações. A análise de dados resulta na descoberta de padrões consistentes; em outras palavras, as relações entre as variáveis embutidas nos dados.

Quando se identifica o surgimento de padrões, fica mais fácil explicar os números. Quando se extraem esses padrões das variáveis, fica mais fácil resolver o problema. Por exemplo, vamos supor que coletamos dados, mediante pesquisa por telefone, de uma amostra de eleitores sobre as preferências numa eleição presidencial. Ao analisar os dados, tentamos encontrar padrões por região que indiquem provável apoio ao candidato em questão. Para descobrir padrões nos dados, dispõe-se de várias técnicas, desde análises básicas, como gráficos, porcentagens e médias, até métodos estatísticos mais elaborados. As características e a complexidade determinam as técnicas específicas a serem adotadas.

Vários são os modelos adotados por analistas e organizações para aplicar raciocínio analítico e para decidir sobre os dados. Se quisermos compreender os tipos de modelos mais eficazes em determinada situação, primeiro precisamos esboçar as características da situação com que se defrontam os analistas. Três são as questões a serem consideradas na identificação do modelo adequado:

- **Quantas variáveis serão analisadas ao mesmo tempo?**
- **Queremos responder a questões de descrição ou de inferência?**
- **De que níveis de medida se dispõe nas variáveis de interesse?**

3.1. 4 tipos de análise de dados

3.1. 4 Tipos de análise de dados

Agora sim, você já tem em mãos tudo o que precisa para esclarecer, de vez, quais são os principais tipos de análise de dados. A seguir, você conhecerá as definições e indicações para aplicação das 4 principais metodologias de avaliação de informações.

a. análise descritiva

Como o próprio nome diz, a análise descritiva é um dos tipos de análise de dados baseados em fatos. Isso significa que, na prática, este tipo de avaliação de dados é feito a partir de resultados obtidos. São exemplos de análise de dados descritiva:

- Relatórios;
- Segmentação e controle de clientes;
- Análises de negócio;
- Aplicação de métricas;
- Avaliação de resultados.

Um dos principais usos para a análise descritiva é orientar a construção de estratégias.

b. análise preditiva

O mais popular dos tipos de análise de dados é justamente o modelo preditivo. Como o nome diz, sua essência está na previsão de cenários futuros com base na análise de padrões revelados pela base de dados.

É importante saber que, em uma análise preditiva, não é possível prever o que vai acontecer, mas sim, o que deve acontecer SE determinadas condições se cumprirem.

Quer ver um exemplo de análise de dados preditiva?

Suponhamos que sua empresa esteja apreensiva quanto à possível entrada de um concorrente no mercado.

A análise preditiva não será capaz de te dizer se o concorrente iniciará ou não suas atividades em breve. Em contrapartida, te ajudará a enxergar o que poderá acontecer SE o concorrente, de fato, entrar no mercado, tomando como base situações anteriores com contextos semelhantes.

Podemos dizer, assim, que o objetivo da análise preditiva é determinar uma tendência, correlação, causa ou probabilidade.

c. análise prescritiva

A análise prescritiva é o próximo passo após os resultados da avaliação preditiva. Isso porque uma prescrição é uma recomendação a algo potencialmente previsto.

Sendo assim, a melhor forma de obter uma análise prescritiva é fazendo projeções (predições) e, então, direcionando esforços para obter o melhor resultado a partir das possibilidades.

Por ser uma análise de dados constantemente mutável (já que está sempre condicionada a previsões e predições), os modelos analíticos prescritivos são comumente apoiados por tecnologias como inteligências artificial, *machine learning* e algoritmos. As ferramentas ajudam a fazer sugestões com base em padrões diferenciados e percepções de objetivos organizacionais, limitações e fatores de influência.

d. análise diagnóstica

Aqui está outro tipo de análise de dados concentrada em algo que já aconteceu (assim como a análise descritiva). A análise diagnóstica, diferentemente da descritiva, tem como objetivo encontrar relações de causa e efeito para destrinchar um acontecimento.

É claro que estabelecer este tipo de relação baseado em um acontecimento passado não é tarefa fácil. Por isso mesmo, o processo é baseado em probabilidades.

Conhecer os principais tipos de análise de dados pode ajudar a sua empresa a dominar as informações-chave do negócio na palma da mão. Lembre-se de que, com a ajuda das melhores ferramentas, é possível automatizar momentos importantes da análise de dados (como a consolidação de relatórios e a criação de gráficos), mantendo a equipe focada naquilo que realmente importa: a estratégia.



