Aprenda com quem faz

Fundamentos em Análise e Ciências de dados

Leandro Lessa

2024

Sumário

> Capítulo 1

Capítulo 1. Introdução à Análise e Ciência de Dados

Cenário atual do Big Data

Com o advento da internet, a geração de dados em escala global experimentou um crescimento substancial ao longo dos anos. A popularização das mídias sociais e o uso excessivo de dispositivos móveis contribuíram ainda mais para o aumento exponencial na produção diária de dados. De acordo com a International Data Corporation (IDC), uma autoridade líder em inteligência de mercado, consultoria e eventos, a quantidade de dados digitais gerados globalmente dobra a cada dois anos. O nosso mundo agora se movimenta impulsionado pelos dados e pela riqueza de insights que podem ser extraídos deles.

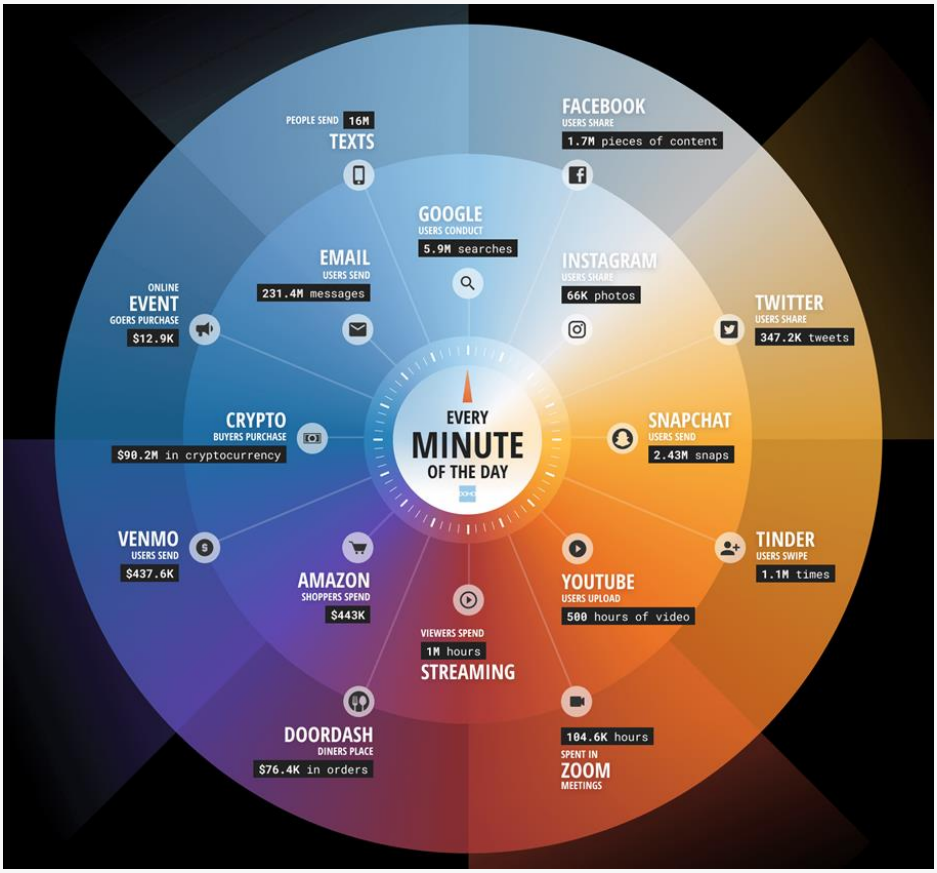
Os dados são gerados em uma taxa constante e crescente. Desde *smartphones*, plataformas de mídia social e até mesmo tecnologias de diagnóstico médico baseadas em imagem, todos esses elementos contribuem para a geração incessante de novos dados. A demanda por armazenamento e processamento em tempo real se torna crucial, à medida que dispositivos e sensores automatizados continuamente produzem informações diagnósticas. A simples gestão desse vasto fluxo de dados já é uma tarefa desafiadora, mas a análise eficaz de volumes enormes de informações se torna ainda mais complexa, especialmente quando esses dados não se encaixam nos modelos convencionais de estrutura, tornando a identificação de padrões significativos e a extração de insights valiosos uma tarefa hercúlea. No entanto, esses desafios inerentes à era do big data também representam oportunidades transformadoras nos âmbitos empresariais, governamentais, científicos e cotidianos.

Vários setores têm liderado o desenvolvimento de suas habilidades de coleta e exploração de dados, e esse progresso tecnológico nos trouxe a

um ponto em que geramos milhões de dados a cada minuto. (EMC EDUCATION, 2015)

A percepção da escala da produção de dados torna-se mais clara quando nos deparamos com comparações reveladoras. A DOMO, empresa especializada em computação em nuvem, realiza anualmente um estudo que estima a quantidade de informações geradas a cada minuto, todos os dias. Além de apresentar estimativas desses dados, o estudo lança luz sobre os padrões de comportamento e os padrões de consumo da população nas principais plataformas de mídia social. A Figura 1 ilustra a quantidade de dados gerados por minuto no ano de 2022.

Figura 1 – Dados produzidos por minuto em 2022.



Fonte: DOMO, 2022.

Conforme ilustrado na Figura 1, observamos as principais plataformas que mais produziram dados por minuto em 2022. Podemos tirar

várias informações dessa imagem, como por exemplo a quantidade de horas gastas em reuniões via *Zoom*, a quantidade de fotos postadas no *Instagram*, ou quanto foi gasto em compras na *Amazon*. Com tantos dados produzidos a cada minuto, os métodos tradicionais de coleta, armazenamento e processamento de dados começaram a não ser suficientes, causando problemas e gastos cada vez maiores para suprir as necessidades do negócio.

Diante disso, surge o conceito do Big Data. Definimos Big Data como um termo genérico para qualquer coleção de conjuntos de dados tão grandes ou complexos que se torna difícil processá-los usando técnicas tradicionais de gerenciamento de dados, como, por exemplo, o SGBDs (Sistema de Gerenciador de Banco de Dados Relacional) (CIELEN; MEYSMAN; ALI, 2016). Em outras palavras, podemos dizer que o Big Data, é uma área do conhecimento que tem como finalidade estudar maneiras de tratar, analisar, processar, armazenar e gerar conhecimento através de grandes volumes de dados de várias fontes, com o objetivo de acelerar a tomada de decisão e trazer vantagem competitiva.

Os SGBDs eram amplamente adotados há muito tempo e é considerado uma solução de tamanho único, mas as demandas de manipulação de Big Data mostraram o contrário. A ciência de dados envolve o uso de métodos para analisar grandes quantidades de dados e extrair o conhecimento que eles possuem. Você pode pensar na relação entre Big Data e Ciência de Dados como a relação entre petróleo bruto e uma refinaria de petróleo. A Ciência de Dados e o Big Data evoluíram da estatística e do gerenciamento de dados tradicionais, mas agora são considerados disciplinas distintas. (CIELEN; MEYSMAN; ALI, 2016)

Ciências de dados

A ciência de dados é uma extensão evolutiva da estatística, capaz de lidar com as enormes quantidades de dados produzidos hoje e acrescenta métodos da ciência da computação ao repertório da estatística. Em uma nota de pesquisa de *Laney e Kart, Emerging Role of the Data Scientist and the Art of Data Science*, os autores vasculharam centenas de descrições de cargos para cientista de dados, estatístico e analista de BI (*Business Intelligence*) para detectar as diferenças entre esses títulos. As principais coisas que diferenciam um cientista de dados de um estatístico são a capacidade de trabalhar com big data e experiência em aprendizado de máquina, computação e construção de algoritmos. Suas ferramentas também tendem a diferir, com descrições de cargos de cientistas de dados mencionando com mais frequência a capacidade de usar *Hadoop, Pig, Spark, R, Python* e *Java*, entre outros. (CIELEN; MEYSMAN; ALI, 2016)

Em outras palavras, a ciências de dados é o estudo que consegue extrair insights importantes para a área de negócios. Por ter uma abordagem multidisciplinar a ciências de dados consegue combinar vários campos, como por exemplo: as áreas de matemática, estatística, métodos científicos, IA (Inteligência Artificial) e *Machine Learning* para analisar e extrair conhecimento de grandes quantidades de dados. Através da Ciências de Dados é possível tomar decisões baseada nas análises e resultados obtidos dos estudos realizados e entender o que aconteceu no passado, o motivo do acontecimento, o que poderá acontecer no futuro e quais estratégias podemos tomar baseado nos resultados. (AWS, 2022)

Profissão Analista de Dados e Cientista de Dados

Nos últimos anos, o mundo tem testemunhado uma explosão de dados gerados a cada segundo. Com o advento da tecnologia e o crescente uso de dispositivos digitais, empresas e organizações têm à sua disposição uma quantidade enorme de informações. Nesse contexto, surgiram duas

profissões fundamentais para extrair valor desses dados: o Analista de Dados e o Cientista de Dados. Ambos desempenham papéis vitais na análise e interpretação dos dados, porém cada um com suas responsabilidades e áreas de atuação específicas.

O Analista de Dados é responsável por coletar, organizar e analisar grandes volumes de dados para identificar padrões e tendências relevantes. Sua principal tarefa é extrair informações valiosas dos dados disponíveis, transformando-os em insights acionáveis para a tomada de decisões. O analista de dados utiliza ferramentas e técnicas estatísticas para processar e visualizar os dados, gerando relatórios e apresentações que auxiliam gestores e tomadores de decisão a compreenderem melhor o cenário atual e planejarem estratégias futuras.

Por outro lado, o Cientista de Dados vai além das habilidades do analista de dados. Além de lidar com grandes volumes de dados, o cientista de dados é especializado em desenvolver modelos preditivos e algoritmos complexos para descobrir padrões ocultos e realizar previsões. Ele utiliza técnicas avançadas de análise estatística, aprendizado de máquina e inteligência artificial para extrair conhecimento dos dados. O cientista de dados também está envolvido no processo de coleta e limpeza dos dados, além de ser responsável por criar e implementar soluções personalizadas para os problemas específicos de uma empresa.

Ambas as profissões têm se tornado cada vez mais relevantes em diversos setores, como finanças, saúde, marketing e tecnologia. A demanda por profissionais qualificados em análise de dados tem crescido exponencialmente e estima-se que essa tendência se mantenha nos próximos anos.

Mercado de Trabalho

Hoje, o mercado de trabalho está aquecido para os setores da tecnologia. Desde os técnicos aos engenheiros, analistas e cientistas de dados, as oportunidades têm surgido frequentemente para aqueles que sabem se posicionar frente ao mercado.

Há quem diga que hoje os dados são tão importantes quanto o petróleo foi no século passado e a tendência é estar cada dia mais valorizado, pois a tecnologia tem avançado e vai se tornar cada vez mais presente no dia a dia da comunidade e do mundo corporativo.

Para aproveitar essas oportunidades, o mais importante é saber como se posicionar, tendo sempre um currículo atualizado, assim como o *LinkedIn*, que hoje pode ser considerado uma das maiores redes de *network*. Há ali a disponibilização de vagas diariamente e se o seu perfil estiver bem atualizado, os próprios recrutadores poderão entrar em contato com você.

Claro que para se mostrar competitivo nesse mercado, será necessário que você esteja sempre buscando as atualizações da área de tecnologia, como já vimos anteriormente, e estar em um crescimento contínuo de conhecimento.

Como é o mercado de trabalho?

As vagas para área de dados estão em crescimento contínuo nos últimos anos. Como o trabalho de um cientista de dados está ligado diretamente em converter um grande volume de dados em *insights* valiosos, sua tarefa é trazer inteligência nos negócios e gerar vantagem competitiva para a empresa. Como a profissão é relativamente nova, hoje a demanda é maior do que o número de profissionais qualificados preparados para atendê-la. Desta forma, as oportunidades de trabalhos estão cada vez maiores para quem está se qualificando.

Tomar decisões baseada em dados é um ativo precioso para qualquer empresa, seja empresas do setor público ou privado. Essas oportunidades podem ser tanto para contratos baseado nas leis trabalhistas CLT como oportunidades como prestador de serviços de pessoa jurídica (PJ).

Afinal, qual o salário de um cientista de dados?

Essa é uma grande pergunta que todo mundo gostaria de responder. Isso vai variar de empresa para empresa, de projeto ou se é um contrato PJ ou CLT. Uma pesquisa realizada pelo site de recrutamento *Glassdoor* mostrou que a média salarial de um cientista de dados no Brasil é de R$26.700,00. No entanto, fora do Brasil existem grandes oportunidades de trabalho com salários muito mais atrativos. A Figura 02 ilustra a média salarial de um cientista de dados no Brasil.

Figura 2 – Média salarial de um cientista de dados no Brasil.

Fonte: Glassdoor (2023).

