Fundamentos em Análise de Dados

Bootcamp Analista de Dados

Fernanda Farinelli

2021

Fundamentos em Análise de Dados

Bootcamp Analista de Dados

Fernanda Farinelli

Sumário

Capítulo 1. Conceitos fundamentais de Big Data:

A evolução das tecnologias de informação e comunicação, além do surgimento da internet mudou o dia a dia das pessoas, trazendo as atividades humanas para o mundo virtual. A *World Wide Web*, a maior rede de informação global, em ritmo evolucionário, passou por fases que ficaram conhecidas como Web 1.0, Web 2.0 e Web 3.0 (SHIVALINGAIAH; NAIK, 2008). Vivemos cercados por uma grande quantidade de dados, que apoiam nossas decisões. Tal disponibilidade oferece oportunidades para a obtenção de conhecimento, ou seja, ao submeter os dados a processos de análise, obtém-se informação e conhecimento útil nos processos decisórios das organizações (HEATH; BIZER, 2011).

Dados, tipos de dados e fontes de dados

Nas últimas décadas, os dados assumiram um papel vital na estratégia das empresas, tornando-se um dos grandes ativos existentes no patrimônio das organizações (DAMA, 2009, p. 1). Veja no Quadro 1 o que são dados, informação e conhecimento:

Quadro 1: Dados, informação e conhecimento.

Dado	Informação	Conhecimento
Simples observações sobre o estado do mundo.	Dados dotados de relevância e propósito.	Informação valiosa da mente humana.
Facilmente estruturado. Facilmente obtido por máquinas. Frequentemente quantificado. Facilmente transferível.	Requer unidade de análise. Exige consenso em relação ao significado. Exige necessariamente a mediação humana.	Inclui reflexão, síntese, contexto. De difícil estruturação. De difícil captura em máquinas. Frequentemente tácito. De difícil transferência.

Fonte: DAVENPORT (1998).

Tal definição enfatiza o papel dos dados em representar fatos sobre o mundo, ou seja, *dados são fatos capturados, armazenados e expressos como texto, números, gráficos, imagens, sons ou vídeos* (DAMA, 2009; 2017). Nossos dados não se resumem a ações, mas podem ser objetos, localizações, quantidades, textos,

imagens e áudios, ou qualquer coisa que possa ser digitalizada e armazenado em algum banco de dados (DAMA, 2017).

As atividades e ações virtuais e os diversos sistemas de informação produzem, coletam e analisam dados. Os dados podem vir de diferentes fontes de dados, ou seja, uma fonte é o local de onde o dado é coletado ou adquirido. Podem ser arquivos, banco de dados, portal de notícias ou até mesmo um *feed* de notícias. As fontes de dados podem ser qualquer dispositivo ou estrutura que forneça dados, localizada ou não no mesmo computador ou dispositivo que o programa de coleta.

Os dados podem assumir diferentes formatos ou tipos conforme sua origem, a Figura 1 apresenta um quadro sumarizando comparativamente estes tipos de dados.

Figura 1: Comparativo dos tipos de dados.

Os dados armazenados nos bancos de dados relacionais são considerados dados estruturados. Como dados semiestruturados, podemos citar os dados armazenados em arquivos XML. Como dados não estruturados, temos os dados originários de mídias sociais, imagens, vídeos e áudios. A maior parte dos coletados atualmente são dados não estruturados. Acredita-se que 95% dos dados gerados hoje são em formato não-estruturado (MAYER-SCHÖNBERGER; CUKIER, 2013:47).

Big data

Esse volume de dados e suas diversas fontes e formatos levou ao fenômeno que ficou conhecido como *Big Data*. Termo cunhado em meados dos anos 90 por Michael Cox e David Ellsworth, cientistas da Nasa, que discutiram sobre os desafios da visualização de grandes volumes de dados, aos limites computacionais tradicionais de captura, processamento, análise e armazenamento (COX; ELLSWORTH, 1997). O termo *Big Data* é usado para descrever esse grande conjunto de dados que desafia os métodos e ferramentas tradicionais para manipulação de dados, considerando um tempo razoável de resposta. *Big Data* é caracterizado pela tríade volume, variedade e velocidade (LANEY, 2001). Ainda temos duas importantes características, veracidade e valor (TAURION, 2013).

O ***volume*** diz respeito à quantidade de dados que são produzidos e coletados pelas organizações. As organizações coletam dados de diversas fontes, implicando na ***variedade*** dos tipos (estruturados, semiestruturados e não estruturados) e formatos dos dados coletados. A ***velocidade*** diz respeito ao quão rápido os dados estão sendo produzidos e quão rápido os dados devem ser tratados para atender a demanda da organização. As decisões são tomadas em tempo real. Temos ainda a ***veracidade*** ou confiabilidade dos dados, ou seja, eles devem expressar a realidade e ser consistentes. Enfim, o ***valor*** diz respeito à utilidade dos dados ao negócio, ou seja, como agregam valor (TAURION, 2013).

Web semântica

A Web Semântica é uma extensão da web que estrutura o significado de seu conteúdo de forma clara e bem definida, permitindo aos computadores interagir entre eles pela troca de informações. Sua principal motivação é ter uma web de dados, no qual tais dados sejam significativos tanto para os humanos quanto para as máquinas (BERNERS-LEE; HENDLER; LASSILA, 2001).

Na Web Semântica, são os vocabulários que definem os conceitos e relacionamentos usados para descrever e representar uma área de interesse. Muitas vezes uma ontologia pode ser empregada quando se tem uma coleção de termos

mais complexa e formal. O papel dos vocabulários, portanto, é o de auxiliar na integração dos dados, tratando os casos de ambiguidade de termos usados em diferentes bases de dados por exemplo.

Dados abertos

O conceito de dados abertos remete à ideia de conteúdo aberto, ou seja, disponível para todos. Dados abertos são dados que podem ser livremente publicados na web, seguindo alguns padrões pré-definidos. A partir de sua publicação, podem ser reutilizados e redistribuídos por qualquer pessoa ou aplicativo, sujeitos, no máximo, à exigência de atribuição da fonte e compartilhamento pelas mesmas regras (ISOTANI; BITTENCOURT, 2015; OKI, 2019).

Linked data e Linked open data

Fundamenta-se na ideia de interligar dados na web em vez de documentos. *Linked data* (LD) ou dados interligados. Trata-se de uma forma de publicar dados na web de forma estruturada, de modo que uma pessoa ou máquina possa explorar esses dados. Relacionado à web semântica, propõe um conjunto de princípios, padrões e protocolos a serem adotados para publicar dados na web e para interligar os dados. As ligações permitem aos usuários da web navegar entre diferentes fontes. Além disso, as ferramentas de busca ficam aptas a indexar a web e fornece recursos de pesquisa mais sofisticados sobre o conteúdo rastreado (BERNERS-LEE, 2006; BIZER; HEATH; BERNERS-LEE, 2009; HEATH; BIZER, 2011).

Adicionalmente, temos o conceito de dados abertos interligados ou *Linked Open Data,* que remete a ideia de conteúdo aberto ou disponível para todos mas com interconexões entre os dados, ou seja, são dados interligados que se encontram disponíveis livremente na web (BERNERS-LEE, 2006).

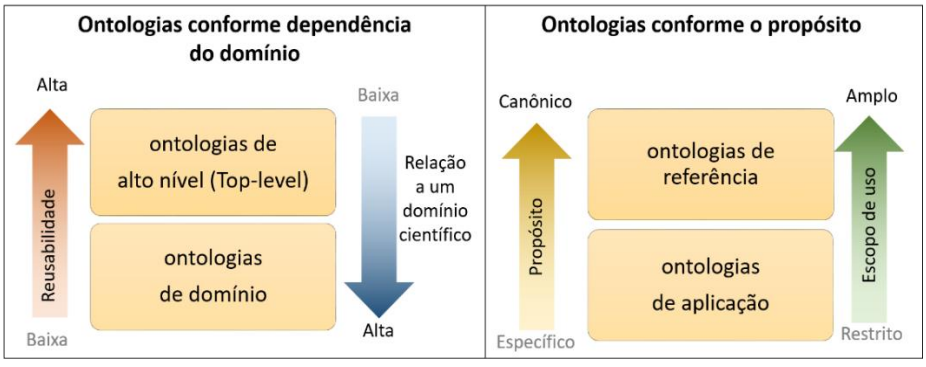
Ontologias

Ontologia é um termo polissêmico e objeto de pesquisa em diversas áreas, como: Filosofia, Ciência da Computação e Ciência da Informação. A palavra "ontologia" é derivada do grego, em que "*Onto"* exprime a noção do ser e "*Logia"* é algo dito ou a maneira de dizer. Ela pode ser entendida como disciplina filosófica ou como artefato representacional. Como disciplina da Filosofia, a Ontologia estuda a natureza da existência das coisas. Como artefato representacional, a ontologia representa conhecimento acerca de vários domínios de conhecimento através da formalização das relações entre termos e conceitos (ALMEIDA, 2013).

As ontologias se classificam conforme descrito abaixo e ilustrado na Figura 2 (FARINELLI, 2017).

- Ontologias de alto nível: descrevem conceitos amplos independentes de um domínio particular. Ex: relacionadas a espaço, tempo, eventos etc.
- Ontologias de referência: descrevem conceitos relacionados à atividade ou a tarefas genéricas, independentes de domínio. Ex: diagnóstico.
- Ontologias de domínio: descrevem conceitos relacionados a domínios específicos, como direito, computação etc. É a categoria mais comum.
- Ontolologias de aplicação: descrevem conceitos dependentes de um domínio e tarefa específicos.

Figura 2: Classificação das ontologias.



Fonte: Traduzido de FARINELLI (2017).

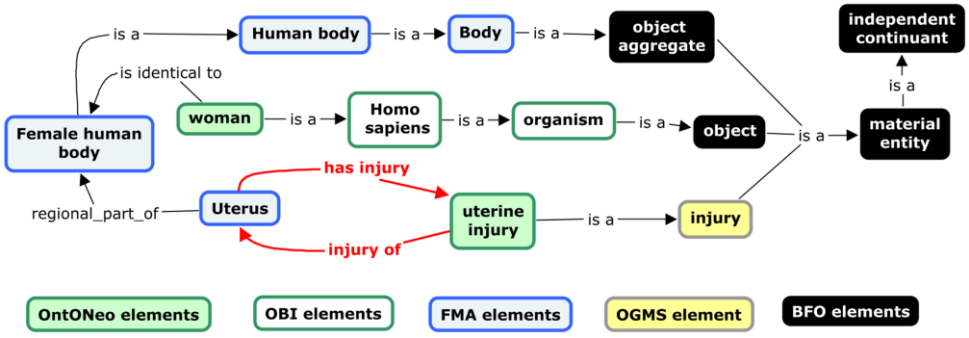
Uma ontologia pode ser muito complexa, com milhares de conceitos ou muito simples, descrevendo apenas um ou dois conceitos. A especificação de uma ontologia inclui os seguintes elementos (FARINELLI, 2017):

- Entidade: é algo que você deseja representar em um domínio particular. Qualquer coisa que exista, existiu ou irá existir. Ex.: eventos, processos, objetos.
- Instância ou indivíduos: representam uma unidade de objetos específicos de uma entidade, ou seja, indivíduos de um determinado universal.
- Atributos: Propriedades relevantes da entidade/classe ou instância que ajudam a descrevê-la.
- Relacionamento: descreve o tipo de interação entre duas classes, duas instâncias ou uma classe e uma instância.
- Cardinalidade: uma medida do número de ocorrências de uma entidade associada a um número de ocorrências em outra.
- Axioma: uma proposição lógica que é considerada verdadeiro. Restringem a interpretação e o uso das classes envolvidas na ontologia.