4. Pós-Processamento

Abrange o tratamento do conhecimento obtido na etapa anterior, objetivando facilitar a interpretação e a avaliação por especialista do domínio da aplicação. A essência dessa etapa é descrever o problema e contar as histórias por trás dele.



4.1. O que é visualização de dados?

A visualização de dados é a apresentação de informações quantitativas em uma forma gráfica. Em outras palavras, as visualizações de dados transformam grandes e pequenos conjuntos de dados em visuais que são mais fáceis para o cérebro humano entender e processar.

As visualizações de dados são surpreendentemente comuns em sua vida cotidiana, mas geralmente aparecem na forma de tabelas e gráficos conhecidos. Uma combinação de múltiplas visualizações e bits de informação é muitas vezes referida como “infográfico”.

As visualizações de dados podem ser usadas para descobrir fatos e tendências desconhecidos. Você pode ver visualizações na forma de gráficos de linhas para exibir as alterações ao longo do tempo. Gráficos de barras e colunas são úteis ao observar relacionamentos e fazer comparações. Gráficos de pizza são uma ótima maneira de mostrar partes de um todo. E mapas são a melhor maneira de compartilhar visualmente dados geográficos.

4.2. O que faz uma boa visualização de dados?

Boas visualizações de dados são criadas quando a comunicação, a ciência de dados e o design colidem. As visualizações de dados feitas corretamente oferecem informações importantes sobre conjuntos de dados complicados de maneiras significativas e intuitivas. Estatístico americano e professor de Yale, Edward Tufte acredita que excelentes visualizações de dados consistem em "ideias complexas comunicadas com clareza, precisão e eficiência".