Áreas de Conhecimento

A profissão de cientista de dados tem ganhado destaque como uma das carreiras mais promissoras e desafiadoras do século XXI. Esses profissionais têm a tarefa de extrair conhecimento valioso a partir de dados complexos, fornecendo insights estratégicos para as empresas e auxiliando na tomada de decisões fundamentadas.

1. Estatística e Matemática

A compreensão profunda de conceitos estatísticos é fundamental para um cientista de dados. Desde a análise exploratória até a construção de modelos preditivos, as técnicas estatísticas são a base para a interpretação correta dos dados. Além disso, sólidos conhecimentos em matemática, especialmente álgebra linear e cálculo, são essenciais para entender e desenvolver algoritmos complexos utilizados na ciência de dados.

2. Programação e Ciência da Computação

A habilidade de programar é uma das competências centrais para um cientista de dados. Linguagens como Python e R são amplamente utilizadas na comunidade de ciência de dados, permitindo a manipulação de dados, a criação de modelos e a visualização dos resultados. Além disso, conhecimentos sólidos em ciência da computação são importantes para otimizar a eficiência dos algoritmos e desenvolver soluções escaláveis (MCKINNEY, 2018).

3. Aprendizado de Máquina e Inteligência Artificial

Essas áreas estão no cerne da ciência de dados. O aprendizado de máquina envolve a criação de algoritmos que permitem aos computadores aprenderem a partir dos dados, reconhecerem padrões e fazerem previsões. Já a inteligência artificial busca desenvolver sistemas que possam simular a capacidade humana de raciocínio e tomar decisões. O domínio dessas áreas permite que o cientista de dados crie modelos preditivos e sistemas inteligentes (GOODFELLOW; BENGIO & COURVILLE, 2016).

4. Bancos de Dados e Gerenciamento de Dados

Para lidar com a enorme quantidade de dados disponíveis atualmente, o conhecimento de bancos de dados é essencial. Aprender a manipular e acessar os dados de forma eficiente, seja em bancos de dados relacionais ou não relacionais, é crucial para o trabalho do cientista de dados. Além disso, entender conceitos de gerenciamento de dados é importante para garantir a integridade e a segurança das informações (HAN; KAMBER & PEI, 2011).

5. Visualização de Dados

A capacidade de comunicar os insights obtidos a partir dos dados é uma habilidade valiosa para um cientista de dados. A visualização de dados permite apresentar informações complexas de maneira clara e compreensível, tornando os resultados mais acessíveis para os tomadores de decisão. Dominar técnicas de visualização é crucial para transmitir os insights de forma impactante (CAIRO, 2016).

As áreas de conhecimento mencionadas são pilares fundamentais para a atuação de um cientista de dados. Essa profissão multifacetada exige habilidades em estatística, programação, aprendizado de máquina, gerenciamento de dados e visualização, entre outras. A combinação dessas competências permite que o cientista de dados explore e aproveite o potencial dos dados, fornecendo insights para as empresas e contribuindo para o avanço da ciência e da tecnologia.

Ciência de Dados e Estatística

Precisamos entender o papel central da estatística dentro da Ciência de Dados. Ela é utilizada desde a definição do tipo de experimento, na coleta dos dados de maneira eficiente, nos testes de hipóteses, estimação de parâmetros até a interpretação dos resultados. Foi a estatística em sua evolução que proporcionou a ciência de dados chegar a um lugar tão estratégico hoje em dia.

A estatística valida a análise dos dados, pois é através dela que o profissional poderá convencer os demais da confiabilidade do processo que foi tomado até o resultado apresentado. É a estatística que proporciona a forma e as ferramentas para encontrar a disposição dos dados e então gerar os *insights* sobre o que de fato está apresentado ali. As medidas estatísticas de média, mediana, moda, desvio padrão e distribuição, atuam revelando o comportamento das variáveis que estão sendo observadas e apontam as anomalias daqueles dados. A estatística é o que explica o porquê que o *Machine Learning* funciona, assim como outras ferramentas, o que permite ao cientista escolher qual delas se aplica melhor ao seu objetivo (IA EXPERT, 2022).

Áreas de atuação do Cientista de Dados

A escolha de uma área de atuação vai depender da aptidão de cada profissional. Afinal de contas, cada profissional tem tipos de personalidades diferentes e possui habilidades distintas que podem se encaixar em possibilidades que melhor se adequam àqueles cargos. Abaixo segue uma lista de áreas em que esse profissional pode atuar:

- Consultoria especialista em tecnologia.
- Pesquisa de mercado ou científica.
- Empresas de manutenção de computadores.
- Instituições financeiras.
- Instituições de ensino superior.
- Agências de publicidade e propaganda.
- Indústrias.

O trabalho de um cientista de dados está mais presente no nosso dia a dia do que podemos perceber ou imaginar. Recomendação de produtos em um *e-commerce*, anúncios que aparecem no nosso navegador de internet, propagandas direcionadas no *YouTube* ou até mesmo disponibilização de itens em um supermercado são exemplos de um trabalho desenvolvido por esse profissional. O cientista de dados é responsável por gerar esses valores para empresa utilizando dados e tecnologia, apoiado por algoritmos de *Machine Learning*.

Onde Aplicar a Ciência de Dados

Um cientista de dados poderá exercer sua função em qualquer lugar que gere dados e que precisa que eles sejam tratados e examinados. Hoje, com empresas adeptas ao *Data Driven*, um universo que gira ao redor de dados que são gerados aos milhões, não existe um único nicho para a ciência de dados, pois ela tem conhecimento e aplicação interdisciplinar.

Setores como educação, comércio, produção, *marketing*, financeiro ou mesmo jurídico podem se beneficiar da ciência de dados para melhorar seus indicadores ou então tomar decisões estratégicas acerca de seus negócios. É a ciência de dados que permite às empresas criarem planos estratégicos e de negócios com base nas informações e análises do comportamento de seu cliente ou mercado, além de avaliar as tendências e suas concorrentes.

O grande ponto de importância é que, independentemente do ramo de negócio da empresa, os cientistas de dados devem estar alinhados com a estratégia da empresa, para que possam definir bem suas métricas e desempenhar um trabalho relevante para a companhia. Abaixo seguem outros exemplos reais em que a ciência de dados pode atuar:

Automotivo: aplicam análises avançadas de dados de motorista, veículo, suprimentos e IoT (Internet of Things) para melhorar a eficiência de

fabricação de peças automotivas e para melhorar a segurança e a Inteligência artificial para carros com motoristas autônomos.

Financeiros: utilizam dados de clientes e transações para reduzir o risco de fraudes, aumentar retorno e melhorar a satisfação de clientes.

Mídia e entretenimento: análise de dados de público e conteúdo para aprofundar o envolvimento do público nas programações, reduzir a rotatividade e otimizar as receitas de publicidade.

Saúde: utilizam dados para monitoramento de pacientes em tempo real. Análise de padrões de doenças, extração de informação em imagens médicas, descoberta e desenvolvimento de novos medicamentos e análise de dados genéticos.

Telecomunicações: utilizam os dados de clientes e da rede para melhorar os serviços e o desempenho da rede, analisar registros de chamadas, alocação de banda em tempo real, planejamento da rede, redução de rotatividade de clientes.

Varejo: utilizam dados de clientes e produtos para realizarem análise de sentimento, segmentação de mercado e cliente, *marketing* personalizado e previsão de demandas.

Principais competências e habilidades necessárias para atuar na área

A área de análise de dados tem se tornado cada vez mais relevante em diferentes setores da economia. Com o crescimento exponencial da quantidade de dados disponíveis, empresas estão buscando profissionais capacitados para extrair insights dessas informações. Logo abaixo exploraremos as principais competências e habilidades necessárias para atuar nessa área:

1. Conhecimento em Estatística e Matemática

A compreensão dos conceitos estatísticos é fundamental para interpretar e analisar dados com precisão. Além disso, conhecimentos em matemática, especialmente em álgebra linear e cálculo, são úteis para desenvolver modelos estatísticos mais complexos (DeGroot & Schervish, 2012).

2. Domínio de Linguagens de Programação

A habilidade de programar é essencial para manipular e analisar grandes volumes de dados. Linguagens como Python, R e SQL são amplamente utilizadas na área de análise de dados. O conhecimento dessas linguagens permite a implementação de algoritmos, a criação de visualizações e a automatização de tarefas (VANDERPLAS, 2016).

3. Familiaridade com Ferramentas e Tecnologias

O domínio de ferramentas e tecnologias específicas é importante para trabalhar com eficiência na área de análise de dados. Exemplos incluem o uso de bibliotecas como Pandas e NumPy para manipulação de dados em Python, o uso de softwares como Tableau e Power BI para visualização de dados e o conhecimento de bancos de dados relacionais e não relacionais (KELLEHER; MAC NAMEE & D'ARCY, 2015).

4. Capacidade de Comunicação e Storytelling de Dados

Um analista de dados eficaz deve ser capaz de comunicar seus insights e resultados de forma clara e concisa. A habilidade de contar histórias com dados, por meio de visualizações e relatórios, é crucial para transmitir informações de maneira impactante e compreensível para diferentes públicos (CAIRO, 2016).

5. Pensamento Analítico e Resolução de Problemas

Um analista de dados deve possuir uma abordagem analítica para identificar e resolver problemas complexos. A capacidade de decompor problemas em etapas menores, realizar análises exploratórias e aplicar métodos estatísticos e algoritmos apropriados são competências essenciais nessa área (DAVENPORT & KIM, 2013).

Desafios de um Cientista de Dados

Diante do cenário já apresentado até o momento, veremos agora os principais desafios que um cientista de dados pode enfrentar no curso de sua profissão.

1. Dados Não Estruturados e Grandes Volumes

Um dos principais desafios enfrentados pelos cientistas de dados é lidar com dados não estruturados, como imagens, áudios e textos. Esses tipos de dados requerem técnicas avançadas de processamento e análise, como aprendizado profundo (*deep learning*) e processamento de Linguagem Natural (NLP) (GOODFELLOW; BENGIO & COURVILLE, 2016). Além disso, o volume de dados gerados diariamente pode ser esmagador, exigindo soluções escaláveis de armazenamento e processamento.

2. Limpeza e Qualidade dos Dados

A qualidade dos dados é fundamental para obter resultados precisos e confiáveis. Os cientistas de dados frequentemente enfrentam o desafio de lidar com dados incompletos, inconsistentes e ruidosos. A etapa de limpeza e preparação dos dados é demorada e requer atenção cuidadosa para garantir que os resultados da análise sejam confiáveis e relevantes (MANNING; RAGHAVAN & SCHÜTZE, 2008).

3. Privacidade e Segurança dos Dados

À medida que as empresas coletam e armazenam cada vez mais dados, a privacidade e a segurança se tornam preocupações cruciais. Os cientistas de dados precisam garantir que as informações confidenciais dos usuários estejam protegidas durante todo o processo de análise. Além disso, é necessário estar em conformidade com as regulamentações de proteção de dados em vigor.

4. Interpretação e Comunicação dos Resultados

Transformar dados em *insights* é apenas metade do desafio. A tarefa de interpretar os resultados da análise e comunicá-los de forma clara e compreensível para diferentes públicos é crucial para o sucesso de um cientista de dados. A habilidade de contar histórias com dados e criar visualizações impactantes é essencial para que os *insights* sejam efetivamente utilizados pelos tomadores de decisão.

5. Aprendizado Contínuo e Atualização Tecnológica

A ciência de dados é uma área em constante evolução, com novas técnicas, ferramentas e algoritmos surgindo regularmente. O cientista de dados deve estar disposto a aprender continuamente e acompanhar as tendências tecnológicas para se manter relevante e competitivo no mercado de trabalho.

6. Comunicação Interpessoal

Compreender o negócio é um fator crítico para o sucesso de um Cientista de Dados. Não basta apenas dominar habilidades técnicas como comunicação, coleta e preparação de dados, além do conhecimento em tecnologias e ferramentas. É fundamental que o profissional tenha uma sólida compreensão das regras e do contexto do negócio em que está atuando.

Por onde começar? Passos para ser um cientista de dados

Se você está aqui, te garanto que o primeiro passo já foi dado. Mas o que mais é necessário para continuar caminhando em direção a essa profissão tão relevante? Vamos conversar um pouco sobre isso.

Primeiramente, você precisará desenvolver as habilidades que já comentamos anteriormente. Verifique dentre os pontos que já citamos quais você já possui algum conhecimento ou prática e veja quais precisam ser descobertos por você nessa nova jornada. Pense e avalie com cuidado aqueles que precisam de mais esforço e atenção da sua parte e trabalhe para que suas habilidades sejam desenvolvidas. Este será seu primeiro desafio.

Em segundo lugar, tenha um foco especial para as habilidades técnicas. *Machine Learning* e linguagens de programação são assuntos muito densos, e exigirão esforço da sua parte para ter domínio sobre eles. Mas não se assuste, faça progresso no seu ritmo e vá aprendendo esses temas com pessoas qualificadas, tentativa e erro e leia materiais instrutivos. À medida que você for se ambientando com esses temas eles serão mais fáceis para você.