As ontologias descrevem entidades sobre a perspectiva dos universais e particulares. Os particulares ou indivíduos são ocorrências únicas de algo existente na realidade. Por exemplo, cada um de nós é uma única ocorrência ou indivíduo de um "homo sapiens". Os universais ou tipos são entidades reais que generalizam os particulares existentes no mundo, por exemplo, "homo sapiens" é uma entidade geral ou universal referente aos particulares que cada um de nós é.

Um exemplo de ontologia é mostrado na Figura 3. Nesta ontologia, é descrita uma pequena fração do domínio médico obstétrico, especificamente para descrever a relação de lesão uterina com o corpo humano feminino. É possível identificar as entidades ou classes representadas por elipses e as relações pelas setas

Figura 3: Parte da ontologia Ontoneo.



Fonte: FARINELLI (2017).

Organizações orientada por dados

Segundo Tom Peters (2001 apud DAMA, 2012), "as organizações que não entenderem a enorme importância da gestão de dados e de informações como ativos tangíveis na nova economia, não sobreviverão." É dentro deste contexto que surge, nas organizações, a ideia da cultura de dados (*data culture*) ou cultura orientada a dados (*data-driven culture*).

A expressão "orientado por dados" ou orientado a dados" remete à ideia que a forma como uma atividade ou processo acontece depende dos dados que servem de entrada (*input*) para que ela ocorra, ou seja, a execução das atividades é orientada pelos dados que a alimentam. Assim, entende-se que uma empresa orientada a dados é aquela que utiliza sua base de dados e informações como insumo para suas decisões.

A orientação por dados remete ao desenvolvimento de ferramentas, habilidades e de uma cultura que atua nos dados (ANDERSON, 2015). Pensar em cultura de dados ou cultura por dados nas organizações não é apenas focar nas tecnologias de banco de dados ou nos profissionais que lidam com seus dados. Envolve acima de tudo uma visão holística dos dados dentro da empresa, desde sua infraestrutura à compreensão do papel dos dados para seu negócio, estratégias e atividades (PATIL; MASON, 2015).

Em suma, uma organização orientada a dados é uma empresa que se preocupa e desenvolve sua cultura por dados ou tem a cultura orientada por dados. A cultura orientada por dados nas organizações envolve a interação entre os dados que a empresa possui, as tecnologias usadas para compor a infraestrutura de dados e como as pessoas trabalham com os dados, seja para solução de problemas ou para obtenção de valor para o negócio da empresa (PATIL; MASON, 2015). Uma organização orientada a dados é uma empresa que desenvolve uma cultura orientada por dados, ou seja, é uma empresa onde as decisões são baseadas nos seus dados. As ações da empresa orientada a dados refletem o resultado das análises obtidas em seus dados.

A cultura orientada por dados nas organizações é um processo gradual que exige mudanças nos hábitos, atitudes, pessoal e até recursos. É preciso criar um entendimento de que os dados são fundamentais para cada atividade, cada processo e para cada decisão. Ou seja, os dados devem ser vistos como um ativo valioso da organização e devem ser tratados como tal durante todo seu ciclo de vida. Desenvolver a capacidade de analisar os dados disponíveis e como eles se relacionam com as mais diversas variáveis é primordial para uma organização que deseja tomar uma decisão orientada por dados (ANDERSON, 2015; GILES, 2013; PATIL; MASON, 2015).

Uma organização orientada a dados adquire, processa e aproveita os dados em tempo hábil para criar eficiência, definir estratégias e desenvolver novos produtos, navegando tranquilamente em seu ambiente competitivo (PATIL, 2011). Nesse sentido, a organização precisa coletar e adquirir os dados, estabelecer estratégias e atividades para limpeza e promoção da qualidade de dados, definir a melhor infraestrutura para armazenamento, acesso e disponibilização dos dados, além de trabalhar as estratégias e a infraestrutura para análise e visualização de dados. Alinda, envolvendo todos estes processos, a organização deve promover um programa de governança de dados a fim de maximizar a cultura de dados da organização.

Tomada de decisão baseada em dados (Data Driven)

Os dados permeiam nossas vidas, seja a vida de um ser humano ou de um ser organizacional. Conhecer esses dados nos ajuda a tomar decisões que impactam

a qualidade de nossas vidas. Ao tomar uma decisão, estamos escolhendo um caminho a seguir ou uma ação a executar, seja para enfrentar um problema ou para extrair vantagem de uma oportunidade. Tomar uma decisão baseada em dados pode otimizar as escolhas melhorando a qualidade de vida de um ser humano ou propiciando uma organização mais eficaz (DATNOW; PARK, 2014).

De acordo com Ashish Thusoo, uma cultura orientada por dados combina processos, pessoas e tecnologia, permitindo às empresas colocar os dados no âmbito da comunicação de seu dia a dia e substituindo o processo de tomada de decisão puramente intuitivo por intuição fundamentadas em dados (BRUNER, 2016). Para se manterem competitivas, as empresas devem parar de tomar decisões com base em intuitos ou instintos e cada vez mais, aplicar análises para obter informações úteis de seus dados (DAMA, 2017).

A tomada de decisão nas organizações envolve mais que intuição ou instintos dos gestores, abordando tecnologia, informações, estrutura organizacional, métodos e pessoal. A tomada de decisão requer quatro etapas (DAVENPORT, 2009):

1. Identificação das decisões que devem ser tomadas e priorização de quais são as mais importantes.

2. Examinar os fatores envolvidos nas decisões identificadas/priorizadas para entender quais decisões precisam ser melhoradas e quais processos podem torná-las mais efetivas. Tais fatores podem ser respostas às seguintes questões:

   – Quem desempenha o papel na decisão?

   – Com que frequência isso ocorre?

   – Quais informações estão disponíveis para apoiá-lo?

   – Quão bem a decisão normalmente é tomada?

3. Projetar os papéis, processos, sistemas e comportamentos que sua organização deveria usar para melhorar as decisões.

4. Institucionalizar a abordagem de uma melhor tomada de decisões fornecendo continuamente aos decisores ferramentas e assistência adequada para que eles possam tomar uma decisão.

Além dessas quatro etapas, deve-se avaliar a qualidade das decisões tomadas após o fato, buscando identificar resultados tanto comerciais e financeiros, mas também avaliar o processo de tomada de decisão e de qualidade das informações que o decisor se baseou (DAVENPORT, 2009).

Falar em tomada de decisões baseadas em dados (*Data-Driven Decision Making* - DDDM) significa adotar uma estratégia em que a análise de dados se torna o cerne dos processos de tomada de decisões na organização, em qualquer nível. Um ponto relevante que se deve ter em mente é que coletar dados por si só não traz o benefício esperado, os dados devem ter alta qualidade e deve-se prover acesso amplo aos dados ao longo de toda organização. Além disso, a organização deve ser capaz de usar dados de diferentes fontes, internas e externas à organização, além de ser capaz de misturar ou combinar esses dados de forma analítica. É preciso criar uma cultura orientada por dados, que permita a todos os colaboradores explorar informações para terem novas ideias.

O processo decisório baseado em dados é uma variação do método científico de pesquisa, pois lida com uma abordagem de levantamento de hipótese e resolução de problemas. Tal método envolve (PATIL; MASON, 2015):

- Coletar dados.
- Identificar hipóteses e/ou intuições sobre os dados, além de questões que esses dados podem responder.
- Formular questões que deseja obter resposta.
- Validar a(s) questões para verificar se é plausível de resposta por meio de seus dados.
- Definir uma massa para teste e realizar um teste para validar sua questão.
- Analisar os resultados para extrair informações sobre a questão.

Outra abordagem apresentada por Bambrick-Santoyo (2010), propõe um fluxo básico de DDDM, envolvendo: coleta e aquisição de dados; análise ou mineração de dados; consolidação e divulgação de dados e tomada de decisão baseada nas análises de dados realizadas.

Fato é que a tomada de decisão envolve atividades tanto de coleta de dados como de análise desses dados, sejam tais atividades manuais ou automáticas. Entretanto, Davenport (2009) aponta para a crescente adoção pelas empresas de sistemas automatizados nas análises de seus dados, a fim de incrementar sua capacidade analítica e consequentemente a tomada decisões, que em muitos casos são realizadas virtualmente e/ou em tempo real, como por exemplo, uma análise de crédito ou uma simulação de um seguro a partir de respostas inseridas em um formulário de adesão/simulação.

Em suma, o processo de tomada de decisão orientado por dados possui sete etapas, conforme ilustrado na Figura 4. Tomar decisões baseadas em dados requer, de forma essencial, o conhecimento do problema que precisa ser resolvido. A identificação do problema refere-se à compreensão do problema no qual uma decisão é envolvida. Ao nos referir a um problema, esse pode ser de fato um problema que necessita de uma solução, mas pode ser também uma inovação que a organização vem buscando. Identificar o problema contribui para um melhor direcionamento dos esforços envolvidos em todo o processo de tomada de decisão, pois torna mais claro quais os dados e informações necessárias no processo.

Figura 4: Etapas da tomada de decisões orientada por dados.

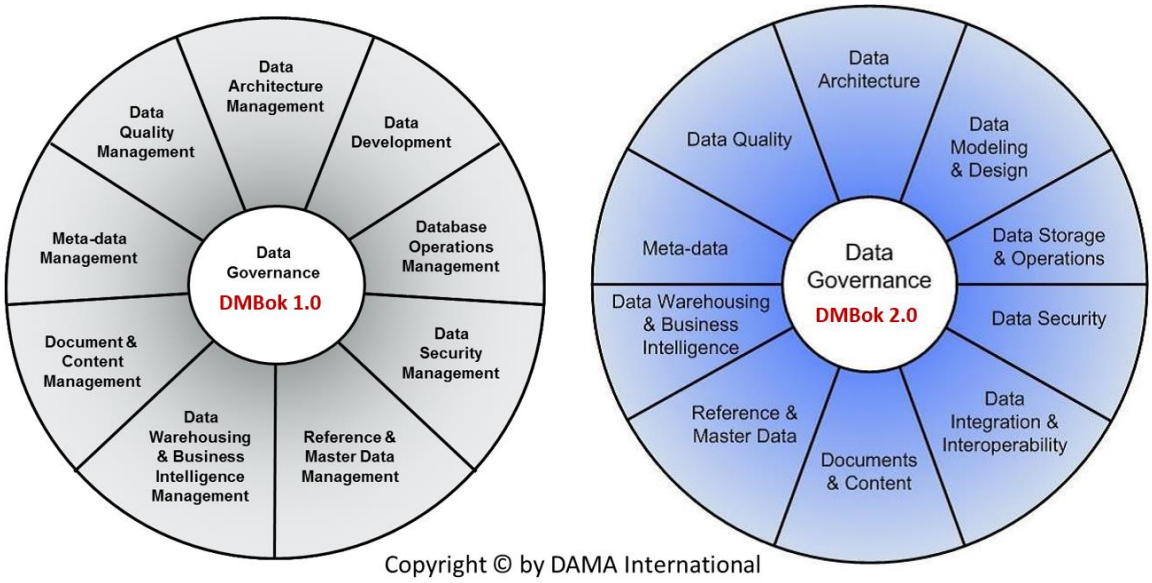
Após a identificação do problema, deve-se estabelecer quais indicadores o ajudaram a obter uma resposta ao problema, o que se deve mensurar para entender seu problema. Em seguida, deve-se determinar que variáveis podem ser usadas na análise, ou seja, os dados necessários para a análise e quais as fontes que fornecem esses dados. A próxima etapa é coletar os dados. Nesse momento, é importante observar questões como o formato e qualidade destes dados. A preparação dos dados consiste na organização dos dados que em geral são originados de múltiplas fontes, em um formato compatível para ser analisado. Finalmente, passa-se a etapa de análise dos dados e consequentemente a tomada de decisão. Durante a análise, pode-se usar diferentes elementos, desde uma análise manual à análise complexa, envolvendo sistemas de informação sofisticados. Essa etapa envolve decisões sobre a publicação e/ou visualização dos resultados da análise.