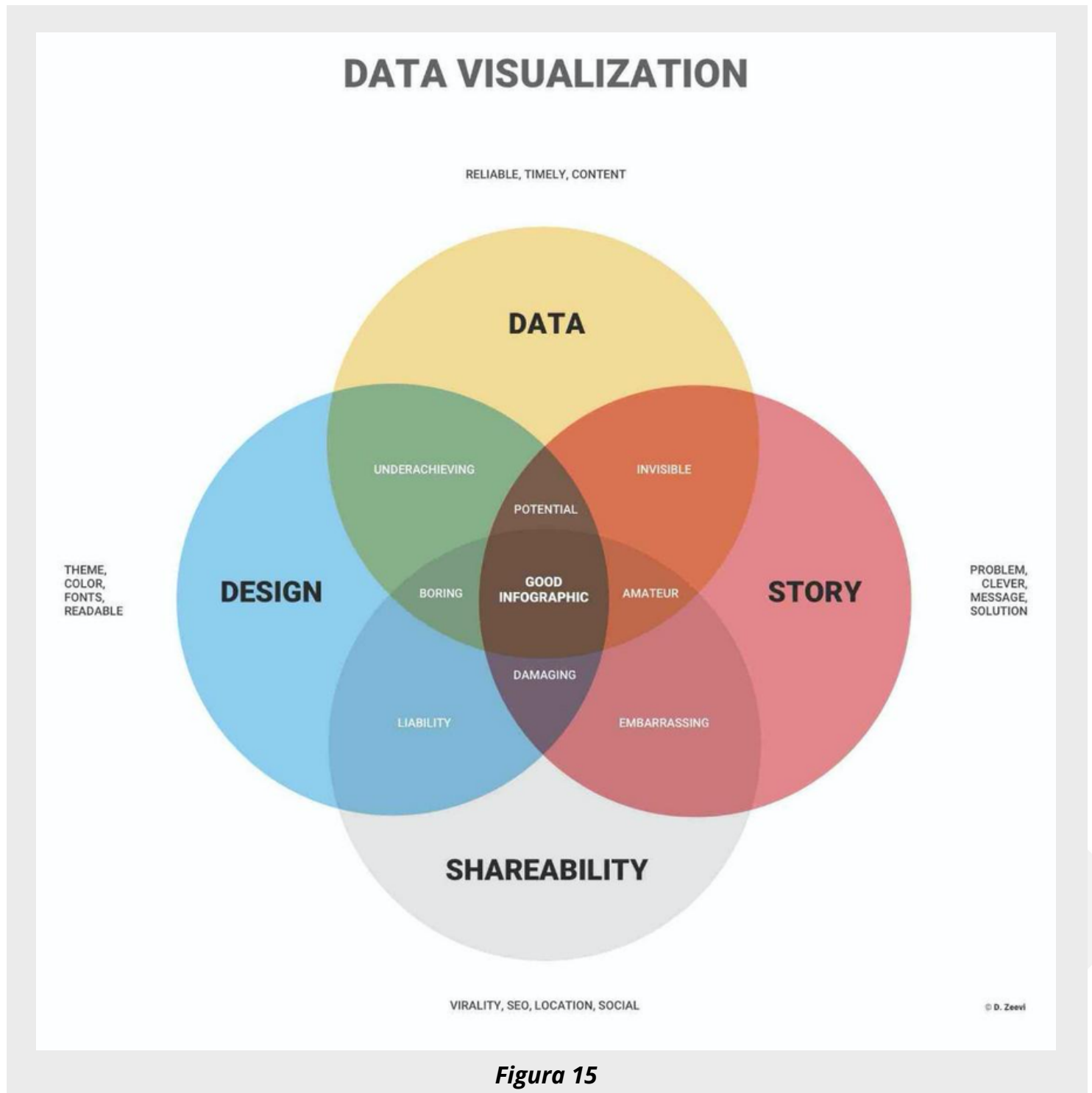


Figura 15

Para criar **boas visualizações de dados**, você precisa começar com dados limpos, bem fornecidos e completos. Quando seus dados estiverem prontos para serem visualizados, você precisará escolher o gráfico correto. Isso pode ser complicado, mas há muitos recursos disponíveis para ajudá-lo a escolher o tipo certo de gráfico para seus dados.

Depois de ter decidido qual tipo de gráfico é melhor, você precisa projetar e personalizar sua visualização ao seu gosto. Lembre-se de que a simplicidade é fundamental. Você não deseja adicionar elementos que distraiam a atenção sobre os dados.

Agora que sua visualização está completa, é hora de publicar e compartilhar com seus colegas, clientes ou leitores.

4.3. Por que a visualização de dados é importante?

- **Melhor tomada de decisão**

Hoje, mais do que nunca, as organizações estão usando visualizações de dados e ferramentas de dados para fazerem melhores perguntas e tomar melhores decisões. Novas tecnologias de computação e novos programas de software fáceis de usar facilitaram o aprendizado sobre suas empresas e as ajudaram a tomar melhores decisões de negócios baseadas em dados.

A forte ênfase nas métricas de desempenho, nos painéis de dados e nos Key Performance Indicators (KPIs) mostram a importância de medir e monitorar os dados da empresa. Informações quantitativas comuns medidas por empresas incluem unidades ou produto vendido, receita por trimestre, despesas de departamento, estatísticas de funcionários e participação de mercado da empresa.

- **Storytelling Significativo**

As visualizações de dados e gráficos de informações (infográficos) tornaram-se uma ferramenta essencial para a grande mídia atual. O jornalismo de dados está em ascensão e os jornalistas dependem consistentemente de ferramentas de visualização de qualidade para ajudá-los a contar histórias sobre o mundo ao nosso redor. Muitas instituições respeitadas adotaram completamente as notícias baseadas em dados, incluindo o The New York Times, o The Guardian, o Washington Post, a Scientific American, a CNN, a Bloomberg, o The Huffington Post e The Economist.

Os profissionais de marketing também se beneficiam enormemente da combinação de dados de qualidade e narrativa emocional.

Bons profissionais de marketing tomam decisões baseadas em dados diariamente, mas compartilhar com seus clientes requer uma abordagem diferente – que os atinja tanto de maneira inteligente quanto emocional. As visualizações de dados ajudam os profissionais de marketing a compartilhar sua mensagem usando estatísticas e informações.

O Professor de Marketing da Universidade de Stanford Jennifer L. Aaker disse:

"Quando dados e histórias são usados juntos, eles ressoam com o público tanto em um nível intelectual e emocional."

Jennifer L. Aaker



4.4. Alfabetização de dados

Ser capaz de compreender e ler as visualizações de dados tornou-se um requisito necessário para o século XXI. Como as ferramentas e recursos de visualização de dados se tornaram prontamente disponíveis, espera-se que cada vez mais profissionais não-técnicos sejam capazes de coletar informações a partir de dados.

4.5. A história da visualização de dados

A visualização de dados existe há séculos, e muitos concordariam que começou no final dos anos 1700 com William Playfair – mais conhecido como o "pai da estatística". Acredita-se que a Playfair tenha inventado a linha, a barra e o gráfico que usamos muitas vezes hoje.

Florence Nightingale é famosa por seu trabalho como enfermeira durante a Guerra da Criméia, mas também foi jornalista de dados, conhecida por seus diagramas "coxcomb" ou "rose". Esses gráficos revolucionários a ajudaram a lutar por melhores condições hospitalares, salvando as vidas dos soldados.