E por último e não menos importante: faça projetos, leia livros, blogs e conteúdos relacionados a *Data Science* e participe de comunidades. A maioria das ferramentas necessárias para um cientista de dados é gratuita, como informamos anteriormente, então desafie a si mesmo criando projetos. Isso colocará em prática todos os conteúdos que você já estudou e fará seu conhecimento ganhar forma.

Além disso, sempre busque novas fontes de conhecimento, como livros e blogs. A área da tecnologia está sempre em constante movimentação e por isso acompanhar as novidades é uma boa ideia. Faça parte de comunidades. Você verá que nas comunidades sempre existem pessoas dispostas a ajudar os outros em seus desafios e estão sempre compartilhando vários conhecimentos úteis entre si.

> Capítulo 2

Capítulo 2. Fundamentos da Ciência de Dados e Análise de Dados

Dados, informação e conhecimento

No primeiro capítulo falamos e exemplificamos como os dados são criados, suas formas e seus tipos. Agora vamos aprofundar um pouco mais no conceito dos dados e a importância que ele possui na tomada de decisão.

Conceito de dados

Dados são os registros soltos, aleatórios e sem qualquer análise. São informações não tratadas que ainda não apresentam relevância. São códigos que isoladamente não possuem nenhum significado, mas quando agrupados podem transmitir uma mensagem, representar algo ou até mesmo um conhecimento. Podemos definir “dado” como uma sequência de símbolos quantificados ou quantificáveis. Os dados são tudo aquilo que pode ser quantificado, como, por exemplo, imagens, sons, textos ou animações. Na Figura 3, temos uma exemplificação dessa ideia.

Figura 3 – Exemplo de dados.

O que podemos dizer dessa imagem?

1. São bolas de golfe.
2. As bolas são de cor branca.
3. Percebemos que elas já foram usadas.

4. São do mesmo tamanho.

5. Possuem um relevo na superfície da bola.

6. Possuem identificações diferentes nas bolas.

Podemos identificar várias características observando essa imagem. No entanto, imagine que essa imagem não foi disponibilizada. Apenas foi disponibilizado um dado, como por exemplo, um número ou uma cor. Perceba que não é possível saber o que ele significa ou o que ele representa, pode ser algo muito relevante ou pode ser nada. Porém, no momento que existir uma agregação com outro dado, ele passa a ser ou não uma informação. No exemplo ilustrado na Figura 5, podemos identificar os dados contidos na imagem, agregá-los, interpretá-los e, assim, obter uma informação. Os dados podem ser do tipo numérico, textual, data e hora, bits e vários outros.

Conceito de Informação

Dizemos que a informação é o dado estruturado ou organizado que possui algum sentido. A informação é a matéria-prima utilizada para o conhecimento, ela traz significado e compreensão sobre um determinado assunto ou situação. Se os dados agregados fazem sentido para quem o lê, então dizemos que existe um valor naquela informação, e é por meio da informação que podemos tomar decisões.

1. Requer unidade de análise.

2. Exige consenso em relação ao significado.

3. Exige necessariamente a mediação humana.

Conceito de conhecimento

Conhecimento é a informação processada e transformada em experiência pelo indivíduo. Ou seja, é o resultado de várias informações organizadas de forma lógica. O conhecimento é a capacidade, adquirida por alguém, de interpretar e operar sobre um conjunto de Informações. Se informação é o dado trabalhado, então o conhecimento é informação trabalhada.

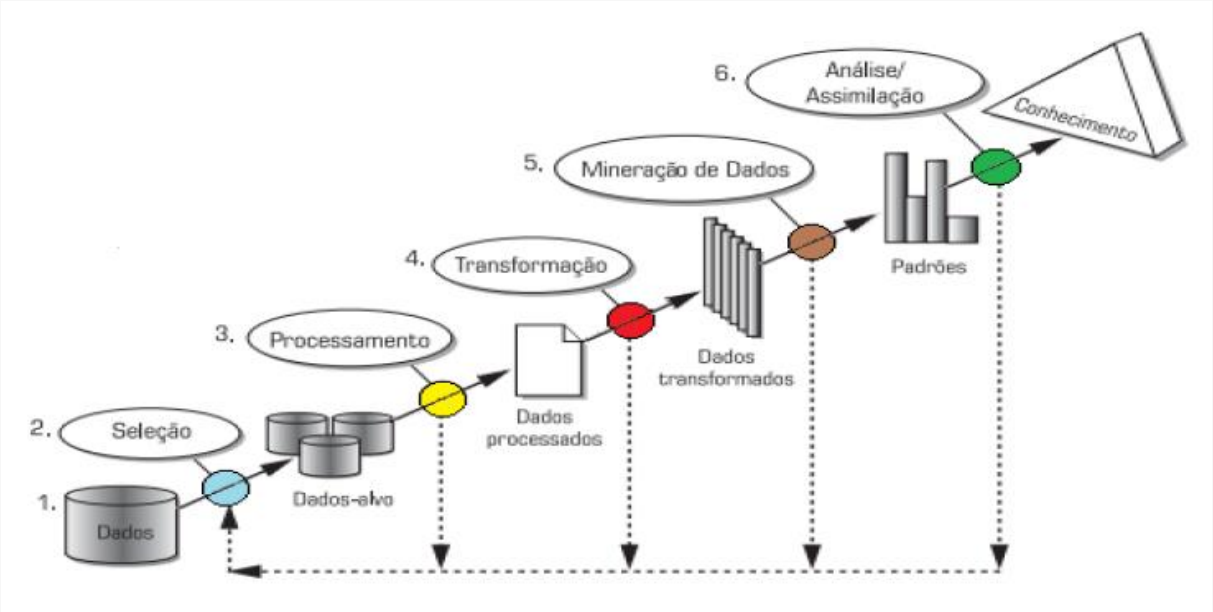
Decisões baseada em dados

Agora que entendemos a importância de conhecer bem os dados e gerar conhecimento sobre eles, o próximo passo é tomar decisões baseada no conhecimento adquirido através da análise dos dados. Mas como podemos tomar decisões baseada nos dados? Para isso precisamos entender o processo de descoberta de conhecimento, só assim vamos ter *insights* suficientes sobre o negócio e tomar as decisões assertivas.

KDD

O processo de descoberta de conhecimento, é ilustrado pela Figura 6, que envolve diversas fases do KDD (*Knowledge Discovery Databases*) proposta por Fayyad (FAYYAD et al., 1996). O objetivo é extrair, de grandes bases de dados, informações válidas e acionáveis, desconhecidas, úteis para tomada de decisão. De uma forma breve, o processo envolve três etapas iniciais: seleção, pré-processamento e transformação, as quais compõem o que é denominado de preparação dos dados. Após essas etapas vem a fase de mineração de dados e, por fim, o conhecimento gerado deverá ser analisado, o que acontece na etapa de análise e assimilação dos resultados (COLAÇO JR., 2004).

Figura 4 – Processo de descoberta de conhecimento.



Fonte: Colaço Jr. 2004

- Seleção de dados: são identificadas as bases de dados a serem utilizadas nas descobertas de conhecimento, leva-se em consideração os objetivos do processo.

- Pré-processamento de dados: como a informação pode vir de diversas bases distintas, podem surgir problemas de integração entre os dados. Isso deve ser resolvido nessa etapa. Por exemplo: suponha que que a informação sobre o sexo dos clientes de uma loja esteja armazenada em um banco como “M” e “F” e em outra, como “H” e “M”. Neste caso o pré-processamento é feito corrigindo e atualizando os dados.

- Transformação de dados: o objetivo é transformar os dados já pré-processados, de modo a torná-los compatíveis com as entradas de diversos algoritmos de mineração existentes.

- Mineração de dados: caracteriza pela escolha e aplicação do algoritmo e da técnica de mineração.

- Análise e assimilação dos resultados: o conhecimento gerado deve ser analisado nesta etapa, de maneira a verificar se é realmente útil para a tomada de decisão. Caso a resposta não for satisfatória, então deve-se repetir todo ou parte do processo de KDD.

Data Mining

O termo mineração de dados vem do inglês *Data Mining* (DM) que tem o objetivo principal um processo mais amplo denominado descoberta de conhecimento em base de dados. A mineração de dados consiste em utilizar dados de estatísticas e de inteligência artificial bem estabelecidas que constroem modelos que predizem os padrões relevantes em um banco de dados. O DM identifica e interpreta padrões de dados que serão utilizados pelos gestores na tomada de decisão (COLAÇO JR. 2004). Embora KDD e *Data Mining* sejam frequentemente entendidos como sinônimos, é importante frisar que, enquanto o KDD compreende todas as etapas para a descoberta do conhecimento a partir da existência de dados, a mineração de dados é apenas e tão somente uma das etapas do processo.

Tipo de dados

Em ciência de dados e Big Data você encontrará muitos tipos diferentes de dados, e cada um deles tende a exigir ferramentas e técnicas diferentes. As principais categorias de dados são: estruturados, não estruturados e semiestruturados.

Dados estruturados

Refere-se a todos os dados que estejam em conformidade com um determinado formato. Tem como característica ser bem-definidos, inflexíveis e pensados antes da própria criação dos dados. Dessa forma, não é possível que tipos de dados diferentes das estruturas preestabelecidas sejam carregados.

Por exemplo, se uma coluna de uma tabela for criada com o tipo de dado numérico, essa coluna não aceitará dados textuais. Um exemplo básico são as planilhas, onde geralmente há linhas e colunas que seguem um determinado padrão.

Os dados estruturados sempre são claramente organizados e mais fáceis de analisar. Em uma planilha, por exemplo, você conseguiria facilmente indicar valores e quantidades listadas, por essa razão muitos dos dados com os quais as organizações trabalham podem ser categorizados como estruturados (ELMASRI; SHAMKANT, 2019).

Figura 5 – Representação dos dados estruturados.

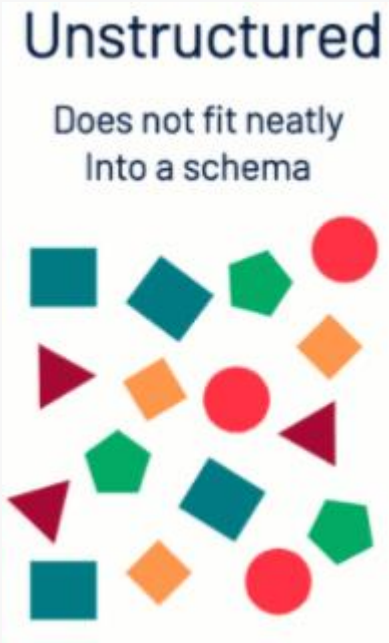
Dados não estruturados

Os dados não estruturados são o oposto dos dados estruturados. Eles não possuem uma estrutura predefinida, alinhada ou padronizada. Os dados não estruturados se caracterizam por possuir uma estrutura flexível e dinâmica ou, até mesmo, nenhuma estrutura. Esses dados podem ser compostos por vários elementos diferentes, como: imagens, áudios, vídeos, gráficos e textos. Eles são difíceis de processar devido a sua complexibilidade e formatação. Os dados não estruturados podem ser

encontrados em mídias sociais, e-mails, fotos, vídeos, *chats*, arquivos de logs, sensor de *IoT*, entre outros.

Hoje os dados não estruturados são os mais difundidos, alcançando cerca de 90% do total dos dados produzidos. Por isso, muitas organizações têm lutado para tentar entender esses dados principalmente para usá-los em estratégias e ideias em seus negócios. Nesse ponto, a Inteligência Artificial tem um grande papel na análise dos dados, já que as análises conterão vídeos, postagens em mídia social, fotografias, *e-mails*, arquivos de áudio e imagens.

Figura 6 – Representação dos dados não estruturados.



Dados semiestruturados

Os dados semiestruturados se encaixam entre as duas definições anteriores. Eles não residem em uma tabela formatada, porém possuem um certo nível de organização. Esses dados possuem uma estrutura heterogênea, não sendo uma estrutura completamente rígida e nem exclusivamente flexível. Um exemplo desse nível de organização é o código

HTML, onde você consegue extrair muitas informações dentro de uma forma específica de expressar os dados.