A tomada de decisões baseadas em dados é uma abordagem que estabelece um fluxo de coleta/aquisição, análise e comunicação de dados para suportar processos decisórios. Quando executado corretamente, o processo de tomada de decisão orientado por dados ajuda as organizações a alcançar melhores resultados. Em contrapartida, a qualidade dos dados é um fator relevante para se considerar, pois uma análise sob dados errados ou de baixa qualidade pode implicar em decisões errôneas, acarretando em diversos problemas às organizações.

DAMA-DMBOK

O DAMA-DMBOK é um “corpo de conhecimento” sobre Gestão de Dados, escrito com a participação de mais de 100 profissionais de dados espalhados ao redor do mundo todo. É uma referência sobre gestão de dados reconhecida e utilizada no exterior e vem crescendo bastante no Brasil. Este guia estabelece as principais diretrizes para tratar os dados dentro das organizações, estabelecendo funções (DMBOK versão 1.0) ou áreas de conhecimento (DMBOK versão 2.0) conforme ilustrado na Figura 5. Neste material, adotei a terminologia “área de conhecimento” para ambas as versões apenas para padronizar e tornar a leitura mais homogênea.

Figura 5 – Áreas de conhecimento do DAMA-DMBOK versões 1 e 2.



Fonte: DAMA, 2012; 2017.

Especificamente sobre a área de conhecimento Arquitetura de Dados, de acordo com o DMBOK, os objetivos da Arquitetura de Dados devem englobar visões ou modelos de dados por áreas ou assuntos, com foco na visão corporativa e de aplicações. Ela deverá dar suporte ao desenvolvimento e manutenção do modelo de dados corporativo, procurando manter a coerência do significado dos dados para a empresa. Dentro da perspectiva do gerenciamento de dados, a área de conhecimento arquitetural de dados é fundamental, pois ajuda as organizações a compreenderem melhor seus ativos de dados e representá-los conforme os diferentes níveis de abstração (DAMA, 2012; 2017).

Os artefatos de arquitetura de dados incluem especificações usadas para descrever o estado existente, definir requisitos de dados, orientar a integração de dados e controlar os ativos de dados durante todo seu ciclo de vida. Deve ainda incluir os padrões que regem como os dados são coletados, armazenados, documentados, organizados, usados e removidos.

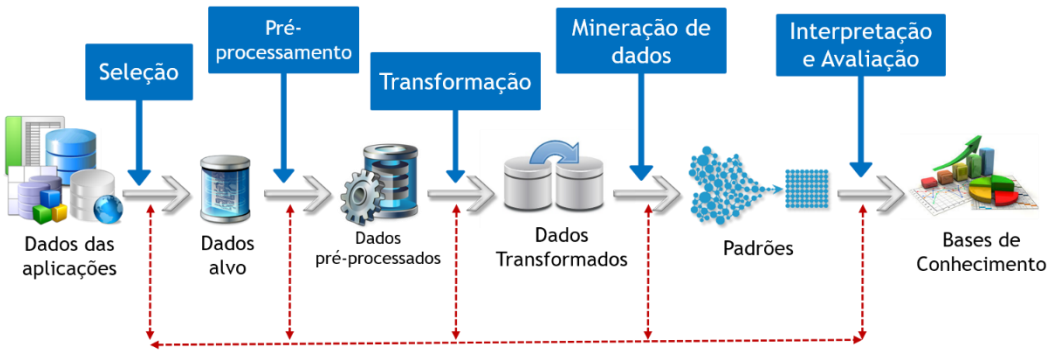
Metodologias de mineração de dados

Em linhas gerais, mineração de dados (data mining) é o processo de extrair conhecimento a partir de grandes volumes de dados. O crescimento rápido do volume e da dimensão das bases de dados criou a necessidade e a oportunidade de se extrair sistematicamente o conhecimento nelas contido e de se produzir novos conhecimentos. Ao longo das últimas 3 décadas, surgiram várias abordagens para mineração de dados, dentre elas, cita-se as metodologias KDD, CRISP-DM e SEMMA.

KDD - Knowledge Discovery in Database

O processo de Descoberta de Conhecimento em Banco de Dados (*Knowledge Discovery in Database* - KDD) é um processo que envolve a busca e a interpretação de padrões em dados através da utilização de algoritmos e análise dos resultados obtidos. O KDD pode ser definido como um processo de extração de conhecimentos válidos, novos, potencialmente úteis e compreensíveis, visando melhorar o entendimento de um problema. É um processo iterativo e interativo, com etapas que partem da preparação de dados, a procura por padrões entre os dados, a avaliação dos resultados e refinamento.

Figura 6: Etapas do processo KDD.



Fonte: tradução adaptada de (FAYYAD *et al.* 1996a).

A Figura 6 demonstra não só a sequência de etapas do processo do KDD, como pode ser observada também a iteração entre as fases pelas setas tracejadas.

Considera-se como um processo interativo dada a participação dos usuários em todas as etapas, tanto nas tarefas como nas decisões. O fato das etapas estarem

conectadas, torna o processo iterativo, além disso, na etapa de avaliação o resultado pode não ser satisfatório, assim, pode iníciar uma nova instância de processo. A Figura 6 apresenta as etapas do processo KDD (FAYYAD *et al.*, 1996a).

No processo de KDD, cada fase possui uma intersecção com as demais. Desse modo, os resultados produzidos numa fase são utilizados para melhorar os resultados das próximas fases. O KDD engloba, portanto, as etapas que produzem conhecimentos a partir de dados relacionados. Sua principal característica é a extração não-trivial de informações e conhecimentos implicitamente contidos em uma base de dados. Em suma, o KDD é capaz de detectar dados e informações armazenadas nas bases de dados, transformando-as em conhecimento.

Etapas da descoberta do conhecimento

As fases ou etapas do KDD, conforme ilustrado na Figura 6, são: seleção dos dados; pré-processamento (limpeza e integração dos dados); transformação dos dados; mineração dos dados; interpretação e avaliação dos padrões (apresentação e assimilação do conhecimento).

Conforme descrito por FAYYAD *et al.* (1996a), o processo KDD inicia-se com o entendimento dos desejos do usuário, buscando conhecer as necessidades tanto do negócio quanto do usuário. Uma vez definidas as necessidades, define-se o objetivo do KDD, ou seja, que conhecimento é necessário descobrir. Para isso, deve-se realizar o levantamento dos requisitos de negócio, que mais tarde guiará a escolha da técnica de mineração de dados a ser aplicada.

O KDD é composto por um conjunto de etapas que, em geral, podem ser reunidas em três fases: preparação, análise e interpretação. As três fases iniciais, que envolvem a seleção, o pré-processamento e a transformação, são também chamadas de preparação dos dados e exigem bastante tempo, sendo que a maior parte desse tempo é consumida com a limpeza dos dados.

Seleção de dados:

Finalizado a identificação dos requisitos de análise que se deseja realizar, parte-se para a obtenção do conjunto de dados que será trabalhado durante todo o processo KDD. Tal conjunto precisa conter as informações necessárias para sejam

submetidas aos algoritmos de mineração, e assim, alcançar o objetivo da análise. Esta etapa de obtenção de dados é chamada de “Seleção”.

A seleção de dados consiste na criação de um conjunto de dados-alvo ou dados selecionados. Nesta fase, escolhem-se apenas atributos relevantes do conjunto de dados no qual vão participar da descoberta que será efetuada.

Em suma, a seleção de atributos consiste na escolha de um subconjunto de atributos relevantes para o objetivo da tarefa. O subconjunto selecionado é então fornecido para o algoritmo de mineração dos dados. Uma motivação para essa seleção é otimizar o tempo de processamento do algoritmo minerador, visto que ele apenas trabalhará com um subconjunto de atributos, desse modo diminuindo o seu espaço de busca.

Pré-processamento:

Esta é uma parte crucial no processo, pois a qualidade dos dados vai determinar a eficiência dos algoritmos de mineração. As grandes bases de dados são altamente susceptíveis a ruídos, valores faltantes e inconsistentes. Dados limpos e compreensíveis são requisitos básicos para o sucesso da mineração dos dados. Dessa forma, o objetivo do pré-processamento é preparar os dados em um formato acessível aos algoritmos da etapa de análise.

No pré-processamento dos dados, pode-se realizar duas tarefas com o objetivo de melhorar a qualidade dos dados da análise: limpeza dos dados e enriquecimento dos dados.

- Limpeza dos dados: Compreende os tratamentos realizados sobre os dados coletados, visando a garantir sua a qualidade, ou seja, sua completude, veracidade e integridade. Consiste em resolver problema com os dados pré-existentes, como a retirada de dados duplicados, correção de dados corrompidos ou inconsistentes, tratamento de valores ausentes.

- Enriquecimento dos dados: consiste em agregar aos dados existentes mais detalhes de modo que essas possam contribuir no processo de descoberta de

conhecimento. Geralmente esses detalhes são dados externos, ou seja, que não estão na base de dados, porém são conhecidos pelos analistas de dados.

Com as aplicação destas técnicas de pré-processamento conseguimos remover os ruídos dos dados, são erros e observações fora do padrão, definimos estratégias para tratar os valores ausentes, e encontramos os melhores formatos para os dados conforme estabele os requisitos dos algoritmos e ferramenta de mineração de dados.

Nesta etapa, deverão ser realizadas tarefas que eliminem dados redundantes e inconsistentes, recuperem dados incompletos e avaliem possíveis dados discrepantes ao conjunto (outliers). Mais uma vez, o auxílio do especialista do domínio é fundamental.

Transformação dos dados:

Após serem selecionados, limpos e pré-processados, os dados necessitam ser armazenados e formatados adequadamente para que os algoritmos de aprendizado possam ser aplicados. Geralmente, os algoritmos utilizados na mineração de dados requerem que os dados se apresentem em um formato apropriado, existindo assim a necessidade da aplicação de operações de transformação desses dados na fase de transformação.

Assim, essa etapa consiste na transformação dos dados brutos em dados transformados para aplicação da técnica inteligente. Essa fase depende do algoritmo a ser aplicado na fase seguinte. Na transformação de dados, pode-se utilizar operações do tipo normalização de dados, conversões de valores simbólicos para valores numéricos, discretização e composição de atributos.

Segundo HAN *et al.* (2011), a normalização consiste em converter os valores de atributos para faixas de -1 a 1 ou de 0 a 1, sendo de grande utilidade para algoritmos de classificação como redes neurais ou baseados em distância, tais como, *nearest-neighbor* e de *clustering*.

No que tange a conversões de valores simbólicos para valores numéricos, a técnica usada depende da capacidade e necessidade da técnica de mineração

utilizada, já que algumas técnicas como árvores de decisão, podem manipular valores simbólicos, enquanto outras como redes neurais, podem manipular somente uma representação numérica de um valor simbólico. Devido ao fato de todas as técnicas de mineração de dados poderem manipular dados numéricos, mas algumas não poderem manipular dados simbólicos, torna-se preciso aplicar algum método de transformação de valores simbólicos em uma representação numérica apropriada (PYLE, 1999).

Por fim, a discretização consiste em transformar valores contínuos de atributos em valores simbólicos (discretos). Assim como também equivale à formação de categorias de valores que já sejam discretos, mas que se apresentam em grande número. De um modo geral, os métodos de discretização existentes podem ser classificados sob diferentes abordagens descritas a seguir por HUSSAIN *et al.* (1999).