Uma das visualizações de dados históricos mais conhecida vem de Charles Joseph Minard. Minard era um engenheiro civil francês famoso por sua representação de dados numéricos em mapas. Seu trabalho mais famoso é o mapa da campanha russa de Napoleão de 1812, mostrando a perda dramática de seu exército sobre o avanço em Moscou e o seguinte retiro.

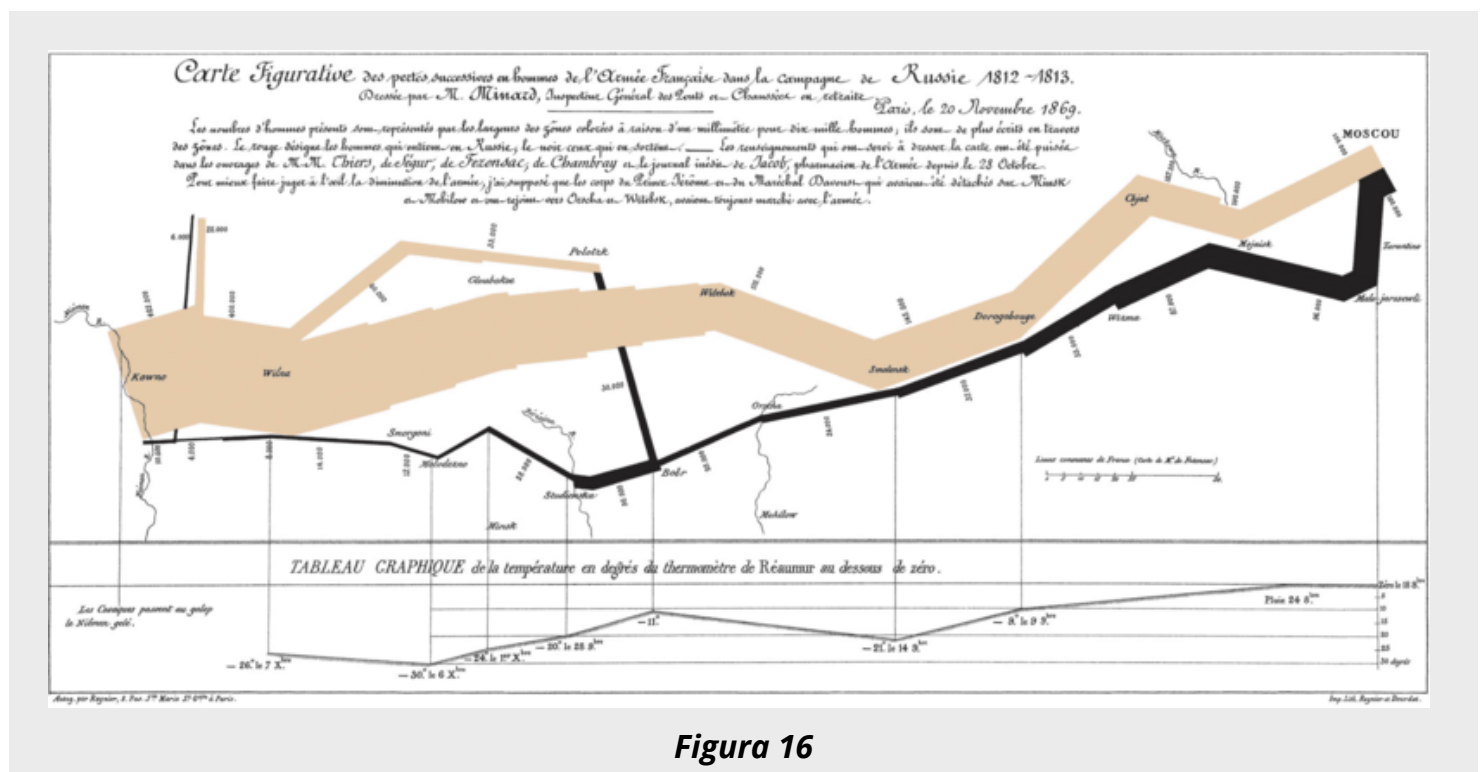


Figura 16

5. Gestão de Dados

A Gestão de Dados é também conhecida no mercado por diversos termos:

- Gestão da informação;
- Gestão da informação empresarial;
- Gestão dos dados empresariais.