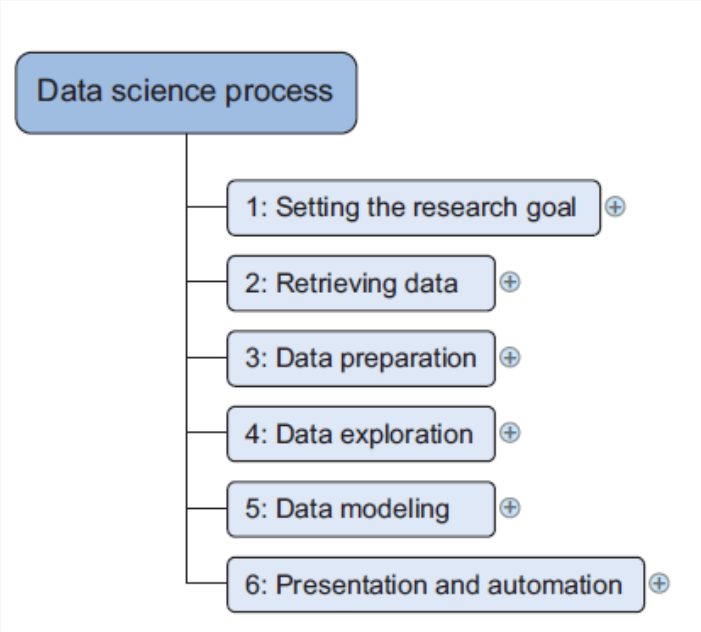
Em muitos casos os dados dispõem de uma definição regular (por exemplo um catálogo de produtos), em outros, um padrão estrutural que pode ser identificado ou não existem informações descritivas relacionadas (por exemplo, um arquivo de imagem) (ABITEBOUL S., 1997).

Figura 7 – Representação dos dados não estruturados.

Etapas de processamento da ciência de dados

Seguir uma abordagem estruturada para ciência de dados ajuda você a maximizar suas chances de sucesso em um projeto de ciência de dados com o menor custo. Também torna possível assumir um projeto como uma equipe, com cada membro da equipe focando no que faz de melhor. Mas tome cuidado: essa abordagem pode não ser adequada para todo tipo de projeto ou ser a única maneira de fazer uma boa ciência de dados. O processo típico de ciência de dados consiste em seis etapas pelas quais você irá iterar. A Figura 8 ilustra as etapas de processamento da ciência de dados.

Figura 8 – Etapas de processamento da ciência de dados.



Fonte: Livro ciências de dados, *datamining* etc.

Definindo o objetivo da pesquisa

A ciência de dados é aplicada principalmente no contexto de uma organização. Quando a empresa solicitar que você execute um projeto de ciência de dados, você primeiro preparará um termo de abertura do projeto. Esta carta contém informações como o que você vai pesquisar, como a empresa se beneficia disso, quais dados e recursos você precisa, um cronograma e entregas.

Recuperando dados

O segundo passo é coletar dados. Você declarou no termo de abertura do projeto quais dados você precisa e onde pode encontrá-los. Nesta etapa, você garante que pode usar os dados em seu programa, o que significa verificar a existência, a qualidade e o acesso aos dados. Os dados também podem ser fornecidos por empresas terceirizadas e assumem muitas formas, desde planilhas do Excel até diferentes tipos de bancos de dados (DAVY CIELEN et al.).

Preparação de dados

A coleta de dados é um processo propenso a erros. Nesta fase você aprimora a qualidade dos dados e os prepara para uso nas etapas subsequentes, que consiste em três subfases: limpeza de dados, removendo valores falsos de uma fonte de dados e inconsistências entre fontes de dados; integração de dados, enriquecendo as fontes de dados combinando informações de várias fontes de dados; e transformação de dados, garantindo que os dados estejam em um formato adequado para uso em seus modelos (DAVY CIELEN et al.).

Exploração de dados

A exploração de dados está preocupada em construir uma compreensão mais profunda de seus dados. Você tenta entender como as variáveis interagem umas com as outras, a distribuição dos dados e se existem discrepâncias. Para conseguir isso, você usa principalmente estatísticas descritivas, técnicas visuais e modelagem simples. Essa etapa geralmente é conhecida pela abreviação EDA, para Análise Exploratória de Dados (DAVY Cielen et al.).

Modelagem de dados ou construção de modelos

Nesta fase você usa modelos, conhecimento de domínio e *insights* sobre os dados encontrados nas etapas anteriores para responder à pergunta de pesquisa. Você seleciona uma técnica das áreas de estatística, aprendizado de máquina, pesquisa operacional e assim por diante. A construção de um modelo é um processo interativo que envolve a seleção das variáveis para o modelo, a execução do modelo e o diagnóstico do modelo (DAVY CIELEN et al.).

Apresentação e automação

Por fim, você apresenta os resultados para o seu negócio. Esses resultados podem assumir muitas formas, desde apresentações até relatórios de pesquisa. Às vezes, você precisará automatizar a execução do processo porque a empresa desejará usar os insights obtidos em outro projeto ou permitir que um processo operacional use o resultado do seu modelo (DAVY CIELEN et al.).

Data Driven

A cultura de *Data Driven* consiste em adotar estratégias e tomar decisões baseadas na análise de informações e não em intuições ou simples experiências. Essa cultura não é como uma ferramenta que pode ser utilizada em alguns momentos, mas uma metodologia bem estruturada que permite que as organizações tenham ideias mais precisas sobre seus negócios e assim elas são capazes de aproveitar melhor as oportunidades. Um dos pilares da cultura é realmente excluir quaisquer influências pessoais ou externas e basear as ações e estratégias nos dados que são apresentados. Dessa forma, o índice de assertividade se torna bastante elevado, embora o conceito possa soar um tanto impessoal.

Um dos objetivos do *Data Driven* é coletar dados de diversas fontes, tanto internas quanto externas, e cruzar as informações de forma a obter um panorama mais claro sobre o mercado e a própria instituição. O *Data Driven* surgiu como um tipo de extensão da ciência de dados, utilizando os métodos científicos e os algoritmos transformando dados em conhecimento. Uma das principais diferenças entre empresas que aderem ao *Data Driven* e as que optam pelo modelo tradicional é o uso de dados de forma integrada em seus processos e operações. Os dados geralmente ficam em nuvem e não em servidores particulares. Dessa forma, todos os envolvidos possuem acesso às informações a qualquer instante.

O resultado fica ligado à inteligência coletiva e não apenas na produtividade dos colaboradores de forma individual. Isso confere maior agilidade na rotina e maior propensão de avanço. Por se tratar de uma mudança profunda de rotina, é primordial ter profissionais capacitados e especializados nesse assunto para que consigam trazer essa transformação para o cotidiano. O *Chief Data Officer* (CDO) é um exemplo, já que ele é o responsável por liderar as mudanças dentro da empresa e trazer um novo *mindset* aos colaboradores e terá ao seu lado os cientistas de dados, que já são profissionais que se relacionam diretamente com as informações para sugerir ou indicar melhorias e resultados.

Os dois últimos pontos principais do *Data Driven* são fundamentais para que todo o resto ocorra bem: dados e tecnologia. É de suma importância ter dados organizados, acessíveis e integrados para que o processo caminhe como deve. Esse é o ponto que irá conceder aos profissionais aquilo que é necessário para extraírem o máximo de tudo que estiver disponível. A tecnologia será a parte responsável por gerar soluções eficientes que irão sustentar toda a nova cultura organizacional. Com a tecnologia será possível gerar ferramentas eficientes e pensadas para etapas e processos específicos dentro da organização, seja na gestão ou nas atividades de análise comuns, e isso gerará um negócio sustentável a longo prazo.

> Capítulo 3

Capítulo 3. Fundamentos da Big Data

A origem dos dados do Big Data

A origem dos dados que compõem o Big Data pode ser gerada por diversas fontes. Telefones celulares, mídias sociais, inteligências artificiais, tecnologias de imagem e muitos outros produzem dados que precisam ser armazenados em algum lugar e servir a alguma finalidade. Quando falamos em Big Data, nos referimos não apenas aos dados específicos gerados por uma amostra tímida de usuários, mas principalmente sobre a enorme quantidade de dados produzidos e armazenados diariamente e que não estão necessariamente padronizados, e esse é um dos grandes desafios encarados atualmente. Quando pensamos na geração dos dados em si, temos os principais fatores em:

Dados gerados por pessoas

São dados criados pela atividade humana na tecnologia, como:

1. Postagens nas redes sociais.
2. Mensagens enviadas em aplicativos.
3. Textos escritos em blogs, revistas ou páginas da *web*.
4. Áudios ou vídeos compartilhados.
5. E-mails e afins.

As redes sociais merecem destaque no quesito produção de dados. Cada vez que um indivíduo posta qualquer informação, compartilha *links*, realiza compras ou classifica um conteúdo, ele está gerando incontáveis dados que serão utilizados dentro do Big Data.

Dados gerados por máquinas

São aqueles criados a partir das máquinas e sem a necessidade da intervenção humana, pois já são programados para extrair tais dados, como por exemplo:

- Sensores em veículos, eletrodomésticos e máquinas industriais.
- Câmeras e sistemas de segurança.
- Satélites.
- Dispositivos médicos.
- Ferramentas pessoais, como aplicativos de smartphone e afins.

Dados gerados por empresas

São dados gerados por empresas: aqueles que as organizações obtêm à medida que administram seus negócios, como por exemplo: Registros gerados toda vez que você faz uma compra em uma loja *online* ou física – registros como números exclusivos de clientes, os itens que você comprou, a data e hora em que você comprou os itens e quantos de cada item você comprou.

No mundo do Big Data você verá esse dado ser chamado de “dado transacional”. O que precisamos ter em mente é que nem todo dado gerado é necessariamente considerado Big Data. O fato de enviar mensagens de texto ao longo do dia, por exemplo, só poderia ser considerado Big Data caso o número esteja perto dos milhões.

Benefícios e usos da ciência de dados e do *Big Data*

A ciência de dados e o *Big Data* são usados em quase todos os lugares, tanto em ambientes comerciais quanto não comerciais. Empresas

comerciais em quase todos os setores usam ciência de dados e Big Data para obter *insights* sobre seus clientes, processos, equipe, conclusão e produtos. Muitas empresas usam a ciência de dados para oferecer aos clientes uma melhor experiência do usuário, bem como para fazer vendas cruzadas, vendas adicionais e personalizar suas ofertas. Um bom exemplo disso é o *Google AdSense*, que coleta dados de usuários da *Internet* para que mensagens comerciais relevantes possam ser correspondidas à pessoa que navega na Internet.

Outro exemplo é a publicidade personalizada em tempo real. Profissionais de recursos humanos usam análise de pessoas e mineração de texto para selecionar candidatos, monitorar o humor dos funcionários e estudar redes informais entre colegas de trabalho. *People analytics* é o tema central do livro *Moneyball: The Art of Winning an Unfair Game*. No livro (e no filme) vimos que o processo tradicional de observação do beisebol americano era aleatório, e substituí-lo por sinais correlacionados mudou tudo. Confiar nas estatísticas permitiu que eles contratassem os jogadores certos e os colocassem contra os adversários onde teriam a maior vantagem.

As instituições financeiras usam a ciência de dados para prever os mercados de ações, determinar o risco de emprestar dinheiro e aprender como atrair novos clientes para seus serviços. Atualmente, pelo menos 50% dos negócios em todo o mundo são realizados automaticamente por máquinas baseadas em algoritmos desenvolvidos por *quants*, como são chamados os cientistas de dados que trabalham em algoritmos de negociação, com a ajuda de big data e técnicas de ciência de dados.

As organizações governamentais também estão cientes do valor dos dados. Muitas organizações governamentais não apenas contam com cientistas de dados internos para descobrir informações valiosas, mas também compartilham seus dados com o público. Você pode usar esses dados para obter insights ou criar aplicativos orientados por dados. Um

cientista de dados em uma organização governamental trabalha em diversos projetos, como detectar fraudes e outras atividades criminosas ou otimizar o financiamento de projetos. Um exemplo bem conhecido foi fornecido por Edward Snowden, que vazou documentos internos da Agência de Segurança Nacional Americana e da Sede de Comunicações do Governo Britânico que mostram claramente como eles usaram ciência de dados e big data para monitorar milhões de indivíduos. Essas organizações coletaram 5 bilhões de registros de dados de aplicativos difundidos, como *Google Maps*, *Angry Birds*, *e-mail* e mensagens de texto, entre muitas outras fontes de dados. Em seguida, eles aplicaram técnicas de ciência de dados para destilar informações (DAVY CIELEN et al.).

As organizações não governamentais (ONGs) também não são estranhas ao uso de dados. Eles o usam para arrecadar dinheiro e defender suas causas. O *World Wildlife Fund* (WWF), por exemplo, emprega cientistas de dados para aumentar a eficácia de seus esforços de captação de recursos. As universidades usam a ciência de dados em suas pesquisas, mas também para aprimorar a experiência de estudo de seus alunos. A ascensão dos cursos online abertos massivos (MOOC) produz muitos dados, o que permite que as universidades estudem como esse tipo de aprendizado pode complementar as aulas tradicionais. Os MOOCs são um ativo inestimável se você quiser se tornar um cientista de dados e um profissional de *Big Data* (DAVY CIELEN et al.).