A composição ou combinação de atributos consiste em construir um pequeno conjunto de novos atributos a partir de atributos originais, de forma que os resultados construídos usando os novos atributos apresentem maior acurácia e concisão que aqueles criados diretamente usando os atributos originais (BARANAUSKAS, 2001, ZHENG, 1996).

A etapa de transformação de dados é potencialmente a tarefa que requer grande habilidade no processo de KDD. Tipicamente, essa etapa exige a experiência do analista de dados e seu conhecimento nos dados em questão. Embora o processo de KDD possa ser executado sem essa fase, nota-se que quando efetivada, os resultados obtidos são mais intuitivos e valiosos, além de que, na maioria das vezes, facilitam a construção do modelo.

São vantagens à realização da transformação de dados:

- Melhora a compreensão do conhecimento descoberto.
- Reduz o tempo de processamento para o algoritmo minerador.
- Facilita a construção e execução do algoritmo a tomar decisões globais, pois os valores dos atributos foram englobados em faixas.

Como desvantagem, cita-se a redução da medida de qualidade de um conhecimento descoberto, perdendo-se, assim, detalhes relevantes sobre as informações extraídas.

A identificação de padrões consistem em ajustar os modelo de análise aos dados ou identificar a própria estrutura do conjunto de dados. Os padrões estão escondidos nos dados e precisam ser novos para o sistema, preferencialmente para o usuário, devem ser válidos em relação aos dados já armazenados e às regras do negócio, além de ser útil para sua aplicação nas tarefas para o qual foi demandado.

De uma maneira geral, a complexidade do processo de KDD encontrar-se na dificuldade em perceber, compreender e interpretar adequadamente os diversos fatos observáveis durante o processo de descoberta, além da dificuldade em conjugar tais interpretações dinamicamente, de forma a deliberar quais as ações carecem de ser efetivadas em cada caso.

Mineração de dados:

Após o pré-processamento e a transformação dos dados, inicia-se a etapa mineração de dados para alcançar os objetivos definidos inicialmente. Devido à sua importância e ao nível de detalhamento necessário, essa etapa foi tratada na seção seguinte.

Interpretação/Avaliação do Conhecimento:

A etapa final também é conhecida como pós-processamento. Consiste em interpretar e avaliar os padrões identificados pela etapa de mineração de dados. Essa é mais uma fase que deve ser feita em conjunto com um ou mais especialistas no assunto. O conhecimento adquirido através da técnica de data mining deve ser interpretado e avaliado para que o objetivo final seja alcançado.

Uma maneira genérica de obter a compreensão e interpretação dos resultados é utilizar técnicas de visualização. As técnicas de visualização estimulam tanto a percepção quanto a inteligência humana, de modo a incrementar a capacidade de compreensão e a associação dos novos padrões (GOLDSCHMIDT; PASSOS, 2005). Algumas técnicas de visualização de dados comumente conhecidas são os

dashboards e as narrativas de dados (*data storytelling*) (FEW; EDGE; 2007. KNAFLIC, 2015).

Dessa forma, essa etapa inclui a visualização dos padrões extraídos dos dados ou dos modelos capazes de resumir tanta a estrutura quanto as informações existentes nos dados. Em geral, a principal meta dessa fase é melhorar a compreensão do conhecimento descoberto pelo algoritmo minerador, validando-o através de medidas da qualidade da solução e da percepção de um analista de dados. Assim, além da visualização, utilizamos medidas técnicas e subjetivas para ponderar sobre os padrões obtidos. Tais medidas técnicas são informações sobre à precisão, erro médio, erro quadrático e às taxas de falsos positivos e negativos. Já as medidas subjetivas referem-se a informações como utilidade, entendimento ou complexidade.

Muitas vezes, o resultado obtido nessa etapa não é satisfatório, nesse caso, pode-se retornar a qualquer um dos estágios anteriores ou até mesmo recomeçar todo o processo. Quando isso ocorre, as duas das ações mais comuns são: 1) modificar o conjunto de dados inicial e/ou 2) trocar a técnica e, consequentemente, o algoritmo de mineração de dados ou ao menos alterar seus parâmetros de configurações.

CRISP-DM

O CRISP-DM (Cross-Industry Standard Processo of Data Mining) é uma técnica para mineração de dados, uma das mais utilizadas, composta por seis fases dispostas de maneira cíclica, conforme mostra a Figura 7 e descritas a seguir. Além disso, apesar de ser composto por fases, o fluxo não é unidirecional, podendo ir e voltar entre as fases (CHAPMAN *et al*. 2000).

- Entendimento do negócio: essa fase inicial se concentra no entendimento dos objetivos do projeto e requisitos sob a perspectiva de negócio, transformando então esse conhecimento em uma mineração de dados. A definição de um problema e um plano preliminar concebido para atingir os objetivos. O entendimento sobre o negócio irá ajudar nas próximas etapas.

- Entendimento dos dados: a fase de entendimento dos dados consiste em uma primeira análise para se familiarizar com os dados, identificar as fontes de

dados, identificar os dados relevantes para o problema em questão e identificar problemas de qualidade de dados. O foco é descobrir as primeiras ideias sobre os dados ou detectar subconjuntos interessantes para formar hipóteses para informações ocultas.

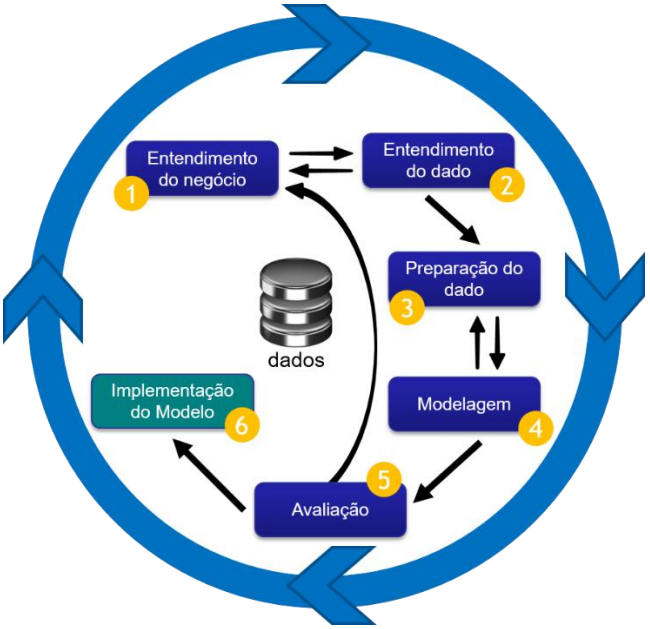
- Preparação dos dados: abrange todas as atividades para a construção do conjunto de dados final a partir dos dados brutos iniciais. Envolve tarefas como seleção de dados que serão usados no modelo, além de limpeza, combinação e integração dos dados.

- Modelagem: nesta fase, são selecionadas e aplicadas várias técnicas de modelização que sejam mais aderentes ao objetivo do projeto, seja ele uma predição, classificação, agrupamento ou regressão.

- Avaliação: Por fim, nesta fase, o modelo de análise obtido é submetido a testes e validações, visando obter a confiabilidade nos modelos.

- Implementação do modelo: por fim, o modelo de análise é entregue/implantado como um relatório ou como outro mecanismo de visualização, de forma que o resultado das análises ou mineração possa ser usado.

Figura 7: Esquema do processo CRISP-DM.

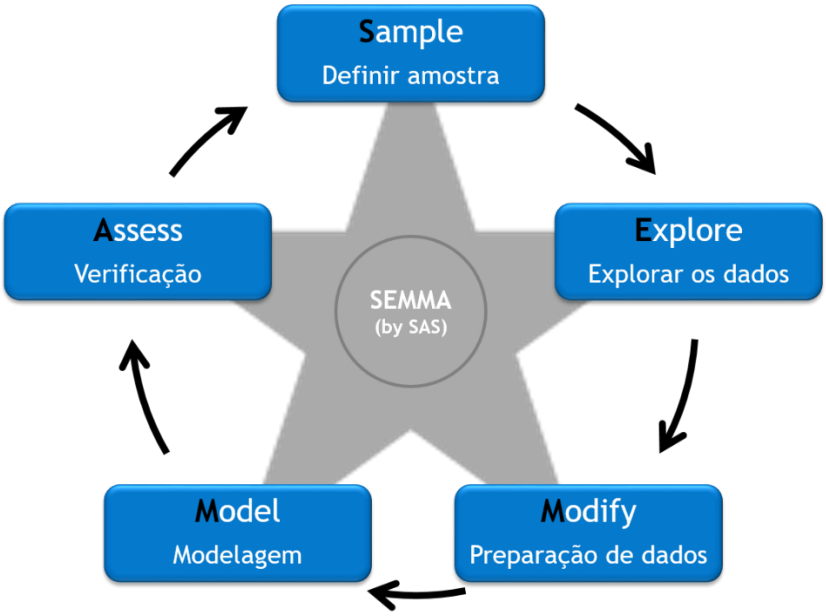


SEMMA

O acrónimo SEMMA significa *Sample, Explore, Modify, Model, Assess*. Trata-se de um processo desenvolvido pelo SAS Institute que considera um ciclo com 5 etapas para o processo de mineração de dados.

- Explorar os dados (Explore): consiste na exploração dos dados através da procura de padrões e tendências imprevistas ou anomalias, buscando compreensão sobre o objeto em análise.
- Preparação de dados (Modify): essa etapa consiste na transformação dos dados para se adequar ao modelo de análise.
- Modelagem (Model): é a fase de definição do modelo de análise, similar à CRISP-DM.
- Verificação (Assess): similar à fase de avaliação do CRISP-DM, consiste em avaliar os dados através da avaliação do modelo.
- Figura 8 foca principalmente nas tarefas de criação do modelo, deixando as questões de negócio de fora.
- Definir amostra (Sample): consiste na amostragem dos dados extraindo uma porção de um conjunto de dados grande suficiente para conter as informações significativas, mas pequeno o suficiente para manipular rapidamente.
- Explorar os dados (Explore): consiste na exploração dos dados através da procura de padrões e tendências imprevistas ou anomalias, buscando compreensão sobre o objeto em análise.
- Preparação de dados (Modify): essa etapa consiste na transformação dos dados para se adequar ao modelo de análise.
- Modelagem (Model): é a fase de definição do modelo de análise, similar à CRISP-DM.
- Verificação (Assess): similar à fase de avliação do CRISP-DM, consiste em avaliar os dados através da avaliação do modelo.

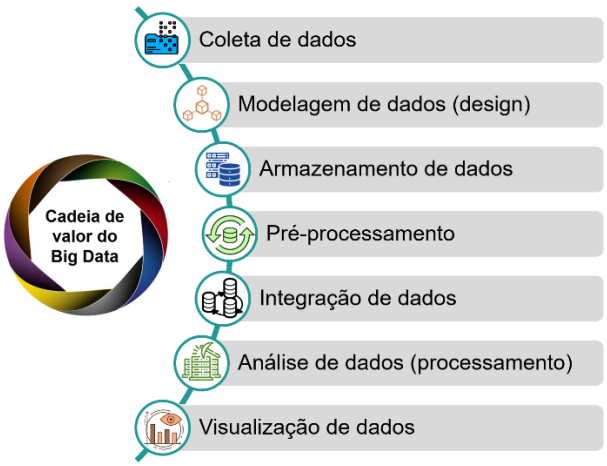
Figura 8: Esquema do processo SEMMA.