Seguem abaixo algumas definições de Gestão de Dados:

"Gestão de Dados é a disciplina responsável por definir, planejar, implantar e executar estratégias, procedimentos e práticas necessárias para gerenciar de forma efetiva os recursos de dados e informações das organizações, incluindo planos para sua definição, padronização, organização, proteção e utilização." [BARBIERI, 2011]

"Gestão de Dados é a função na organização que cuida do planejamento, controle e entrega dos ativos de dados e de informação. Esta função inclui: as disciplinas do desenvolvimento, execução e supervisão de planos, políticas, programas, projetos, processos, práticas e procedimentos que controlam, protegem, distribuem e aperfeiçoam o valor dos ativos de dados e informações." [BARBIERI, 2011]

Em suma, a disciplina Gestão de Dados é responsável por zelar da melhor forma possível, por intermédio de seus profissionais de tecnologia e de negócios, os dados e metadados das organizações, fazendo com que sejam aderentes às necessidades do negócio, únicos, íntegros, confiáveis, manuteníveis, conhecidos, performativos, legíveis e disponíveis a quem realmente precisa ter o acesso.

Apoiada por organizações internacionais voltadas para o desenvolvimento dos assuntos ligados à Gestão de Dados, tais como o Data Governance Institute e a DAMAS – Data Management International–, aos poucos a Gestão de Dados vem surgindo no mercado brasileiro de forma muito mais abrangente, englobando funções anteriormente esquecidas ou mal gerenciadas pelas organizações.

O escopo de atuação da disciplina "Gestão de Dados" e a escala de sua implementação nas empresas variam amplamente de acordo com o tamanho, os meios e a experiência das organizações. Por outro lado, a principal intenção da adoção da Gestão de Dados nas empresas geralmente é a mesma: fornecer mecanismos de utilização do conhecimento das informações para tomar decisões ágeis e corretas.

A Gestão de Dados é importante para as organizações independentemente do seu tamanho, área de atuação e finalidade.

5.1. Princípios

O guia DAMA-DMBOK estabelece cinco princípios básicos que orientam a adoção da Gestão de Dados nas organizações. O conjunto desses princípios estabelece uma filosofia de trabalho que deve sempre ser levada em consideração pelos profissionais que atuam nesta área.

Os princípios estabelecidos pelo guia DAMA-DMBOK são os seguintes:

- Dados e informações são ativos valiosos das organizações;
- Como todo ativo, os dados devem ser gerenciados, assegurando qualidade adequada, segurança, integridade, proteção, disponibilidade, compreensão e uso efetivo;
- A responsabilidade da Gestão de Dados é compartilhada entre os gestores de dados de negócio e os profissionais de gestão de dados de tecnologia;
- Gestão de dados é uma disciplina de negócios e um conjunto de funções relacionadas;
- Gestão de dados é uma profissão emergente e em amadurecimento.

5.2. Principais Funções

A versão atual do guia DAMA-DMBOK estabelece dez funções primárias:

- Governança de Dados: função que representa o exercício da autoridade e o controle de estratégias, políticas, regras, procedimentos, papéis e atividades envolvidos com os ativos de dados. A Governança de Dados é considerada a função central do framework e influencia todas as demais funções do guia DAMA-DMBOK.

- Gestão da Arquitetura de Dados: função responsável por definir as necessidades de dados (geralmente corporativos) da empresa. A função também é responsável por criar e manter a Arquitetura Corporativa de Dados de acordo com os objetivos estratégicos da empresa.

- Gestão do Desenvolvimento dos Dados: função que representa as atividades de dados dentro do ciclo de desenvolvimento de sistemas, tais como: Modelagem de Dados (incluindo as avaliações em modelos de dados), análise de requisitos de dados, projeto de banco de dados, implantação e manutenção dos bancos de dados.

- Gestão de Operação de Dados: função responsável por manter armazenados os dados ao longo do seu ciclo de vida após a criação das estruturas para este propósito. O ciclo se inicia na criação e/ou aquisição dos dados e vai até o arquivamento final ou a sua eliminação.

- Gestão da Segurança dos Dados: função responsável por definir e manter as políticas de segurança e procedimentos a fim de prover a adequada autenticação, utilização, acesso e auditoria de dados.

- Gestão de Dados Mestres e Dados de Referência: função responsável por definir e controlar atividades para garantir a consistência e disponibilização de visões únicas dos dados mestres e de referência da empresa.

- Gestão de Data Warehousing e Business Intelligence: função responsável por definir e controlar processos para prover dados de suporte à decisão, geralmente disponibilizados em aplicações analíticas.

- Gestão da Documentação e Conteúdo: função dedicada a planejar, implementar e controlar atividades para armazenar, proteger e acessar os dados não estruturados da empresa.