Os 5 Vs do Big Data: Compreendendo as Dimensões Fundamentais

Os "5 Vs do Big Data" são um conjunto de características que definem a natureza única e desafiadora dos dados em larga escala (HAN, HAIHONG & LE, 2011). Esses “Vs” representam as dimensões essenciais que os cientistas de dados e profissionais de análise precisam considerar ao lidar com conjuntos de dados extensos e complexos.

Volume

O primeiro V, Volume, refere-se à imensa quantidade de dados gerados a uma taxa exponencial. Essa característica é uma das razões pelas quais o termo "big" é utilizado. Dados volumosos podem surgir de várias fontes, como redes sociais, sensores, transações comerciais e muito mais.

Velocidade

O segundo V, Velocidade, destaca a velocidade vertiginosa com que os dados são gerados e precisam ser processados. Isso é especialmente relevante em cenários de processamento em tempo real, como análises de tráfego de internet e transações financeiras.

Variedade

O terceiro V, Variedade, refere-se à diversidade de tipos e formatos de dados. Além dos dados estruturados, como tabelas em bancos de dados, há também dados não estruturados, como textos, imagens, áudios e vídeos. Lidar com essa variedade exige abordagens flexíveis de armazenamento e análise.

Veracidade

O quarto V, Veracidade, diz respeito à confiabilidade e qualidade dos dados. Com a ampla gama de fontes de dados, é fundamental garantir que os dados sejam precisos, confiáveis e livres de distorções. A veracidade é crucial para tomadas de decisão precisas.

Valor

O último V, Valor, representa o objetivo final da coleta e análise de big data. O verdadeiro valor reside na capacidade de transformar esses dados em informações acionáveis, que impulsionam a tomada de decisões informadas, identificam oportunidades de negócios e melhoram processos.

Os 5 Vs do Big Data fornecem um *framework* abrangente para abordar os desafios e oportunidades únicos apresentados pelo processamento de dados em larga escala. Ao considerar volume, velocidade, variedade, veracidade e valor, os profissionais podem desenvolver estratégias eficazes para extrair insights significativos dos dados e aproveitar seu potencial máximo.

Armazenamento de Dados no Big Data

No contexto do Big Data, existem várias formas de armazenamento de dados, cada uma adequada para diferentes tipos de informações e necessidades específicas. Algumas das principais formas de armazenamento de dados no contexto do Big Data incluem:

Data Lake: é uma estrutura que permite armazenar grandes volumes de dados em seu formato bruto original, incluindo dados estruturados, não estruturados e semiestruturados. Essa abordagem é flexível e escalável, permitindo a captura de dados de diversas fontes sem a necessidade de estruturá-los previamente. O Data Lake é ideal para análises avançadas e descoberta de insights valiosos.

Data Warehouse: é uma estrutura centralizada que armazena dados estruturados, processados e otimizados para consultas analíticas. Essa abordagem é ideal para fornecer uma visão consolidada do negócio, suportar relatórios gerenciais e fornecer respostas rápidas a perguntas predefinidas.

Data Mart: é uma versão especializada e menor do Data Warehouse, que atende a necessidades específicas de um departamento ou equipe dentro da organização. O Data Mart fornece dados relevantes e personalizados para uma área específica, permitindo acesso rápido às informações relevantes para suas atividades diárias.

Banco de Dados NoSQL: os bancos de dados NoSQL (Not Only SQL) são projetados para lidar com grandes volumes de dados não estruturados

e semiestruturados. Eles oferecem escalabilidade horizontal e flexibilidade para lidar com diferentes tipos de informações, tornando-os adequados para cenários de Big Data.

Armazenamento em Nuvem: a computação em nuvem oferece soluções de armazenamento altamente escaláveis e flexíveis para o Big Data. As plataformas de armazenamento em nuvem permitem que as organizações paguem apenas pelo uso, tornando-as uma opção econômica para o gerenciamento de grandes volumes de dados.

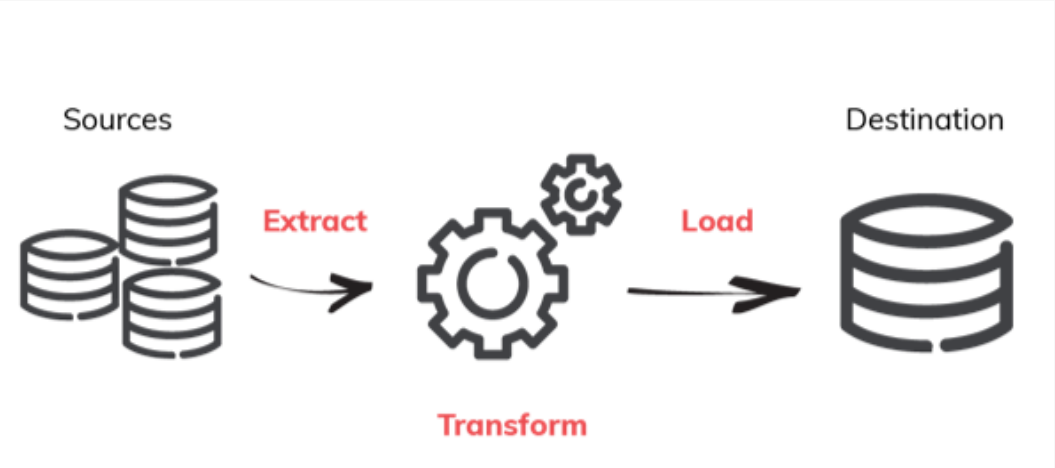
Técnicas de Integração de Dados

A integração de dados é uma etapa crucial no ecossistema do Big Data, que envolve a coleta, transformação e carga de informações de diversas fontes em um único repositório, permitindo a análise holística e abrangente dos dados. Duas abordagens amplamente utilizadas na integração de dados no contexto do Big Data são ETL (Extração, Transformação e Carga) e ELT (Extração, Carga e Transformação).

ETL (Extração, Transformação e Carga):

No processo ETL, os dados são extraídos de várias fontes, como bancos de dados, arquivos, APIs e aplicativos, e em seguida passam por um estágio de transformação, onde são limpos, organizados, enriquecidos e padronizados de acordo com as necessidades da análise. Posteriormente, os dados transformados são carregados em um repositório de destino, como um Data Warehouse ou Data Mart. O ETL é ideal para cenários em que a transformação dos dados precisa ser feita antes do carregamento no repositório de destino, garantindo a consistência e qualidade dos dados para análise. A imagem 09 ilustra a representação da técnica do processo do ETL.

Figura 9 – Representação da técnica do processo do ETL.



Vantagens

Qualidade dos Dados: o processo de transformação antes do carregamento permite garantir a qualidade, integridade e consistência dos dados, resultando em análises mais confiáveis e precisas.

Personalização: é possível personalizar a transformação dos dados de acordo com as necessidades específicas da análise, garantindo que os dados sejam adaptados às exigências do repositório de destino.

Redução da Carga no Repositório: ao realizar a transformação antes do carregamento, é possível reduzir a carga e o volume de dados no repositório de destino, tornando-o mais eficiente e de fácil manutenção.

Desvantagens

Demanda de Recursos: o processo ETL pode exigir recursos significativos, especialmente quando lida com grandes volumes de dados, tornando-o mais demorado em algumas situações.

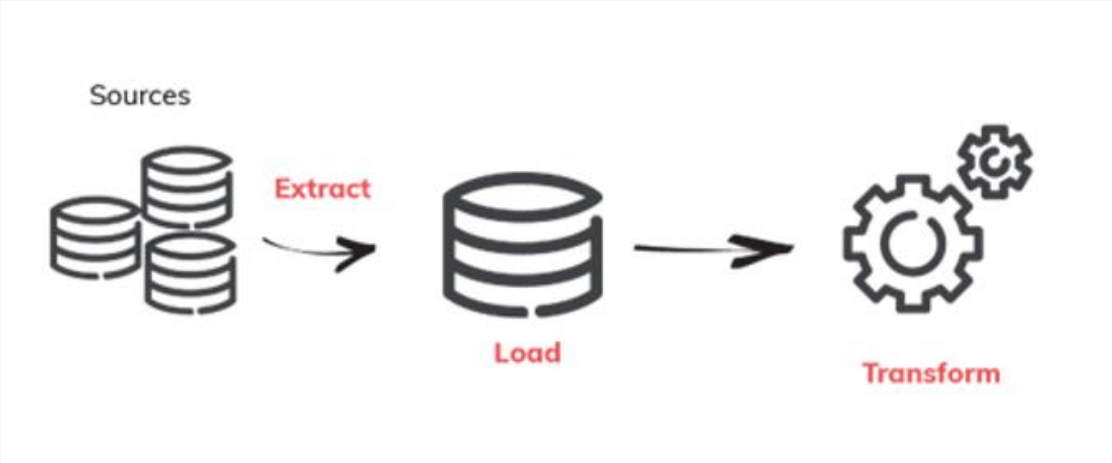
Dificuldade em Lidar com Dados Não Estruturados: a transformação pode ser mais complexa ao lidar com dados não estruturados ou semiestruturados, exigindo esforços adicionais para organizar e padronizar esses dados.

ELT (Extração, Carga e Transformação)

A abordagem ELT é semelhante ao ETL, mas com uma diferença significativa: os dados são extraídos de suas fontes e, em seguida, carregados diretamente no repositório de destino, como um Data Lake ou um banco de dados NoSQL. Em seguida, a transformação dos dados é realizada dentro do repositório, aproveitando a escalabilidade e o poder de processamento dos sistemas de armazenamento de dados. A abordagem ELT é mais adequada para casos em que os dados brutos precisam ser armazenados em sua integridade antes da transformação, permitindo uma análise flexível e exploratória posteriormente.

A imagem 10 ilustra a representação da técnica do processo do ELT

Figura 10 – Representação da técnica do processo do ELT.



Vantagens

Escalabilidade: o ELT permite tirar proveito da escalabilidade dos sistemas de armazenamento de dados, como Data Lakes e bancos de dados NoSQL, para processar grandes volumes de dados em paralelo.

Flexibilidade de Análise: o armazenamento dos dados brutos permite análises mais flexíveis e exploratórias, pois a transformação é

realizada posteriormente, facilitando a adaptação dos dados às necessidades da análise.

Velocidade de Carregamento: o ELT pode ser mais rápido no carregamento inicial dos dados, uma vez que eles são armazenados sem transformação prévia.

Desvantagens

Desafios de Qualidade dos Dados: como os dados brutos são carregados diretamente, pode haver menos garantias em relação à qualidade dos dados, exigindo esforços adicionais para limpeza e organização posteriormente.

Consumo de Espaço de Armazenamento: o armazenamento de dados brutos pode consumir mais espaço, especialmente quando não há uma seleção criteriosa dos dados carregados, o que pode levar a custos adicionais de armazenamentos e tecnologias adequadas para otimizar o processo.

No contexto do Big Data, ambas as abordagens são relevantes e suas escolhas dependem de vários fatores, como o volume de dados, a velocidade de processamento, a complexidade das transformações necessárias e os objetivos da análise. A implementação bem-sucedida da integração de dados requer uma compreensão profunda das necessidades e dos requisitos específicos da organização, bem como a escolha das ferramentas.

> Capítulo 4

Capítulo 4. Fundamentos da Análise de Dados

Aplicações da Análise de dados

A análise de dados desempenha um papel vital em diversos setores, possibilitando a transformação de informações brutas em insights acionáveis para tomadas de decisão informadas e estratégias eficientes (HAN, HAIHONG & LE, 2011). A crescente geração e acumulação de dados em larga escala tornaram a análise de dados uma ferramenta indispensável para empresas, organizações e profissionais que buscam se destacar em um ambiente competitivo.

Identificação de Padrões e Tendências

A análise de dados permite identificar padrões e tendências que, muitas vezes, passariam despercebidos. Ao compreender as relações entre variáveis e eventos, as organizações podem antecipar mudanças no mercado, comportamento do consumidor e outros fatores cruciais para o sucesso.

Tomada de Decisão Informada

A tomada de decisão informada é fundamental para o crescimento sustentável e eficácia operacional. A análise de dados fornece insights baseados em evidências, permitindo que gestores e líderes tomem decisões mais confiantes e alinhadas com os objetivos estratégicos.

Otimização de Processos

A análise de dados permite identificar ineficiências e gargalos em processos internos e externos. Isso leva à otimização de operações, redução de custos e melhoria da qualidade dos produtos e serviços oferecidos.

Personalização e Experiência do Cliente

Com a análise de dados, é possível entender melhor as preferências e necessidades dos clientes. Isso permite personalizar ofertas e melhorar a experiência do cliente, aumentando a satisfação e fidelização.

Previsão e Planejamento Estratégico

A análise de dados oferece a capacidade de prever cenários futuros com base em dados históricos e padrões atuais. Isso é fundamental para um planejamento estratégico sólido e a tomada de medidas proativas.