Cadeia de valor do *Big Data*

Dados tradicionais e Big Data demandam processos de coleta, armazenamento, processamento, análise e visualização. Porém a diferença se volta para as três características: volume, variedade e velocidade (TAURION, 2013). Ao longo das últimas três décadas, surgiram várias abordagens para mineração de dados, dentre elas, as metodologias KDD, CRISP-DM e SEMMA (AZEVEDO; SANTOS, 2008; SHAFIQUE; QAISER, 2014). O pipeline de dados é o conjunto de processos e atividades que uma organização executa para extrair informações dos seus dados capazes de subsidiar seu processo de tomada de decisão. Um exemplo de pipeline fundamentado nos processos de mineração de dados KDD, SEMMA e CRISP-DM, além da cadeia de valor proposta por CURRY (2016), aqui nomeado de cadeia de valor do *Big Data,* é apresentada na Figura 9. Nesta cadeia de valor, o fluxo de informações é descrito como uma série de etapas necessárias para gerar valor e informações úteis dos dados.

Figura 9: Cadeia de Valor de Big Data.



Capítulo 2. Tecnologias aplicadas à análise de dados

Neste capítulo, apresentamos algumas tecnologias que são encontradas no dia a dia dos analistas de dados.

Fundamentos de banco de dados

Os sistemas de informação não são nada sem os dados, e estes precisam ser armazenados em algum repositório. O conceito de armazenamento de dados é muitas vezes relacionado à persistência de dados, ou seja, o dado deve ser persistido ou armazenado em local não volátil, de forma que eles possam ser recuperados posteriormente para consulta e análise. Isso significa armazenar esses dados em um local que possa garantir a integridade dos dados por um período indeterminado, até que eles sejam atualizados ou descartados propositalmente. O conjunto de dados armazenados é conhecido como Banco de Dados. Um banco de dados é “uma coleção de dados inter-relacionados, representando informações sobre um domínio específico”. Um sistema de banco de dados apresenta o conjunto de quatro componentes básicos: dados, hardware, software e usuários (ELMASRI; NAVATHE, 2005; SILBERSCHATZ; KORTH; SUDARSHAN, 2012).

Para suportar as necessidades dos sistemas de bancos de dados, foram criados os Sistema Gerenciadores de Banco de Dados (SGBD ou em inglês Data Base Management System - DBMS). SGBDs são sistemas ou softwares utilizados para gerir os bancos de dados, permitindo: i) criar, modificar e eliminar bases de dados; ii) realizar as operações básicas com os dados (inserir, alterar, excluir e consultar); iii) garantir a segurança de acesso aos dados; iv) garantir a integridade de dados, controle de concorrência e possibilidades de recuperação e tolerância a falhas (SILBERSCHATZ; KORTH; SUDARSHAN, 2012).

Os objetivos de um sistema de banco de dados são o de isolar o usuário dos detalhes internos do banco de dados (promover a abstração de dados) e promover a independência dos dados em relação às aplicações, ou seja, tornar independente da aplicação, da estratégia de acesso e da forma de armazenamento. Existem diversos paradigmas tecnológicos de SGBDs, como por exemplo, os SGBDs em rede,

hierárquicos, relacionais, NoSQL. Vamos tratar dos SGBDs relacionais e NoSQL nos capítulos seguintes.

APIs de coleta de dados

O acrônimo API corresponde às palavras em inglês "*Application Programming Interface*", em português "Interface de Programação de Aplicações". Uma API é um middleware ou software intermediário que permite que dois aplicativos se comuniquem. Quando você usa um aplicativo de mídia social como o Facebook ou Twitter, ou envia uma mensagem instantânea ou verifica o clima em app no seu telefone, você está usando uma API.

Uma API é composta por um conjunto de rotinas (programas) que são responsáveis por realizar várias operações previamente conhecidas e divulgadas pelo fornecedor da própria API. Por meio das APIs, é possível utilizar suas funcionalidades seguindo os protocolos previamente definidos. As APIs permitem que os desenvolvedores economizem tempo, aproveitando a implementação de uma plataforma para realizar o trabalho. Isso ajuda a reduzir a quantidade de código que os desenvolvedores precisam criar e garante maior consistência entre aplicativos para a mesma plataforma.

Visão geral ferramentas de análise de dados

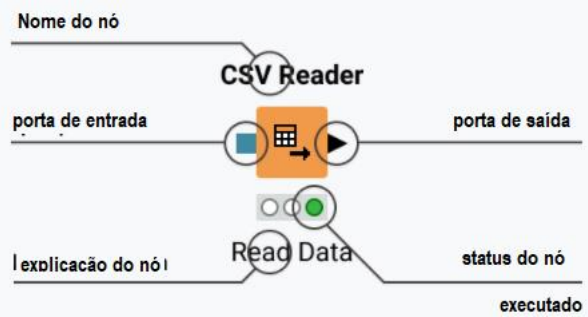
Podemos utilizar diversas ferramentas para analisar dados, a seguir apresentamos algumas que serão trabalhadas neste curso.

Plataforma Knime Analytics

A plataforma KNIME Analytics é um software de código aberto muito potente utilizado para construir fluxos de trabalho (workflows) capazes de coletar, tratar, analisar e exibir seus resultados. O Knime (a letra K é silenciosa na pronuncia) possui muitas integrações com outros produtos, como bancos de dados relacionais e NoSQL, linguagens como R e Python, ferramentas como Weka, APIs como do Twitter e Google.

O Knime trabalha com o conceito de nós (Figura 10) onde cada nó realiza uma ou mais tarefas, como leitura e gravação de arquivos e bancos de dados; preparação e transformação de dados; execução de modelos preditivos; análises estatísticas; criação de gráficos, dentre outros.

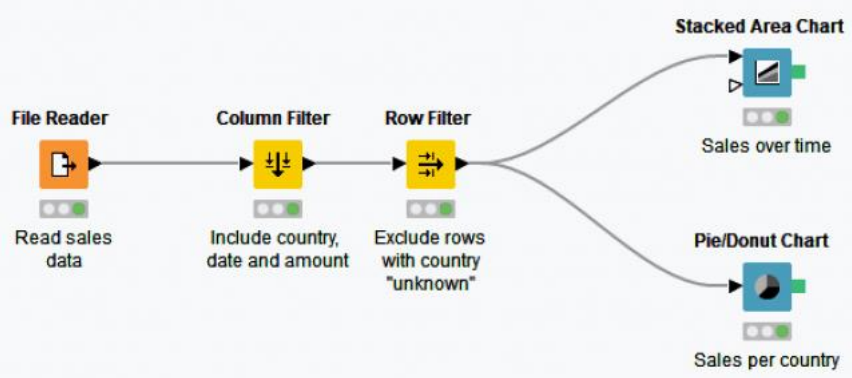
Figura 10: Exemplo de nó.



Fonte: Traduzido de https://www.knime.com/getting-started-guide.

Os fluxos de trabalho (Figura 11) são construídos ao se unirem os nós conectando as portas de saída de um nó, à porta de entrada de outro nó.

Figura 11: Exemplo de workflow.



Fonte: https://www.knime.com/getting-started-guide.

Para download do Knime acesse https://www.knime.com, além deste link, recomenda-se a leitura do guia introdutório disponível em: https://www.knime.com/getting-started-guide.

Linguagem R

R é uma linguagem de programação e ambiente de software para análise estatística, representação gráfica e relatórios. Essa linguagem foi criada por Ross Ihaka e Robert Gentleman na University of Auckland, Nova Zelândia, e foi chamada de R, com base na primeira letra do nome dos dois autores. O R está disponível gratuitamente sob a GNU (*General Public License*), para vários sistemas operacionais como Linux, Windows e Mac.

O R fornece uma ampla variedade de técnicas estatísticas, como modelagem linear e não linear, testes estatísticos clássicos, análise de série temporal, classificação, agrupamento, entre outras, além de técnicas gráficas. Uma das vantagens fortes do R é a facilidade com que gráficos de qualidade podem ser produzidos, incluindo símbolos matemáticos e fórmulas quando necessário.

Para obter o R para instalação, acesse o site do projeto *The Comprehensive R Archive Network*, disponível em: https://cran.r-project.org/. Nesse site, além do R para download, estão disponíveis materiais de apoio, como manuais e códigos, inclusive em Português.

Para desenvolver utilizando a linguagem R é recomendado utilizar uma aplicação IDE (*Integrated Development Environment*). Neste curso, recomenda-se o uso da ferramenta RStudio, disponível para download em: www.rstudio.com. Como alternativa, podemos utilizar o Google Colaboratory com ambiente R, disponível em: https://colab.research.google.com/#create=true&language=r.

Ferramenta Weka

O WEKA (Acrônimo para *Waikato Environment for Knowledge Analysis*) é um software livre com licença GNU General Public License (GPL), desenvolvido pela Universidade de Waikato na Nova Zelândia para utilização em tarefas de Mineração de Dados. O software foi escrito na linguagem Java™ e contém uma GUI para interagir com arquivos de dados e produzir resultados visuais (pense em tabelas e curvas). Ele também tem uma API geral, assim é possível incorporar o WEKA, como

qualquer outra biblioteca, a seus próprios aplicativos para fazer coisas como tarefas de mineração de dados automatizadas ao lado do servidor (FRANK *et al.* 2016; PIMENTA *et al.* 2009).

O WEKA contém uma série de algoritmos que são desenvolvidos pela comunidade que contribuí com a ampliação do Software, já que é desenvolvido em Java e o projeto é código aberto, o que significa que dia após dia o projeto aumenta cada vez mais uma vez que não há restrições de bibliotecas, bem como não há nenhum tipo de corporação por trás de uma iniciativa exclusivamente acadêmica. O sistema possui uma interface gráfica amigável e seus algoritmos fornecem relatórios com dados analíticos e estatísticos do domínio minerado.

Além disso, ele conta com uma grande flexibilidade na utilização de suas técnicas de mineração, nas quais há uma ampla variedade de algoritmos que contêm a sua respectiva descrição. De acordo com o conhecimento do analista, pode representar um diferencial de acordo com a escolha do algoritmo para a base que será analisada, na qual uma representação de um algoritmo pode ter um resultado distinto de acordo com a técnica escolhida.

Capítulo 3. Análise Exploratória de dados

Após aquisição de dados, é recomendável realizar a Análise Exploratória De Dados (AED). Seu intuito é analisar os dados antecipadamente, buscando conhecê-los, antes da aplicação de outras técnicas estatísticas (DATA SCIENCE GUIDE, 2021). Assim, o analista obtém uma compreensão geral de seus dados e das possíveis relações existentes entre as variáveis em análise. Em suma, a AED consiste em sumarizar e organizar os dados coletados por meio de tabelas, gráficos ou medidas numéricas, e a partir desta sumarização/organização procura por alguma regularidade ou padrão nas observações, ou seja, faz a interpretação dos dados [1](HECKERT,FILLIBEN, 2003; TUKEY, 1977).

População e amostra

Entende-se como população o conjunto dos elementos que representam pelo menos uma característica comum, no qual deseja-se analisar o comportamento. Ou seja, a população é o conjunto global sobre o qual se deseja chegar a conclusões. A amostra refere-se ao subconjunto finito de uma população sobre o qual são feitas observações. A amostra é qualquer conjunto de elementos retirado da população, não vazio e que tenha um menor número de elementos que a população. Uma amostra tem que ser representativa em relação à população para que os resultados não sejam distorcidos (MEDRI, 2011).

Variável

Entende-se como variável qualquer característica de interesse associada aos elementos de uma população. Essas variáveis podem ser (MEDRI, 2011):

Variáveis qualitativas: variáveis que assumem valores categóricos, de classes ou de rótulos, ou seja, por natureza, dados não numéricos. Essas variáveis denotam características individuais das unidades sob análise, tais como sexo, estado civil, naturalidade, raça, grau de instrução, dentre outras, permitindo estratificar as unidades para serem analisadas de acordo com outras variáveis. Podem ser

[1] Para detalhes, acesse: http://leg.ufpr.br/~fernandomayer/aulas/ce001n-2016-01/02_Analise_Exploratoria_de_Dados.html.

subdivididas como nominal, em que as categorias não possuem uma ordem natural (por exemplo: nomes, cores, sexo) ou como ordinal, em que as categorias podem ser ordenadas (por exemplo: classe social, grau de instrução, estado civil) (MEDRI, 2011).