- Gestão de Metadados: função responsável por gerir e armazenar os metadados da empresa, além de viabilizar formas de acesso.

- Gestão da Qualidade dos Dados: função dedicada à gestão das atividades para aplicação de técnicas de Qualidade de Dados com o propósito de medir, avaliar, melhorar e garantir a qualidade dos dados da empresa. A próxima versão do guia DAMA-DMBOKO já prevê a inclusão de mais uma função em seu framework. Trata-se da função de Integração de Dados.



Funções do DAMA-DMBOK



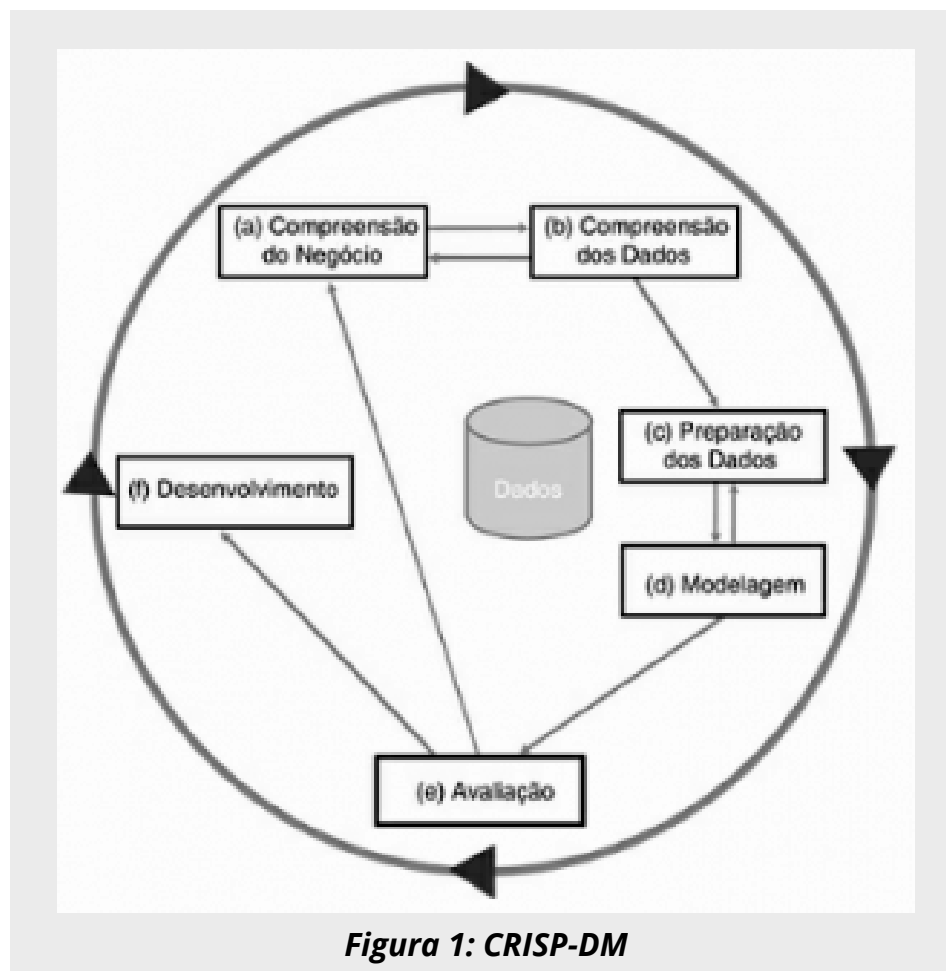
Figura 17

6. CRISP-DM – Conceitos

A Mineração de Dados é uma atividade que tem agregado muito valor à descoberta da informação, e seus conceitos ganham força quando se utiliza o CRISP-DM como modelo de processo.

O CRISP-DM – Cross Industry Standard Process for Data Mining (Processo Padrão Inter-Indústrias para Mineração de Dados) e surgiu em 1996 como forma de apoio ao processo de descoberta do conhecimento, o famoso KDD – Knowledge Discovery in Databases (Descoberta de Conhecimento em Bases de Dados).