Inovação e Novas Oportunidades

A análise de dados também é uma fonte de inovação, permitindo que empresas identifiquem novas oportunidades de negócios, desenvolvam produtos inovadores e explorem mercados emergentes.

Tipos de Análise de Dados

A análise de dados abrange diversas abordagens que se adaptam a diferentes objetivos e contextos (HAN, HAIHONG & LE, 2011). Cada tipo de análise possui características únicas que atendem a necessidades específicas, fornecendo informações cruciais para tomadas de decisão informadas e estratégias eficientes.

Análise Descritiva

A análise descritiva é o ponto de partida, envolvendo a exploração e resumo das principais características dos dados. Por meio de estatísticas de tendência central, dispersão e visualizações, essa abordagem fornece uma compreensão inicial dos dados e identifica padrões evidentes.

Análise Exploratória

A Análise Exploratória de Dados (AED) é uma etapa crítica na análise de dados, onde o objetivo é explorar os dados de forma mais profunda e detalhada para obter insights e padrões que podem não ser evidentes à primeira vista. A AED é uma abordagem flexível e criativa, que envolve o uso de técnicas estatísticas e gráficas para descobrir informações valiosas nos dados.

Alguns dos principais objetivos e técnicas utilizadas na análise exploratória de dados incluem:

Identificar Padrões e Relações: a AED permite identificar padrões, tendências e relações entre variáveis que podem não ser óbvios em uma análise descritiva inicial. Por exemplo, ao explorar dados de vendas de uma empresa, a análise exploratória pode revelar sazonalidades ou correlações entre vendas de diferentes produtos.

Encontrar Outliers e Anomalias: a análise exploratória é útil para detectar valores extremos (*outliers*) ou dados incomuns que podem afetar a análise. Identificar e entender esses outliers é importante para entender melhor o comportamento dos dados.

Análise de Correlação: a AED pode incluir a análise de correlações entre diferentes variáveis, ajudando a entender quais variáveis estão relacionadas entre si. Por exemplo, em um conjunto de dados de saúde, podemos explorar se existe correlação entre a quantidade de exercícios físicos realizados e os níveis de colesterol no sangue.

Visualização Avançada: gráficos mais complexos e avançados são utilizados na AED para representar a distribuição e relação entre variáveis de forma mais detalhada. Gráficos como *scatter plots, heatmaps, box plots e pair plots* são comuns na análise exploratória.

Análise de Clusterização: a análise exploratória pode incluir técnicas de clusterização para agrupar dados semelhantes em clusters ou grupos distintos. Isso pode revelar padrões de segmentação de clientes, por exemplo.

Análise de Componentes Principais: outra técnica comum na AED é a Análise de Componentes Principais (PCA), que ajuda a reduzir a dimensionalidade dos dados e encontrar as principais características que explicam a maior parte da variância dos dados.

Análise Diagnóstica

A análise diagnóstica busca entender as relações causais entre variáveis, identificando as razões por trás dos padrões observados. Ao examinar as correlações e relações, essa abordagem ajuda a identificar fatores que influenciam comportamentos e eventos.

Análise Preditiva

A análise preditiva envolve a criação de modelos estatísticos e algoritmos de aprendizado de máquina para prever eventos futuros com base em dados históricos. Essa abordagem é crucial para antecipar tendências e tomar medidas proativas.

Análise Prescritiva

A análise prescritiva vai além da previsão, recomendando ações específicas a serem tomadas para atingir os resultados desejados. Usando algoritmos avançados, essa abordagem oferece orientações para otimizar decisões e estratégias.

> Capítulo 5

Capítulo 5. Fundamentos de Banco de Dados SQL e NoSQL

Introdução aos Bancos de dados SQL e NoSQL

A gestão de dados é uma parte essencial da ciência de dados e da análise de informações. Os bancos de dados são ferramentas fundamentais para armazenar, organizar e recuperar dados de maneira eficiente (HAN, HAIHONG & LE, 2011). Duas categorias principais de bancos de dados são os bancos de dados SQL (*Structured Query Language*) e os bancos de dados NoSQL (Not Only SQL), cada um projetado para atender a diferentes necessidades e desafios.

Os bancos de dados SQL, também conhecidos como bancos de dados relacionais, têm sido amplamente utilizados desde a década de 1970. Eles são baseados em um modelo de dados relacional, onde as informações são organizadas em tabelas com linhas e colunas. O SQL é a linguagem padrão utilizada para realizar consultas e operações nesses bancos de dados. Exemplos populares de sistemas de gerenciamento de bancos de dados SQL incluem o MySQL, PostgreSQL, Oracle e Microsoft SQL Server.

Esses bancos de dados oferecem uma estrutura rígida, garantindo a integridade dos dados por meio de chaves primárias, chaves estrangeiras e restrições de integridade referencial. Eles são ideais para cenários em que a consistência e a precisão dos dados são primordiais, como sistemas financeiros, sistemas de gerenciamento de recursos humanos e sistemas de controle de estoque.

Por outro lado, com o advento da *web* 2.0 e das aplicações de larga escala, surgiu a necessidade de lidar com enormes quantidades de dados e com uma alta demanda por escalabilidade e flexibilidade. Nesse contexto, os bancos de dados NoSQL se destacaram como uma alternativa viável.

Os bancos de dados NoSQL adotam modelos de dados diferentes dos bancos SQL tradicionais. Eles podem ser baseados em documentos (*document-oriented*), famílias de colunas (*column-family*), pares de chave-valor (*key-value*) ou grafos (*graph*). Essa diversidade de modelos permite que os bancos de dados NoSQL sejam mais adaptáveis a diferentes tipos de dados e a casos de uso específicos.

Esses bancos de dados são projetados para oferecer alta escalabilidade, flexibilidade e velocidade de acesso aos dados. Eles são amplamente utilizados em aplicações I, mídias sociais, sistemas de análise de Big Data e Internet das Coisas (*IoT*).

Diferenças entre os Banco de Dados SQL e NoSQL

Estrutura de Dados

SQL: os bancos de dados SQL são baseados no modelo relacional, onde os dados são organizados em tabelas com linhas e colunas. Cada tabela tem um esquema fixo que define os nomes e tipos de colunas, garantindo a consistência dos dados. A relação entre as tabelas é estabelecida usando chaves primárias e chaves estrangeiras.

NoSQL: os bancos de dados NoSQL adotam diferentes modelos de dados, como documentos, pares de chave-valor, colunas ou grafos. Eles não têm um esquema rígido, permitindo a inserção de dados com estruturas diferentes em uma mesma coleção (ou tabela, dependendo do modelo).

Linguagem de Consulta

SQL: os bancos de dados SQL utilizam a linguagem SQL (*Structured Query Language*) para realizar operações como inserção, consulta, atualização e exclusão de dados. O SQL é uma linguagem declarativa, onde os usuários especificam o que desejam obter e o banco de dados determina como buscar essas informações.

NoSQL: cada tipo de banco de dados NoSQL possui sua própria API ou linguagem de consulta específica. Algumas podem usar consultas similares ao SQL, mas outras podem ser baseadas em métodos ou operações específicas do modelo de dados utilizado.

Escalabilidade

SQL: os bancos de dados SQL tradicionais são mais difíceis de escalar horizontalmente (adicionar mais servidores ao cluster) devido à necessidade de manter a integridade referencial e a consistência dos dados em todas as réplicas. A escala vertical (aumento da capacidade do servidor) é mais comum nesse caso.

NoSQL: os bancos de dados NoSQL são projetados para serem facilmente escaláveis horizontalmente. Eles podem distribuir os dados em várias máquinas, permitindo maior flexibilidade para lidar com grandes volumes de dados e tráfego.

Modelo de Consistência

SQL: os bancos de dados SQL seguem o modelo ACID (Atomicidade, Consistência, Isolamento, Durabilidade), o que significa que as transações são processadas com rigorosas garantias de consistência, garantindo que os dados permaneçam íntegros mesmo em caso de falhas.

NoSQL: os bancos de dados NoSQL geralmente seguem o modelo BASE (Basicamente Disponível, Suave), que prioriza a disponibilidade e a tolerância a partições em detrimento da consistência estrita. Isso significa que, em sistemas distribuídos, a consistência eventual pode ser aceita e os dados podem levar algum tempo para serem propagados completamente.

Vantagens e Desvantagens de Bancos de Dados SQL e NoSQL

Bancos de dados SQL e NoSQL têm suas próprias vantagens e desvantagens, tornando-os adequados para diferentes tipos de aplicativos e cenários. Vamos explorar as principais características de cada abordagem:

Vantagens de Bancos de Dados SQL

Consistência Estrita: os bancos de dados SQL seguem o modelo ACID, garantindo consistência estrita dos dados em todas as operações. Isso é essencial em aplicações que exigem alta integridade dos dados, como sistemas financeiros e de gerenciamento de estoque.

Relações e Joins: bancos de dados SQL são ideais para aplicativos que precisam realizar operações complexas de junção e consulta em várias tabelas. Essa capacidade permite a realização de análises avançadas e relatórios detalhados.

Maturidade e Suporte: SQL é uma linguagem de consulta padronizada e amplamente adotada, com muitas ferramentas e recursos disponíveis. Bancos de dados SQL são tecnologias maduras com ampla documentação e suporte da comunidade.

Desvantagens de Bancos de Dados SQL

Escalabilidade Vertical Limitada: bancos de dados SQL são mais desafiadores para escalar horizontalmente, o que pode se tornar caro e difícil de gerenciar em cenários com grande crescimento de dados e tráfego.

Rigidez do Esquema: a necessidade de definir um esquema fixo pode ser restritiva em ambientes em que a estrutura dos dados está sujeita a mudanças frequentes.

Vantagens de Bancos de Dados NoSQL

Escalabilidade Horizontal: bancos de dados NoSQL são projetados para serem altamente escaláveis horizontalmente, permitindo que grandes volumes de dados e tráfego sejam distribuídos em múltiplos nós do cluster.

Flexibilidade no Esquema: NoSQL oferece a capacidade de armazenar dados com estruturas diferentes em uma mesma coleção, tornando-os adequados para aplicativos com requisitos de dados variáveis e em evolução.

Alta Performance: o modelo BASE adotado por bancos de dados NoSQL pode resultar em melhor desempenho e menor latência em comparação com bancos de dados SQL em certos cenários de alta concorrência.

Desvantagens de Bancos de Dados NoSQL

Consistência Eventual: alguns bancos de dados NoSQL adotam a consistência eventual, o que pode levar a janelas de tempo em que os dados estão temporariamente inconsistentes entre os nós.

Menor Suporte e Documentação: embora NoSQL tenha ganhado popularidade, a oferta de ferramentas e recursos pode não ser tão extensa quanto a dos bancos de dados SQL devido à sua natureza mais recente e diversificada.

Principais Banco de Dados SQL

Existem diversos sistemas de gerenciamento de banco de dados (SGBDs) que utilizam a linguagem SQL para gerenciar e manipular dados. Abaixo estão alguns dos principais bancos de dados SQL:

MySQL: um dos bancos de dados SQL mais populares, conhecido por sua velocidade, escalabilidade e ampla adoção em diversas aplicações, desde pequenos sites até grandes empresas.

PostgreSQL: um SGBD de código aberto conhecido por sua robustez, recursos avançados, extensibilidade e conformidade com padrões SQL.

Oracle Database: um sistema de gerenciamento de banco de dados líder no setor, amplamente utilizado em empresas para aplicações de missão crítica, com foco em escalabilidade e desempenho.

Microsoft SQL Server: um SGBD da Microsoft que oferece integração com outras tecnologias da empresa, adequado para aplicativos Windows e ambientes empresariais.

SQLite: um SGBD embutido de código aberto, que não requer um servidor separado, sendo adequado para aplicativos móveis e pequenas aplicações.

IBM Db2: um SGBD conhecido por sua confiabilidade e recursos avançados, usado principalmente em empresas para aplicações corporativas.

MariaDB: um *fork* do MySQL, também de código aberto, que mantém compatibilidade com o MySQL enquanto introduz novos recursos e melhorias.

SAP HANA: um SGBD *in-memory* da SAP, que oferece alto desempenho para análise de dados em tempo real e aplicações empresariais.

Microsoft Access: um SGBD de *desktop* da Microsoft, mais adequado para pequenas aplicações ou ambientes individuais.

Teradata: um SGBD usado para gerenciar grandes volumes de dados e processamento de data *warehousing*, com foco em análise de dados.

Principais Banco de Dados NoSQL

Os bancos de dados NoSQL são categorizados de acordo com seus modelos de dados. Cada categoria oferece recursos específicos para atender a diferentes requisitos de aplicativos.

Abaixo veremos os principais bancos de dados NoSQL divididos por categoria.

Bancos de Dados Document-oriented (Orientados a Documentos)

MongoDB: um dos bancos de dados NoSQL mais populares, que armazena dados em documentos JSON-like, permitindo esquemas flexíveis e consultas poderosas.