Variáveis quantitativas: variáveis que assumem valores numéricos, intervalar ou de razão, por exemplo: idade, salário, peso etc. As variáveis quantitativas podem ser classificadas como discretas, quando assumem um número finito de valores, em geral, valores inteiros (por exemplo: número de irmãos, número de passageiros) ou contínuas, quando assume um número infinito de valores dos números reais, geralmente em intervalos (por exemplo: peso, altura, pressão) (MEDRI, 2011).

Medidas

A análise exploratória de dados consiste em um conjunto de cálculos de medidas estatísticas que visam resumir as características dos dados. Dentre as medidas estatísticas as mais utilizadas são as medidas de posição central (de tendência central), de posição, de dispersão e de assimetria (MEDRI, 2011).

Medidas de Posição Central: Representam os fenômenos pelos seus valores médios, em torno dos quais tendem a concentrar-se os dados. Dentre todas as medidas de tendência central, temos: Média, Mediana e Moda.

- *Moda:* é o valor (ou atributo) que ocorre com maior frequência.
- *Média:* soma de todos os valores da variável dividida pelo número de observações.
- *Mediana*: valor que deixa 50% das observações à sua esquerda

Medida de Posição: São medidas que dividem a área de uma distribuição de frequências em regiões de áreas iguais. As principais medidas de posição são: Quartil, Percentil, Mínimo, Máximo.

- *Máximo (max)* e *Mínimo (min):* a maior e a menor observação de valor dos dados.

- *Quartis:* divide um conjunto de valores dispostos em forma crescente em quatro partes. Primeiro Quartil (Q1): valor que deixa 25% das observações à sua esquerda. Terceiro Quartil (Q3): valor que deixa 75% das observações à sua esquerda.

- *Decis:* divide um conjunto de valores dispostos em forma crescente em dez partes, sendo cada parte correspondente a 10% dos dados.

- *Percentil:* divide um conjunto de valores dispostos em forma crescente em cem partes, sendo cada parte correspondente a 1% dos dados.

Medidas de Dispersão: É um valor que busca quantificar o quanto os valores da amostra estão afastados ou dispersos relativos à média amostral. A dispersão é a variabilidade de observações que os dados possuem entre si, ou seja, se todos os valores forem iguais, não existe dispersão, agora, se os dados não são iguais, existe dispersão entre eles. As medidas utilizadas para representar dispersão são:

- *Amplitude:* diferença entre o valor máximo e o valor mínimo.

- *Intervalo-Interquartil:* é a diferença entre o terceiro quartil e o primeiro quartil, ou seja, Q3 - Q1.

- *Variância:* média dos quadrados dos desvios em relação à média aritmética.

- *Desvio Padrão:* mede a variabilidade independentemente do número de observações e com a mesma unidade de medida da média.

- *Coeficiente de Variação:* mede a variabilidade em uma escala percentual independente da unidade de medida ou da ordem de grandeza da variável.

Medidas de Assimetria e Curtose: As medidas de assimetria possibilitam analisar uma distribuição de acordo com as relações entre suas medidas de moda, média e mediana, quando observadas graficamente ou analisando apenas os valores. Ou seja, uma distribuição é considerada simétrica quando apresenta o mesmo valor para a moda, a média e a mediana. Da mesma forma, é considerada assimétrica quando essa igualdade de medidas não ocorre. Curtose é o grau de achatamento da distribuição dos dados em relação a uma distribuição padrão, chamada de curva

normal. Ou seja, o quanto uma curva de frequência será achatada em relação a uma curva normal de referência.

Análises de variáveis

Ao realizar a análise de variáveis com o intuito de tentar compreender e resumir seus dados, pode seguir três linhas: análise univariada, bivariada e multivariada, assim chamadas considerando a quantidade de variáveis envolvidas.

Na análise univariada, o objetivo é descrever a população examinando uma única variável a cada vez. A análise bivariada busca identificar relações entre duas variáveis, como uma variável influencia o comportamento da outra. Por fim, na análise multivariada, a abordagem empregada busca compreender o comportamento resultante da multiplicidade das variáveis.

Em geral, na univariada e bivariada, busca-se identificar a frequência de ocorrências para cada variável ou combinação de variáveis, identificando a moda da variável. Outra preocupação é identificar as medidas de resumo em cada variável quantitativa. O tipo de abordagem utilizada na análise bivariada depende dos tipos das variáveis envolvidas. Existem três possibilidades: qualitativa versus qualitativa, qualitativa versus quantitativa e quantitativa versus quantitativa.

Capítulo 4. Fundamentos de Análise de dados

Análise de dados nada mais é do que um conjunto de técnicas empregadas para a transformação de dados e informações em conhecimento para um propósito específico. A análise de dados visa encontrar uma forma eficiente de conhecimento (padrões) em conjuntos de dados, seja para compreender o comportamento de pessoas ou para identificar novas oportunidades de negócio.

Inicialmente, com o surgimento dos DWs, eram empregadas técnicas de análise a fim de compreender o que aconteceu e o motivo pelo qual eventos aconteceram. Mais tarde, isso já não era suficiente, pois surgiu a necessidade de tentar prever o que poderia acontecer com os negócios antes de acontecer e, assim, antecipar algumas ações. Como o volume de dados ultrapassava a capacidade humana de interpretar e compreender tanta informação, foi necessário criar mecanismos automáticos para processar tantos dados.

O crescimento rápido do volume e da dimensão das bases de dados criou a necessidade e a oportunidade de extrair sistematicamente o conhecimento nelas contido e de produzir novos conhecimentos.

Introdução à análise de dados

A análise de dados definitivamente já não é novidade no mundo dos negócios. A ideia de transformar dados em conhecimento remete ao surgimento dos *data warehouses*. A partir do conhecimento apurado, é possível definir os melhores caminhos a seguir, que apresentem melhores chances de retorno e sucesso para as organizações. Compreender os diversos cenários em que o negócio está inserido é essencial para que haja boas condições no direcionamento de estratégias e tomada de decisão.

A análise de dados ou *Data Analytics* é uma forma de retroalimentar os planejamentos e iniciativas da empresa para que sinais, indicativos e insights se transformem em insumos para a definição dos rumos do empreendimento. Análise de dados nada mais é do que um conjunto de técnicas empregadas na transformação

de dados e informações em conhecimento para um propósito específico. A análise de dados busca:

- Estudar os princípios, métodos e sistemas computacionais para extrair conhecimento de dados;
- Identificar as possibilidades de converter dados brutos em conhecimento;
- Encontrar uma forma eficiente de conhecimento (padrões) em (grandes) conjuntos (fluxos) de dados;
- Compreender o comportamento de pessoas e identificar novas oportunidades de negócio.

No início, com os DWs, eram empregadas duas técnicas de análise, a descritiva e a diagnóstica, a fim de compreender o que aconteceu e o motivo pelo qual eventos aconteceram. Mais tarde, isso já não era suficiente, assim, surgiu técnicas como a análise preditiva e a análise prescritiva para tentar prever o que poderia acontecer com os negócios antecipadamente e, assim, precipitar algumas ações.

Como o volume de dados ultrapassava a capacidade humana de interpretar e compreender tanta informação, foi necessário criar mecanismos automáticos para processá-los. O processo para a descoberta de conhecimento em base de dados foi se consolidando como uma boa metodologia para identificar novos conhecimentos. Este processo é um conjunto de etapas realizadas para analisar os dados, baseado na busca, análise e interpretação de padrões úteis, retirados de grandes bases de dados. Uma das suas principais fases é a mineração de dados, pois é onde efetivamente ocorre a extração de conhecimento a partir de dados previamente preparados.

Com a emergência do fenômeno *Big Data*, as organizações se viram “atoladas” em enormes massas de dados. Como já foi mencionado, a variedade de tipos de dados também incrementou a complexidade das análises, pois novas técnicas para mineração surgiram, destaca-se a mineração de textos como uma técnica empregada para extrair conhecimento de textos livres.

Principais tipos de análise de dados

De acordo com a classificação do Gartner, existe uma cadeia de evolução em análise de dados, variando de descritiva à diagnóstica e à preditiva, culminando com prescritiva. Estes tipos de análise de dados ajudam as organizações a compreender dois momentos sobre seus negócios: passado e futuro.

Análise descritiva

Análise descritiva de dados, às vezes descrita como análise exploratória, tem como objetivo entender o cenário atual da organização a partir da análise de seus dados históricos. Trabalha com histórico de dados, cruzando informações com o objetivo de gerar um panorama claro e preciso dos temas relevantes para a empresa no presente momento a partir de seu passado. Em geral, utilizam métricas e técnicas estatísticas simples ou avançadas para entender e explicar como os dados são, buscando explicar o que está acontecendo ou aconteceu em uma determinada situação (TUKEY, 1977).

Análise diagnóstica

Algumas referências incluem a análise diagnóstica como parte da descritiva, isto porque essa análise visa explicar os eventos que ocorreram e foram descritos no modelo de análise anterior. Esse modelo tentar responder à pergunta "*Por que isso aconteceu?*". Nele o foco está na relação de causas e consequências percebidas ao longo do tempo, sobre de um determinado assunto ou evento, cruzando informações com o objetivo de entender quais fatores influenciaram o resultado atual.

Análise preditiva

A análise preditiva é utilizada para prever tendências baseadas nos dados. Segundo o Gartner, a análise preditiva é uma forma de análise avançada que examina dados ou conteúdo para responder à pergunta "*O que vai acontecer?*", ou mais precisamente, "*O que é provável que aconteça?*". Esse tipo de análise é o mais indicado para quem precisa prever algum tipo de comportamento ou resultado. Essa técnica busca analisar dados relevantes ao longo do tempo, buscando padrões comportamentais e suas variações de acordo com cada contexto, a fim de prever como será o comportamento de seu público ou mercado no futuro, dadas as condições atuais. Muito útil para avaliar tendências de consumo e flutuações

econômicas. É caracterizada por técnicas como análise de regressão, previsão, estatísticas multivariadas, correspondência de padrões, modelagem preditiva e previsão.

Análise prescritiva

A análise prescritiva vai um pouco além da preditiva, porém, a lógica envolvida é semelhante. Essa forma de análise examina dados ou conteúdo para responder à pergunta "*O que deve ser feito?*" ou "*O que podemos fazer para fazer algo acontecer?*". Um pouco mais profunda que a análise preditiva, a análise prescritiva traduz as previsões em planos viáveis para o negócio. Ou seja, foca em predizer as possíveis consequências para as diferentes escolhas que forem feitas, e desta forma, este tipo de análise pode recomendar melhores caminhos a serem seguidos. É caracterizada por técnicas, como análise de gráficos, simulação, processamento de eventos, redes neurais, mecanismos de recomendação, heurística e aprendizado de máquina.

Web mining

Web Mining (Mineração na Web) corresponde à aplicação de técnicas de Data Mining (Mineração de Dados) à Web. Ou seja, Web Mining é o processo de extração de conhecimento a partir dos dados da Web. São três tipos de Web Mining: 1) Mineração de conteúdo da Web (*Web Content Mining*): extrai informação do conteúdo dos recursos Web. 2) Mineração da estrutura da Web (*Web Structure Mining*): tem como objetivo principal extrair relacionamentos, previamente desconhecidos, entre recursos Web. 3) Mineração de uso da Web (*Web Usage Mining*): utiliza técnicas de mineração de dados para encontrar analisar ou descobrir padrões de navegação do usuário nos sites. O objetivo é melhorar a experiência do usuário nas aplicações Web (SCIME, 2005).