O CRISP-DM é constituído por 6 fases: Compreensão do Negócio, Compreensão dos Dados, Preparação dos Dados, Modelagem, Avaliação e Desenvolvimento. Vejamos abaixo uma figura que representa o processo CRISP-DM e como as fases se inter-relacionam:



6.1. As 6 fases do CRISP-DM

6.1.1. Compreensão do Negócio

Conhecer e compreender o problema a ser resolvido é de suma importância neste processo. Muitas vezes nos deparamos com pessoas que fazem parte do negócio e possuem dificuldades de definir o que é de fato o seu negócio. Perceba que no próprio diagrama existem retornos para esta fase a partir de outras etapas do processo, ou seja, podem ocorrer falhas durante o processo por má compreensão do negócio. Para fins de tentar mitigar esses riscos de má compreensão, o CRISP-DM prevê algumas atividades nesta fase, conforme indico a seguir:

- **Identificar os especialistas na organização:** os especialistas da área de negócio (domínio), o pessoal da TI, bem como os responsáveis pela tomada de decisão, precisam conhecer o processo. É comum nesta etapa, realizar treinamentos para fins de nivelamento de conhecimento do pessoal em torno do processo KDD e onde o CRISP-DM estará atuando;
 - **Levantar e esboçar as necessidades e expectativas:** as pessoas que estarão envolvidas no processo, precisam expor suas necessidades, principalmente aqueles que irão lidar com os dados e informações obtidos para posterior análise e tomada de decisão. É comum serem feitas rodadas de reuniões com estas pessoas para fins de elencar os objetivos e as necessidades;
- **Levantamento dos hardwares e softwares:** a organização não é feita apenas de pessoas, mas também de ferramentas e neste sentido se faz necessário conhecer o que existe disponível ou não. O CRISP-DM orienta que o processo seja realizado em plataforma que possua arquitetura expansível, com capacidade de suportar grandes volumes de dados, com grandes chances destes dados serem heterogêneos e que possa ter capacidade de processamento compatível com o volume de dados;
- **Fazer inventário das bases de dados existentes:** é importante conhecer o que a organização possui de bases de dados internas e do acesso e uso de bases de dados externas. Neste momento, se faz necessário também observar potenciais bases de dados externas que guardem relação com o negócio da organização e por isso podem fazer parte do processo;
- **Verificar a existência e Data Warehouses na organização:** caso a organização possua dados armazenados em bases multidimensionais, como os Data Warehouses e Data Marts, é possível que ela já tenha estabelecido um processo ETL e neste caso haverá um ganho de esforço no restante do processo.

6.1.2. Compreensão dos Dados

Esta fase geralmente é executada juntamente com a fase anterior (Compreensão do Negócio), onde o estudo sobre as informações coletadas se faz necessário e deve ser feita de forma minuciosa.

Se a solução do problema de negócios é o objetivo, os dados compreendem a matéria-prima disponível a partir da qual a solução será construída. É importante entender os pontos fortes e as limitações dos dados, porque raramente há uma correspondência exata com o problema. Por exemplo, uma base de dados de clientes, uma base de dados de vendas e-commerce e uma base de dados de respostas de satisfação dos clientes contêm informações diferentes, podem abranger diferentes publicações, que se cruzam, e podem ter, assim, diferentes graus de confiabilidade sobre estes dados, daí a necessidade de compreendê-los.

Vejamos algumas das principais atividades envolvidas nesta fase:

- **Conhecer e entender os dados disponíveis:** é preciso entender bem os atributos e dados que foram levantados, para fins de definir os objetivos do restante do processo. É comum que não haja documentação completa sobre os dados, através de metadados e dicionário de dados, o que vai requerer retorno à etapa anterior para fins de dirimir dúvidas e assim completar ou elaborar as documentações sobre os dados.

- **Avaliação da qualidade dos dados disponíveis:** o propósito para o qual os dados foram disponibilizados é o principal alvo aqui. Os dados atendem ao propósito? Possuem muitos ruídos? Precisam ser transformados? Estes são alguns dos questionamentos feitos durante a avaliação da qualidade dos dados e, para isso, recursos de limpeza dos dados serão usados, porém aderentes ao domínio do negócio.
- **Verificar se a volumetria dos dados atende ao negócio:** refere-se à quantidade de dados que será utilizada, pois amostras pequenas de dados podem não ser úteis para o processo, portanto é importante que seja feito um acordo para o fornecimento de volume de dados adequado.

6.1.3. Preparação dos Dados

Essa é uma fase do processo que antecede a construção de modelos e que irá adequar os dados, compreendendo ações de pré-processamento.

As diversas ferramentas analíticas que podem ser usadas nesta fase, apesar de oferecerem muitos recursos, impõem alguns requisitos sobre os dados quanto ao seu formato, gerando a necessidade de formatações e/ou transformações dos dados.