Couchbase: um banco de dados NoSQL orientado a documentos e com capacidade de armazenamento em memória, que oferece alta velocidade e escalabilidade.

Bancos de Dados Key-Value Stores (Armazenamento de Chave-Valor)

Redis: um banco de dados NoSQL em memória que armazena dados como pares chave-valor e suporta estruturas de dados complexas, tornando-o ideal para armazenamento em cache e gerenciamento de sessões.

Amazon DynamoDB: um serviço de banco de dados NoSQL da Amazon Web Services (AWS), que oferece armazenamento de chave-valor e desempenho rápido, sendo amplamente utilizado em aplicações *web* e móveis.

Bancos de Dados Column-family Stores (Armazenamento de Famílias de Colunas):

Apache Cassandra: um banco de dados NoSQL distribuído, altamente escalável e tolerante a falhas, projetado para gerenciar grandes volumes de dados em vários servidores.

HBase: um banco de dados NoSQL baseado no Apache Hadoop, projetado para armazenar e gerenciar grandes quantidades de dados em formato de família de colunas.

Bancos de Dados Graph Databases (Bancos de Dados de Grafos):

Neo4j: um banco de dados NoSQL de grafos popular, que permite armazenar e consultar dados altamente conectados, sendo adequado para aplicações de redes sociais e sistemas de recomendação.

Amazon Neptune: um serviço de banco de dados NoSQL de grafos oferecido pela AWS, que permite a criação e consulta de grafos altamente conectados.

> Capítulo 6

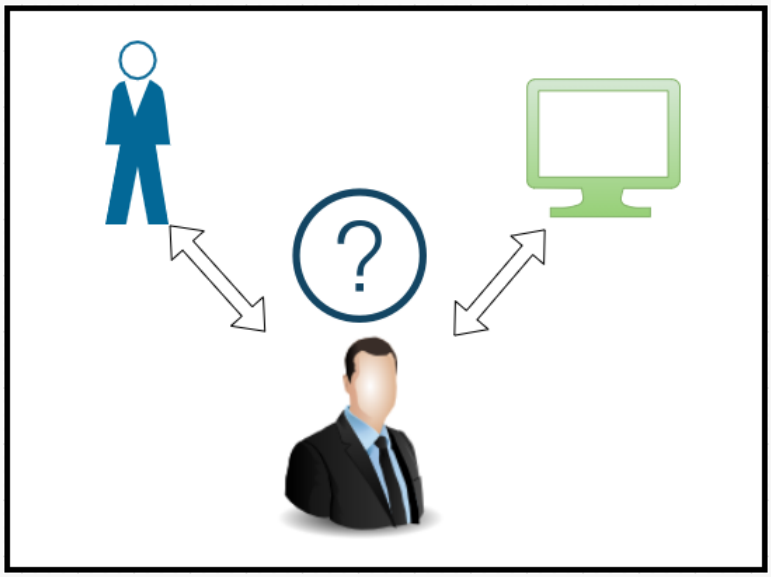
Capítulo 6. Fundamentos de Aprendizagem de Máquina

Para extrair *insights* dos dados é muito comum o cientista de dados utilizar várias técnicas e algoritmos. Neste capítulo vamos realizar uma introdução sobre esse assunto.

Inteligência Artificial

A Inteligência Artificial (IA) é uma parte da ciência da computação onde os sistemas são planejados para a execução de tarefas que necessitam de inteligência humana e possui várias técnicas para realização da atividade. Para entendimento do comportamento de uma IA vamos tomar como exemplo o teste de *Turing*. O teste de *Turing* testa a capacidade de uma máquina apresentar um comportamento inteligente equiparado a um ser humano. Podemos observar na Figura 11 um exemplo prático ilustrativo de um teste de *Turing*.

Figura 11 – Teste de Turing.



1. Um avaliador humano faz uma série de perguntas baseadas em texto para uma máquina e um humano, sem ver nenhum deles.

2. O humano e a máquina respondem à pergunta do avaliador.

3. Se o avaliador não conseguir diferenciar se as respostas dadas foram de um humano ou de máquina, o computador é aprovado no teste de *Turing*. Isso significa que exibiu um comportamento semelhante ao humano ou inteligência artificial.

O *Big Data* é umas das principais fontes para a inteligência artificial, gerando dados para que cada dia mais ela crie sistemas que possuem a capacidade de auxiliar na vida cotidiana. Ao contrário do que se imaginava, a inteligência artificial não resultou em vários robôs realizando atividades humanas, mas tem operado de forma eficiente e um tanto silenciosa.

Com a grande coleta de dados do *Big Data*, é possível a criação de modelos que analisam e antecipam comportamentos e dinâmicas de sistemas complexos. Esses dados provêm não apenas da interação dos indivíduos na rede, mas também pelo rastro que ele deixa na *internet* sem saber.

O volume desses dados produz uma característica muito importante no processo da IA: descobrir quais dados são relevantes para a análise e utilizar ferramentas que são capazes de manipular e estudar essa quantidade exorbitante de dados. Entender esses elementos, suas origens e projetar as condições futuras permite um melhor planejamento estratégico.

Esse é o ponto onde a Inteligência Artificial entra. A quantidade dos dados somada à necessidade da análise de cada um pode ser um processo automatizado através de *Machine Learning* constante, o que significa ter uma máquina capaz de aprender certa informação. O aprendizado de

máquina utiliza códigos para fazer uma varredura em grandes quantidades de dados, buscando padrões. Elevando a quantidade de vezes em que a máquina reproduz esse comportamento, ela será capaz de analisar grandes quantidades muito mais rapidamente se comparado ao processo humano manual.

Aprender é uma atividade inerente ao ser humano, e quando há a tentativa e erro, produz-se outros resultados que podem auxiliar futuramente. O *Machine Learning* segue esse mesmo princípio, o que permite que os resultados se tornem sempre mais assertivos e específicos.

Se tratando de máquinas, o *Big Data* vai fornecer exatamente aquilo que é necessário para o bom desenvolvimento de seu aprendizado: dados não estruturados e contínuos. Isso replica a forma intuitiva do ser humano de produzir novos conhecimentos.

Machine Learning

O aprendizado de máquina (ML) é um subconjunto de inteligência artificial que funciona muito bem com dados estruturados. O objetivo por trás do aprendizado de máquina é que as máquinas aprendam padrões em seus dados sem que você os programe explicitamente para isso. Atualmente, o aprendizado de máquina não pode fornecer o tipo de IA que os filmes apresentam. Mesmo os melhores algoritmos não conseguem pensar, sentir, apresentar qualquer forma de autoconsciência ou exercer o livre arbítrio. O que o aprendizado de máquina pode fazer é realizar análises preditivas com muito mais rapidez do que qualquer ser humano. Como resultado, o aprendizado de máquina pode ajudar os humanos a trabalharem com mais eficiência. O estado atual da IA, então, é o de fazer análise, mas os humanos ainda devem considerar as implicações dessa análise – tomando as decisões morais e éticas necessárias (MUELLER; MASSARON, 2016).

Existem três tipos de aprendizado de máquina. O aprendizado supervisionado, o não supervisionado e o por reforço. O aprendizado supervisionado é baseado na regressão básica e classificação. O humano fornece um banco de dados e ensina a máquina a reconhecer padrões e semelhanças através de rótulos. Por exemplo, a máquina pode reconhecer um carro, mas a cor, tamanho e outras características podem variar. No entanto, a máquina aprende elementos-chave que identificam um carro.

Já no aprendizado não supervisionado, a máquina aprende com dados de teste que não foram rotulados, classificados ou categorizados previamente. Desta forma, não existe supervisão humana. O aprendizado não supervisionado identifica semelhanças nos dados e reage com base na ausência ou presença das semelhanças em cada novo dado. A clusterização, que é uma técnica de aprendizado não supervisionado que permite dividir automaticamente o conjunto de dados em grupos de acordo com uma similaridade.

Aprendizado por reforço é o aprendizado baseado na experiência que a máquina tem e aprende a lidar com o que errou antes e procurar a abordagem correta. Podemos comparar o aprendizado por reforço de uma criança. Por exemplo, quando uma criança começa a engatinhar, ela tenta se levantar e cai várias vezes, e após muitas tentativas ela consegue uma forma de se levantar sem cair. Um outro exemplo são as recomendações de sites de entretenimento como o *YouTube*. Após assistir um vídeo, a plataforma irá mostrar títulos semelhantes que acredita que você também irá gostar. No entanto, se você começa a assistir o recomendado e não o termina, a máquina entende que a recomendação não foi boa e irá tentar outra abordagem da próxima vez.

Para exemplificar ainda mais, vamos ilustrar um exemplo de aprendizado supervisionado para detecção de fraudes. De maneira geral e de alto nível, temos:

1. O analista cria regras para o que constitui fraude (por exemplo, uma conta com mais de 30 transações no mês, compras em diversos setores, saldo médio menor que R$200,00).

2. Essas regras são passadas para o algoritmo que recebe os dados que são rotulados como “fraude” ou “não fraude”. Após isso, a máquina aprende o comportamento dos dados fraudulentos.

3. Com o apoio das regras, a máquina começa a prever as fraudes.

4. Ao final é feita a validação do modelo previsto da máquina. Para isso, um analista investiga e verifica manualmente se as previsões do modelo preveem a fraude.

Algoritmos de Machine Learning

Como vimos anteriormente, temos três tipos de algoritmos de *Machine Learning*. O algoritmo de aprendizado supervisionado, o aprendizado não supervisionado e o aprendizado por reforço. Abaixo vamos apresentar as principais características de cada um desses algoritmos.

Para os algoritmos de aprendizado supervisionados temos basicamente duas classes: classificação e regressão. Os algoritmos de classificação têm como objetivo identificar a qual categoria pertence uma amostra do problema. Por exemplo, podemos classificar se uma transação é uma fraude ou não; se um e-mail é SPAM ou não, se uma mensagem em rede social possui sentimento positivo, negativo ou neutro entre outros. Os principais algoritmos são árvores de decisão, *Naive Bayes* e redes neurais.

Já para os algoritmos de regressão, a ideia é prever um valor de uma variável com base no valor de outra. Para isso, o modelo pode aprender uma função que prevê o preço das ações de um fundo imobiliário, uma demanda de venda, o tempo de desgaste de pneus em uma frota de carros, tempo de baixa de estoque de peças ou qualquer outro valor quantitativo. Os

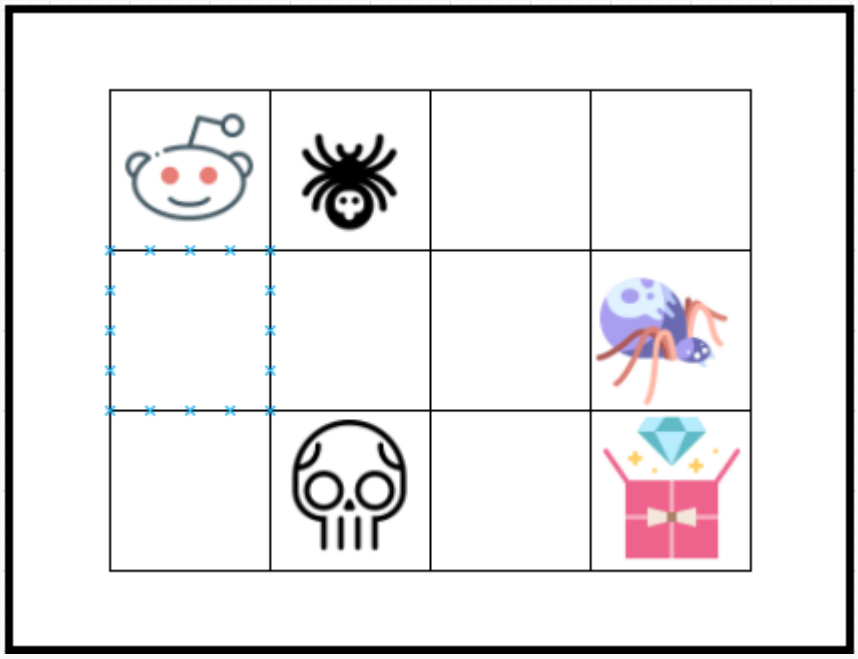
algoritmos mais utilizados são regressão linear, regressão logística e as redes neurais que podem apresentar resultados com valores contínuos.

Os algoritmos de aprendizado não supervisionado podem ser divididos em duas classes: Associação e Clusterização. Os algoritmos de associação permitem o descobrimento de regras e correlação em uma base de dados, identificando conjuntos de itens que ocorrem juntos dentro de uma determinada frequência. É muito utilizado no setor de varejo para analisar carrinhos de compras, assim descobrindo os itens frequentemente mais comprados em conjunto. Desta forma, criando novas estratégias de *marketing* e vendas. O algoritmo mais utilizado nessa classe é o algoritmo de regras de associação.