Text mining

A Mineração de Texto (*Text Mining*), consiste na aplicação de técnicas de mineração de dados para obtenção de informações importantes em um texto. É um processo que utiliza algoritmos capazes de analisar coleções de documentos texto escritos em linguagem natural com o objetivo de extrair conhecimento e identificar padrões. Dentre as técnicas utilizadas, destaca-se o processamento de linguagem natural (SILVA, 2002; TAN, 1999).

Capítulo 5. Mineração de Dados (Data Mining)

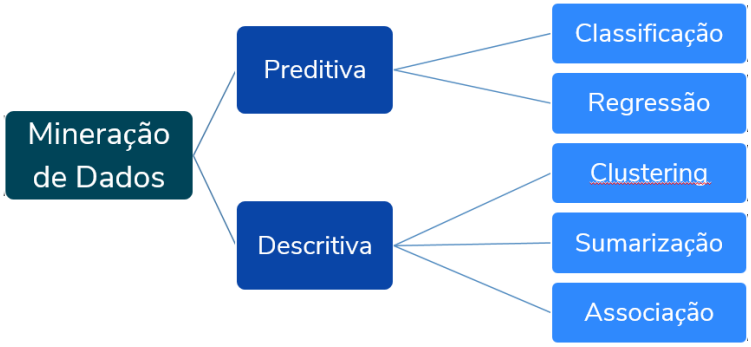
Como já mencionado no Capítulo 1, o *data mining,* ou mineração de dados (DM), é responsável pela seleção dos métodos a serem utilizados para localizar padrões nos dados, seguida da efetiva busca por padrões de interesse numa forma particular de representação, juntamente com a busca pelo melhor ajuste dos parâmetros do algoritmo para a tarefa em questão. Reconhece-se, no entanto, que nem todo processo de DM é conduzido em um contexto de KDD. O termo mineração de dados é muito abrangente, se referindo a dúzias de técnicas e procedimentos usados para examinar e transformar dados (FAYYAD *et al.*, 1996b, HAN *et al*., 2011, WITTEN *et al*., 2016, WU *et al*., 2014).

A mineração de dados pode ser dividida em dois tipos: direcionada e não direcionada. A direcionada tenta prever um ponto de dados em particular, por exemplo, o preço de venda de um imóvel a partir de informações sobre outros imóveis à venda na mesma região. Já na não direcionada não existe um ponto de pesquisa em particular, visa criar grupos de dados, ou encontrar padrões em dados existentes e deixar que estes padrões lhe digam algo que você não conheça.

A mineração de dados caracteriza-se pela existência do algoritmo minerador, que diante da tarefa especificada será capaz de extrair de modo eficiente conhecimento implícito e útil das bases de dados preparadas. Nessa etapa defini-se as principais técnicas e os algoritmos que serão aplicados na necessidade de descoberta conhecimento definida previamente. Dentre os exemplos de técnicas que podem ser aplicadas durante a mineração de dados, existem àquelas baseadas na Inteligência Artificial são elas: Redes Neurais Artificiais, as Árvores de Decisões, a Teoria dos Conjuntos *Fuzzy*, os Algoritmos Genéticos ou, ainda, combinações entre essas técnicas gerando os chamados sistemas híbridos (HAN *et al*., 2011, WITTEN *et al*., 2016).

O tipo de tarefa de KDD que será executado impacta diretamente na escolha da técnica de mineração de dados (FAYYAD *et al*., 1996a; 1996b). A seguir, na Figura 12, apresentamos algumas tarefas de KDD associadas ao tipo de análise. Tais tarefas são sumarizadas em seguida.

Figura 12: Tarefas de mineração de dados.



Descoberta de Associação

Utilizada para descrever características com alto grau de "correlação" no conjunto de padrões a serem reconhecidos em um conjunto de dados (FAYYAD *et al*., 1996a). Ela pode ser exemplificada com o cotidiano de compras dentro de um supermercado, pois pode-se conseguir com o histórico de compras por parte dos clientes, obtendo algumas associações, como quem compra o produto X quase sempre compram o produto Y. Para citar alguns exemplos de ferramentas que implementam a tarefa de descoberta de associações temos algoritmos como o *Apriori*, GSP (*Generalized Sequential Pattern*), DHP (*Direct Hashing and Pruning*), dentre outros (GOLDSCHMIDT, PASSOS, 2005).

Classificação

A técnica de classificação, consiste em identificar uma função que avalie um conjunto de variáveis categóricos pré-definidos, denominados classes. Visa a construção de um modelo de classificação a fim de categorizar os objetos em classes. Pressupõe a existência de características que definem grupos específicos e associa ou classifica um item em uma ou várias classes predefinidas (FAYYAD *et al*., 1996a). Um exemplo seria a avaliação para concessão de crédito em banco, onde, a partir de alguns atributos, pode-se classificar novos pedidos de crédito, julgando-os se seria ou não interessante conceder crédito. Os algoritmos clássicos empregados na função classificação baseiam-se em árvores de decisão, regras de decisão e análise discriminante (GOLDSCHMIDT, PASSOS, 2005).

Regressão

A função regressão é similar à classificação, diferencia-se desta por objetivar a predição de um valor real em vez de um atributo nominal ou uma categoria. Envolve a busca por uma função que avalie os registros de um conjunto de dados em valores reais (FAYYAD *et al*., 1996a). Na regressão, estima-se o valor de uma determinada variável analisando os valores das demais. Um exemplo do uso de regressão refere-se a estimativa da probabilidade de um paciente sobreviver, levando em consideração o resultado de uma série de diagnósticos e exames.

Clusterização

Essa análise tem o objetivo de encontrar grupos de observações intimamente relacionados, de modo que observações dentro de um grupo tenham uma semelhança mais acentuada quando comparadas a outros grupos. Divide os dados de um conjunto em subconjuntos, agrupamento ou clusters, de maneira que tais dados de um agrupamento compartilhem de características comuns que os diferencie dos elementos de outros agrupamentos. Visa maximizar similaridade intra-agrupamento e minimizar similaridade inter-agrupamento. No processo de clusterização ou agrupamento, é necessário identificar automaticamente os agrupamentos de dados aos quais o usuário precisará determinar rótulos (FAYYAD *et al*., 1996a). Como exemplo podemos citar o caso de uma empresa de varejo que deseja agrupar seus clientes conforme o perfil de comprar dos mesmos.

Entre os vários algoritmos empregados na função de clusterização, destacam-se os baseados na teoria de conjuntos nebulosos, como: o fuzzy c-means, o extended fuzzy c-means, o algoritmo de agrupamento participativo, K-Means, K-Modes, K-Prototypes, K-Medoids, Kohonen.

Sumarização

A sumarização busca identificar e determinar características comuns entre conjuntos de dados. Visam localizar uma descrição sintetizada para um subconjunto de dados. Por exemplo, tendo em vista um conjunto de dados com informações sobre clientes que contrataram um determinado produto bancário. O objetivo da sumarização é encontrar características comuns entre os clientes. Características comos: os contratante do serviço X, em geral possuem nível superior, são homens que trabalham na área de tecnologia e estão na faixa etária de 25 a 35 anos.

Estas informaçãoes podem direcionar a áres de marketing da empresa para direcionar a oferta destes produtos para novos clientes. Uma relação comum é combinar a tarefa de sumarização aos agrupamentos obtidos na clusterização. Alguns algoritmos de sumarização são: Lógica Indutiva e Algoritmos Genéticos.

Capítulo 6. Atividades da cadeia de valor do big data

Este capítulo sintetiza algumas das atividades da cadeia de valor do *big data* que também fazem parte do dia a dia das empresas.

Coleta, Preparação e Visualização dos dados são três das atividades relevantes no dia a dia do Analista de Dados. Os dados podem ser coletados de várias fontes diretas e indiretas, internas e externas à organização. A preparação de dados refere-se ao conjunto de atividades realizadas para melhorar a qualidade do dado ou transformar os dados brutos em um formato plausível para ser usado e analisado (REHMAN *et al*., 2016).

Coleta de dados

A coleta precede as demais atividades de análise de dados, pois tudo começa com a obtenção dos dados. Ela nada mais é que o processo de obtenção de dados de uma ou mais fontes (REHMAN *et al.*, 2016). Para realizar a coleta, é necessário identificar as fontes e os respectivos tipos de dados que cada fonte provê.

As fontes podem ser internas, como os dados disponíveis na organização, tais como os oriundos dos CRMs, ERPs, SCMs, entre outros sistemas de processamento de transação usados pela empresa; ou externas, como os dados disponíveis na WEB, adquiridos junto a empresas especializadas e dados abertos disponibilizados, por exemplo, por órgãos governamentais.

A coleta de dados nas bases internas pode ser realizada utilizando a linguagem SQL e aplicações desenvolvidas em ferramentas de ETL, como o Pentaho Data Integration[2], a Plataforma Knime Analytics ou em linguagens, como Java e Python.

[2] Para detalhes, acesse: https://help.pentaho.com/Documentation/7.1/0D0/Pentaho_Data_Integration ou https://www.infoq.com/br/articles/pentaho-pdi/.

Um tipo de coleta de fontes externas que merece destaque é a Web, neste caso, não apenas os dados abertos, mas também as mídias sociais em geral. Em geral, essas coletas podem ser realizadas utilizando APIs de coleta de dados ou rotinas de rastreamento e raspagem (web crawler e web scraping) (BENEVENUTO; ALMEIDA; SILVA, 2011; MUNZERT *et al.*, 2014).

Preparação de dados

A preparação de dados é uma das etapas mais importante do pipeline de big data, pois é nela que são realizadas as operações para melhorar a qualidade dos dados. A qualidade vai determinar a eficiência e acuracidade das análises de dados. Os conjuntos de dados são susceptíveis a ruídos, valores faltantes e outras inconsistências. Desta forma, a preparação deles visa, acima de tudo, transformar os dados brutos em um formato plausível para ser utilizado nas análises.

As operações de preparação consistem em limpar, enriquecer, normalizar, integrar e combinar dados para análise, e incluem uma ampla gama de métodos que são usados principalmente para os seguintes fins:

Limpeza dos dados: Operações de tratamento sobre os dados pré-existentes, de forma a assegurar a qualidade (completude, consistência, veracidade e integridade). Consiste em resolver problemas como a retirada de dados duplicados, correção de dados corrompidos ou inconsistentes, tratamento de valores ausentes ou inaplicáveis; detecção e remoção de anomalias (ruídos, outliers, valores de dados irregulares, incomuns e indesejados).

Enriquecimento de dados: Operações que visam agregar aos dados existentes mais dados (detalhes), tornando os dados mais ricos de modo que possam contribuir no processo de descoberta de conhecimento. Em geral, a partir de dados existentes é possível coletar novos dados externos, que tenham algum grau correlação, por meio de processos de integração e combinação de dados.

Transformação de dados: Operações que visam transformar os dados conforme alguma regra, por exemplo, padronização e normalização dos dados,

conversão de valores categóricos em numéricos e vice-versa, geração de hierarquia de conceitos.

Integração de dados: Operações que visam a fusão de dados de fontes distintas em um único conjunto.

Redução dos dados: Operações que visam criar um conjunto reduzido da série de dados que produz (quase) o mesmo efeito nas análises. Pode ser por redução de dimensionalidade ou redução no volume de dados.

Visualização de dados

Quando falamos de visualização de dados, nosso inquietação é: “Não basta termos os dados, é preciso saber mostrá-los”. Assim, alguns mecanismos de visualização de dados aplicados nas análises são chamados da *dashboards* ou painéis e *data storytelling*. *Storytelling* consiste na ação de contar uma história, no caso, contar uma história sobre seus dados.