Alguns exemplos naturais da preparação de dados são a sua conversão para o formato tabular, retirando ou até inserindo valores ausentes e convertendo dados para diferentes tipos. Algumas técnicas de mineração focam em dados chamados de simbólicos e categóricos, enquanto outras lidam apenas com valores numéricos.

Vejamos algumas atividades importantes nesta fase:

- **Seleção dos dados para análise:** nesta etapa, serão selecionados apenas os dados que irão de fato ser analisados, sendo eles internos ou externos à organização, inclusive planejando quais destes dados serão indicados para carga.
- **Limpeza dos Dados:** inconsistências sempre surgem nos dados e nesta atividade elas serão removidas ou ajustadas, assim como irão completar dados que estejam ausentes com algum padrão de dados estabelecidos e que tenham relação com o negócio da organização.

- **Adequar formato dos dados;**
- **Construir novos atributos com base nos atributos existentes (atributos derivados).**

6.1.4. Modelagem

Nesta fase as atividades estão dentro das características similares às da Mineração de Dados no KDD, onde, por exemplo, serão escolhidas as técnicas mais adequadas para modelagem, com base em algoritmos de mineração, em que testes iniciais voltados à calibração de parâmetros dos algoritmos serão feitos. Pode ser que durante essa atividade, haja necessidade de retorno à atividade de preparação dos dados, visto que algumas técnicas de modelagem apresentam demandas diferentes quanto ao formato do conjunto de dados utilizado, ou até mesmo podem ocorrer falhas durante a construção do modelo.

A modelagem é o principal local onde as técnicas de mineração de dados são aplicadas aos dados. É importante ter alguma compreensão dos conceitos sobre mineração de dados, incluindo os tipos de técnicas e os algoritmos existentes.

6.1.5. Avaliação

O insumo desta fase é a saída da fase anterior em forma de um ou mais modelos. A avaliação vai checar se o modelo elaborado condiz com as expectativas da organização e do que foi definido anteriormente na fase inicial do processo. O resultado desta avaliação pode ser aceitável ou pode resultar na necessidade de revisão das fases anteriores, a fim de redefinir alguns passos.

**6.1.6. Desenvolvimento (implantação)**

Esta fase consiste na definição das fases de implantação do projeto de Mineração de Dados, levando em consideração que o modelo resultante da fase de modelagem precisa ser factível de ser usado, ou seja, o modelo para obtenção de conhecimento precisa, além de ser aderente às necessidades da organização, ser interpretável e ter capacidade operacional.

Será elaborado relatório final do processo, que apresenta os resultados obtidos além de possíveis alternativas de ação no processo de descoberta de conhecimento aplicado na organização.

Abaixo na tabela 01, temos as fases e suas atividades:

ENTENDIMENTO DO NEGÓCIO	ENTENDIMENTO DOS DADOS	PREPARAÇÃO DOS DADOS	MODELAGEM	AValiação	DESENVOLVIMENTO
Determinação dos Objetivos	Coleta Inicial dos Dados	Seleção dos Dados	Seleção de Técnicas de Modelagem	Avaliação dos Resultados	Aplicação do Projeto
Situação a ser Avaliada	Descrição dos Dados	Limpeza dos Dados	Geração de Projetos de Testes	Revisão do Processo	Planejar Entrega
Determinar Metas de Mineração	Exploração dos Dados	Construção dos Dados	Construção do Modelo	Determinar Próximos Passos	Plano de Monitoração e Manutenção
Produzir o Projeto	Verificação da Qualidade dos Dados	Integração dos Dados	Modelo a ser Avaliado		Produção do Relatório Final
		Formatação dos Dados	Ajustes no Modelo		Revisão do Projeto
		Descrição das Bases			Lições Aprendidas

Tabela 01



Referências

ANDRADE, Eduardo. **Introdução à Pesquisa Operacional – Métodos e Modelos para Análise de Decisões**. 4. ed. Rio de Janeiro: LTC, 2014.

BARBIERI, Carlos. **BI2 – Business Intelligence: modelagem e qualidade**. 1ª ed. Rio de Janeiro: Elsevier, 2011.

ELMASRI, R.; NAVATHE, S. **Sistemas de Banco de Dados [BV:PE]**. 7. ed. São Paulo: Pearson, 2018.

MACHADO, Felipe N. R. **Tecnologia e Projeto de Data Warehouse**. 6. ed. São Paulo: Érica, 2013.

SILVA, Leandro. **Introdução à Mineração de Dados com aplicações em R**. 1ª ed. Rio de Janeiro: Elsevier, 2016.



Digital College

ENSINO DE HABILIDADES DIGITAIS

digitalcollege.com.br