No contexto de Análise de Dados, o *Data Storytelling* (ou em português: contar uma história por meio de dados) é uma técnica fundamental que permite apresentar o que foi feito, os resultados obtidos e as respectivas análises, mantendo seu principal interessado envolvido com o conteúdo. Para conhecer um pouco mais sobre Data Storytelling, sugiro que seja realizada a leitura deste material *O que é Data Storytelling?,* indicado no link: https://paulovasconcellos.com.br/o-que-%C3%A9-data-storytelling-ac5a924dcdaf (acesso em Julho/2021)

Segundo Few (2006), os dashboadrs são definidos da seguinte forma: “Mostrador Visual para as principais informações necessárias para se atingir um ou mais objetivos, consolidados e arranjados em uma única tela de modo que a informação seja monitorada de uma só vez.”

A definição de *dashboard* como um mostrador visual de dados reúne em uma única tela gráficos, textos, alertas em forma de cores ou objetos visuais, setas, semáforos e demais recursos visuais com o intuito de tornar os dados apresentados atrativos para quem necessita. Os dashboards são painéis visuais que mostram

métricas e indicadores importantes para alcançar objetivos e metas traçadas de forma visual, facilitando a compreensão das informações geradas.

Referências

ALMEIDA, M. B. d. Revisiting ontologies: A necessary clarification. In: *Journal of the American Society for Information Science and Technology*, v. 64, n. 8, p. 1682-1693, 2013.

AZEVEDO, A. I. R. L.; SANTOS, M. F. KDD, SEMMA and CRISP-DM: a parallel overview. In: *IADS-DM*, 2008.

BARBIERI, C. P. *BI2 - Business Intelligence: Modelagem e Qualidade*. Rio de Janeiro: Elsevier, 2011.

BARBIERI, C. P. *BI-business intelligence: modelagem e tecnologia*. Axcel Books, 2001.

BDW, B. D. W. *Introduction about NoSQL Data Models*. Big Data World. on-line. 2019 2014.

BENEVENUTO, F.; ALMEIDA, J. M.; SILVA, A. S. *Explorando redes sociais online: Da coleta e análise de grandes bases de dados às aplicações*. Campo Grande, Brasil: Sociedade Brasileira de Computação, 2011. p. 63-102.

BERNERS-LEE, T. *Linked Data*. 2006. Disponível em: <https://www.w3.org/DesignIssues/LinkedData.html>. Acesso em: 08 jul. 2021.

BERNERS-LEE, T.; HENDLER, J.; LASSILA, O. The semantic web. In: *Scientific american*, v. 284, n. 5, p. 28-37, 26 abr. 2001.

BIZER, C.; HEATH, T.; BERNERS-LEE, T. Linked data-the story so far. In: *Semantic Services, Interoperability and Web Applications: Emerging Concepts*, p. 205-227, 2009.

CANALTECH. *EMC oferece solução de armazenamento e análise de Data Lake - Infra*. CanalTech, 3 abr. 2015. Disponível em: <https://canaltech.com.br/infra/EMC-oferece-solucao-de-armazenamento-e-analise-de-Data-Lake/>. Acesso em: 08 jul. 2021.

CHEN, P. P. *Modelagem de dados: a abordagem entidade-relacionamento para projeto lógico*. Makron Books do Brasil, 1990.

CHEN, P. P. The entity-relational model toward a unified view of data. In: *ACM Trans, on Database Systems*, v. 1, n. 1, p. 1-49, 1976.

COUGO, P. S. *Modelagem conceitual e projeto de banco de dados*. 1. ed. 18ª Reimp. ed. Rio de Janeiro: Elsevier, 1997.

COULOURIS, G.; DOLLIMORE, J.; KINDBERG, T.; BLAIR, G. *Sistemas Distribuídos: Conceitos e Projeto*. Bookman Editora, 2013.

COX, M.; ELLSWORTH, D., 1997, English, Phoenix, AZ, USA. *Application-controlled demand paging for out-of-core visualization*. IEEE Computer Society Press. 235-ff. Disponível em: <https://www.nas.nasa.gov/assets/pdf/techreports/1997/nas-97-010.pdf>. Acesso em: 08 jul. 2021.

CURRY, E. The Big Data Value Chain: Definitions, Concepts, and Theoretical Approaches. In: CAVANILLAS, J. M.;CURRY, E. e WAHLSTER, W. (Ed.). *New Horizons for a Data-Driven Economy: A Roadmap for Usage and Exploitation of Big Data in Europe*. Springer, 2016. p. 29-37. Disponível em: <https://link.springer.com/content/pdf/10.1007%2F978-3-319-21569-3_3.pdf>. Acesso em: 08 jul. 2021.

DAMA. *DAMA-DMBOK: Data management body of knowledge.* Bas King Ridge, Nova Jersey, EUA: Technics Publications LLC, 2017. p. 624.

DAMA. *The DAMA Guide to The Data Management Body of Knowledge (DAMA-DMBOK Guide)*. Bas King Ridge, Nova Jersey, EUA: Technics Publications LLC, 2009. p. 406.

DATA SCIENCE GUIDE, D. *Exploratory data analysis*. Data science guide. on-line. 2021.

DAVENPORT, T. H. *Ecologia da informação: por que só a tecnologia não basta para o sucesso na era da informação*. Tradução ABRÃO, B. S. 2. ed. São Paulo: Futura, 1998. p. 292.

ELMASRI, R.; NAVATHE, S. B. *Sistemas de banco de dados*. 4. ed. São Paulo: Addison Wesley, 2005.

FARINELLI, F. *Improving semantic interoperability in the obstetric and neonatal domain through an approach based on ontological realism.* Orientador: ALMEIDA, M. B. d. 2017. 256 f. Doctoral (Doctor in Information Science) - School of Information Science Federal University of Minas Gerais at Brazil, Belo Horizonte. Disponível em: <http://www.bibliotecadigital.ufmg.br/dspace/handle/1843/BUBD-AX2J5B>. Acesso em: 08 jul. 2021.

FOWLER, M. *Nosql Definition*. 2016. Disponível em: <https://martinfowler.com/bliki/NosqlDefinition.html>. Acesso em: 08 jul. 2021.

HEATH, T.; BIZER, C. Linked data: Evolving the web into a global data space. In: *Synthesis lectures on the semantic web: theory and technology*, 1, n. 1, p. 1-136, 2011.

HECKERT, N. A.; FILLIBEN, J. J. *NIST/SEMATECH.* e-Handbook of Statistical Methods; Chapter 1: Exploratory Data Analysis. 2003.

HEUSER, C. A. *Projeto de Banco de Dados*. 6. ed. Porto Alegre: Bookman, 2008. p. 282.

ISOTANI, S.; BITTENCOURT, I. I. *Dados Abertos Conectados: Em busca da Web do Conhecimento*. Novatec Editora, 2015.

KAUSAR, M. A.; DHAKA, V. S.; SINGH, S. K. Web crawler: a review. In: *International Journal of Computer Applications*, v. 63, n. 2, 2013.

LANEY, D. 3D data management: Controlling data volume, velocity and variety. In: *META Group Research Note*, v. 6, p. 70-73, 2001.

MAYER-SCHÖNBERGER, V.; CUKIER, K. *Big data: a revolution that will transform how we live, work, and think*. Boston: Houghton Mifflin Harcourt, 2013. p. 242.

MEDRI, W. *Análise exploratória de dados*. Apostila do curso de especialização em Estatística. Londrina: Universidade Estadual de Londrina, 2011.

MITCHELL, R. *Web scraping with Python: Collecting more data from the modern web*. O'Reilly, 2018.

MUNZERT, S.; RUBBA, C.; MEIßNER, P.; NYHUIS, D. *Automated data collection with R: A practical guide to web scraping and text mining*. John Wiley & Sons, 2014.

NASCIMENTO, J. P. B. *A carreira dos profissionais de Ciência de Dados, Engenharia de Dados e Machine Learning*. IGTI Blog, 2017. Disponível em: <http://igti.com.br/blog/carreira-big-data-engenharia-dados-machine-learning/>. Acesso em: 20 mai. 2021.

OKI, O. K. I. *Guia de Dados Abertos*. Open Data Handbook, 2019. Disponível em: <http://opendatahandbook.org/guide/pt_BR/>. Acesso em: 08 jul. 2021.

PARUCHURI, V. *What is Data Engineering?*. Dataquest, 2017. Disponível em: <https://www.dataquest.io/blog/what-is-a-data-engineer/>. Acesso em: 08 jul. 2021.

REHMAN, M. H. u.; CHANG, V.; BATOOL, A.; WAH, T. Y. Big data reduction framework for value creation in sustainable enterprises. In: *International Journal of Information Management*, v. 36, n. 6, Part A, p. 917-928, 2016.

SADALAGE, P. J.; FOWLER, M. *NoSQL distilled: a brief guide to the emerging world of polyglot persistence*. Pearson Education, 2013.

SCIME, A. *Web Mining: applications and techniques*. IGI Global, 2005.

SHAFIQUE, U.; QAISER, H. A comparative study of data mining process models (KDD, CRISP-DM and SEMMA). In: *International Journal of Innovation and Scientific Research*, v. 12, n. 1, p. 217-222, 2014.

SHIVALINGAIAH, D.; NAIK, U. Comparative Study of Web 1.0, Web 2.0 and Web 3.0. In: *6th International CALIBER*, Allahabad, Índia. 2008, p. 499-507.

SILBERSCHATZ, A.; KORTH, H. F.; SUDARSHAN, S. *Sistemas de banco de dados*. 6. ed. São Paulo: Elsevier, 2012.

SILVA, E. M. *Descoberta de conhecimento com o uso de text mining: cruzando o abismo de moore*. 2002. -, Universidade Católica de Brasília Disponível em: <https://bdtd.ucb.br:8443/jspui/handle/123456789/1462>. Acesso em: 08 jul. 2021.

SOUSA, F. R. *et al*. Gerenciamento de dados em nuvem: Conceitos, sistemas e desafios. In: *Topicos em sistemas colaborativos, interativos, multimidia, web e bancos de dados, Sociedade Brasileira de Computacao*, p. 101-130, 2010.

SOUSA, F. R.; MOREIRA, L. O.; MACHADO, J. C. Computação em nuvem: Conceitos, tecnologias, aplicações e desafios. In: *II Escola Regional de Computação Ceará, Maranhão e Piauí (ERCEMAPI)*, p. 150-175, 2009.

STROHBACH, M. *et al*. Big Data Storage. In: CAVANILLAS, J. M.;CURRY, E. e WAHLSTER, W. (Ed.). *New Horizons for a Data-Driven Economy: A Roadmap for Usage and Exploitation of Big Data in Europe*. Cham: Springer International Publishing, 2016. p. 119-141.

TAN, A.-H. *Text mining: The state of the art and the challenges*. 1999. sn. 65-70. Disponível em: <http://www.ntu.edu.sg/home/asahtan/papers/tm_pakdd99.pdf>. Acesso em: 20 mai. 2021.

TAURION, C. *Big data*. Rio de Janeiro: Brasport, 2013.

TAURION, C. *Cloud computing-computação em nuvem*. Brasport, 2009.

THELWALL, M. A web crawler design for data mining. In: *Journal of Information Science*, v. 27, n. 5, p. 319-325, 2001.

TUKEY, J. W. *Exploratory data analysis*. Reading, Mass., 1977.

VANDEN BROUCKE, S.; BAESENS, B. From web scraping to web crawling. In: *Practical Web Scraping for Data Science*. Springer, 2018. p. 155-172.

WHITE, T. *Hadoop: The definitive guide*. O'Reilly Media, 2012.