Os algoritmos de clusterização ou agrupamento permitem que seja feito agrupamento de grupos com base nas semelhanças encontradas. É uma técnica que permite realizar a divisão de grupos em um conjunto de dados de forma automática, baseado em medidas de similaridade ou de distância. Existem vários métodos que permitem obter medidas de similaridade, podemos citar a similaridade de cosseno e a correlação de Pearson, e citar os algoritmos baseado em particionamento, baseado em densidade, o hierárquico aglomerativo e hierárquico divisório.

Os algoritmos de aprendizado por reforço permitem que o modelo aprenda executando ações e avaliando o resultado dessas ações. De maneira geral, o algoritmo funciona da seguinte forma: o algoritmo geralmente conhece as regras, no entanto, desconhece a melhor sequência de ações que devem ser executadas. Desta forma, os algoritmos aprendem de forma interativa. A Figura 12 ilustra um exemplo de problema no qual queremos encontrar o melhor caminho para alcançar o diamante (recompensa).

Figura 12 – Exemplo algoritmo de aprendizado por reforço.



O agente deve encontrar o melhor caminho possível para alcançar a recompensa e quando encontrar um obstáculo, deve ser penalizado (pois ele deve escolher o caminho sem obstáculos). Com a Aprendizagem por reforço, podemos treinar o agente para encontrar o melhor caminho. Os algoritmos de aprendizado por reforço são muito aplicados no campo de estudo da robótica. Podemos citar os seguintes algoritmos: *Multi-Armed Bandits, Contextual Bandits e k-Armed Bandits*.

Algoritmos Supervisionado Árvore de Decisão

Os algoritmos de árvore de decisão são ferramentas essenciais no arsenal dos cientistas de dados, permitindo a tomada de decisões informadas com base em sequências de escolhas lógicas (HAN, HAIHONG & LE, 2011).

O algoritmo de árvore de decisão é uma técnica de aprendizado de máquina supervisionado que é amplamente utilizada para resolver problemas de classificação e regressão. Ele cria uma representação em

forma de árvore das decisões e possíveis resultados com base nos atributos dos dados de treinamento.

É importante considerarmos os seguintes pontos sobre os algoritmos de árvore de decisão:

Estrutura Hierárquica: a árvore de decisão é composta por um conjunto de nós interligados que representam as decisões tomadas ao longo do processo de classificação ou previsão. A raiz da árvore representa a decisão inicial, enquanto os nós internos representam testes nos atributos e os nós folha representam as classes ou valores de saída.

Divisão de Dados: o processo de construção da árvore começa com o conjunto de treinamento completo. Em cada etapa, o algoritmo seleciona o atributo que melhor divide os dados, com base em critérios como ganho de informação ou índice de Gini.

Criação de Nós: a cada divisão, novos nós são criados na árvore, representando diferentes caminhos de decisão com base nos valores do atributo selecionado. O algoritmo continua a dividir os dados em nós sucessivos até que critérios de parada sejam atendidos, como um número mínimo de amostras em um nó ou a profundidade máxima da árvore.

Classificação e Previsão: depois de construída, a árvore pode ser usada para classificar novos dados ou fazer previsões. Começando na raiz, os dados seguem os ramos da árvore até chegarem a um nó folha, que representa a classe prevista ou o valor de saída.

Overfitting: um desafio na construção de árvores de decisão é evitar o overfitting, no qual a árvore se adapta demais aos dados de treinamento e não generaliza bem para novos dados. Isso pode ser controlado por meio de técnicas como a poda (*pruning*) da árvore ou a limitação da profundidade máxima.

Aplicações: as árvores de decisão são usadas em uma variedade de áreas, como diagnóstico médico, análise financeira, previsão de mercado, reconhecimento de padrões, entre outras.

Algoritmos Supervisionado Random Forest

A Random Forest, ou Floresta Aleatória, é uma técnica avançada de aprendizado de máquina que se destaca por sua eficácia em tarefas de classificação e regressão (HAN, HAIHONG & LE, 2011). A Random Forest se baseia na combinação de múltiplas árvores de decisão individuais para criar um modelo robusto e preciso. Ele é usado para tarefas de classificação, regressão e detecção de anomalias.

Aqui estão os principais pontos a serem considerados sobre o algoritmo Random Forest:

Combinação de Árvores: a ideia central da Random Forest é criar várias árvores de decisão independentes durante o treinamento e, em seguida, combinar suas previsões para obter um resultado final. Cada árvore é treinada em um subconjunto aleatório dos dados de treinamento, usando uma técnica chamada de bootstrapping (amostragem com reposição).

Randomização: além da amostragem de dados, a Random Forest também aplica randomização durante a construção de cada árvore. Em cada divisão, em vez de selecionar o melhor atributo, ela escolhe aleatoriamente um subconjunto de atributos para considerar. Isso ajuda a reduzir a correlação entre as árvores individuais e a promover a diversidade.

Votação ou Média: para a tarefa de classificação as árvores individuais “votam” na classe prevista, e a classe com mais votos é escolhida como a previsão final. No caso de regressão, as previsões das árvores são médias para determinar o resultado final.

Redução de Overfitting: uma das principais vantagens da Random Forest é sua capacidade de reduzir o overfitting. A combinação de várias árvores e a randomização durante a construção tornam o modelo mais generalizável e menos suscetível a se ajustar demais aos dados de treinamento.

Importância de Features: a Random Forest pode calcular a importância relativa de cada atributo (feature) ao observar como a precisão do modelo muda quando cada atributo é aleatoriamente embaralhado. Isso ajuda a entender quais atributos são mais informativos para a tarefa.

Aplicações: a Random Forest é amplamente usada em diversas áreas, como classificação de imagens médicas, detecção de fraudes, análise de sentimentos, previsão de demanda e muito mais.

Aprendizado não supervisionado K-means

K-means é um dos algoritmos mais utilizados para realizar agrupamentos de dados numéricos em mineração de dados (DESAI et al., 2016), (KANUNGO et al., 2002). O objetivo do *K-means* é encontrar a melhor divisão de p dados em k grupos, de forma que a distância total entre os dados de um grupo e seu centroide, somados por todos os grupos, seja minimizada. Ele é amplamente utilizado em análise de cluster e segmentação de dados, permitindo a identificação de padrões subjacentes nos dados sem a necessidade de rótulos prévios.

Aqui estão os principais pontos a serem considerados sobre o algoritmo K-Means:

Objetivo do Agrupamento: o objetivo do K-Means é agrupar um conjunto de pontos de dados em K clusters, onde K é um número predefinido de clusters a serem criados. O algoritmo tenta encontrar K centroides (pontos que representam o centro de um cluster) que minimizam a distância entre os pontos de dados e o centroide correspondente.

Passos do Algoritmo: o K-Means opera em etapas iterativas:

- Inicialização: inicializa K centroides aleatoriamente no espaço dos dados.
- Atribuição: cada ponto de dados é atribuído ao centroide mais próximo.
- Atualização: os centroides são recalculados como a média dos pontos de dados atribuídos ao cluster.
- Reatribuição: os pontos de dados são novamente atribuídos aos centroides mais próximos.
- Convergência: os passos de atribuição e atualização são repetidos até que os centroides e as atribuições se estabilizem.

Critério de Distância: a métrica de distância mais comum usada pelo K-Means é a distância euclidiana, mas outras métricas também podem ser usadas, dependendo do domínio do problema.

Número de Clusters (K): um dos desafios do K-Means é determinar o número ideal de clusters (K). Isso pode ser feito por meio de métodos como o método do cotovelo (Elbow Method), que avalia a variação explicada à medida que K aumenta.

Sensível à Inicialização: a convergência do K-Means pode depender da inicialização dos centroides. Inicializações ruins podem levar a resultados subótimos, portanto é comum executar o algoritmo várias vezes com diferentes inicializações e selecionar o melhor resultado.

Limitações: o K-Means pode ter dificuldade em lidar com clusters de diferentes formas e tamanhos. Ele também assume que os clusters têm variações semelhantes, o que nem sempre é verdade em todos os conjuntos de dados.

Aplicações: o K-Means é aplicado em várias áreas, como análise de mercado, segmentação de clientes, agrupamento de documentos, análise de imagens e detecção de anomalias.

Naive Bayes

O algoritmo Naive Bayes é um método de aprendizado de máquina baseado na teoria probabilística de Bayes, utilizado principalmente para classificação de textos e análise de sentimentos, mas também aplicado em uma variedade de outras tarefas de classificação. É conhecido por sua simplicidade, eficiência e boa performance em muitos cenários.

Aqui estão os principais pontos a serem considerados sobre o algoritmo Naive Bayes:

Teorema de Bayes: o algoritmo Naive Bayes é fundamentado no Teorema de Bayes, que descreve como as probabilidades condicionais podem ser invertidas. A ideia básica é calcular a probabilidade de uma hipótese (ou classe) dado um conjunto de evidências (atributos ou *features*).

"Naive" (Ingênuo): o termo "naive" é usado porque o algoritmo faz uma suposição simplificada de que os atributos são independentes, ou seja, que a presença ou ausência de um atributo não está relacionada à presença ou ausência de outros atributos. Essa suposição, embora raramente seja verdadeira em situações do mundo real, muitas vezes funciona bem e simplifica os cálculos.

Classificação Bayesiana: na classificação de um novo exemplo, o Naive Bayes calcula a probabilidade de cada classe possível para esse exemplo, usando as probabilidades condicionais dos atributos para cada classe. A classe com a maior probabilidade é escolhida como a previsão final.

Distribuição de Probabilidades: o Naive Bayes assume diferentes distribuições de probabilidade para os atributos, dependendo da natureza

dos dados. As distribuições mais comuns incluem Bernoulli (para dados binários), Multinomial (para dados discretos) e Gaussiana (para dados contínuos).

Laplace Smoothing: para evitar a probabilidade zero em cenários onde um atributo não foi observado em uma classe durante o treinamento, o Laplace Smoothing (também conhecido como adição de Laplace) é usado para suavizar as probabilidades.

Aplicações: o Naive Bayes é frequentemente usado em tarefas de classificação de texto, como detecção de spam, análise de sentimentos e categorização de documentos. Também é aplicado em áreas como diagnóstico médico, detecção de fraudes e reconhecimento de padrões.

Simplicidade e Eficiência: o Naive Bayes é rápido e eficiente, sendo capaz de lidar com grandes volumes de dados de maneira escalável.

Limitações: a suposição de independência ingênua nem sempre é válida, o que pode levar a resultados subótimos quando as correlações entre atributos são importantes. Além disso, o Naive Bayes pode ser sensível a atributos irrelevantes ou com alta cardinalidade.

Correlação Linear

Em pesquisas, frequentemente procura-se verificar se existe relação entre duas ou mais variáveis, isto é, saber se as alterações sofridas por uma das variáveis são acompanhadas por alterações nas outras. Por exemplo, qual a relação entre a p tempo de trabalho da pessoa e seu salário ou o consumo de bebidas com a renda mensal. A correlação permite verificar se duas variáveis independentes estão associadas uma com a outra. O termo correlação significa relação em dois sentidos (co + relação), e é usado em estatística para designar a força que mantém unidos dois conjuntos de valores. A verificação da existência e do grau de relação entre as variáveis é o objeto de estudo da correlação.

Existem graus de força entre a relação entre duas variáveis. Para calcular essa força podemos utilizar o coeficiente de correlação de Pearson. Os graus de correlação vão de -1 a 1. A correlação pode ser positiva ou negativa. Se o valor for mais próximo de -1, dizemos que existe uma correlação forte negativa (quando uma variável aumenta a outra diminui), por exemplo, o preço do dólar aumenta, diminui as compras de produtos internacionais. Se o valor for mais próximo de 1, dizemos que existe uma correlação forte positiva (quando uma variável aumenta a outra também aumenta), por exemplo, promoção no preço da carne, aumenta o volume de compra.

Técnicas para balanceamento de dados

O balanceamento de dados é uma etapa importante no processo de treinamento de modelos de aprendizado de máquina, especialmente quando as classes de saída estão desequilibradas. Desequilíbrios podem levar a problemas como viés no modelo, baixo desempenho na classe minoritária e resultados não realistas.

Existem várias técnicas para lidar com esse desequilíbrio e melhorar o desempenho do modelo. Algumas delas incluem:

Oversampling (Supremoamostragem): essa técnica envolve a replicação de amostras da classe minoritária para criar um equilíbrio entre as classes. Pode ser feita aleatoriamente ou usando técnicas mais avançadas, como SMOTE (*Synthetic Minority Over-sampling Technique*), que gera novos exemplos sintéticos com base em vizinhos próximos.

Undersampling (Subamostragem): aqui, reduz-se o número de amostras da classe majoritária, buscando um equilíbrio entre as classes. Isso pode ser feito aleatoriamente ou usando algoritmos para selecionar amostras que mantenham as características da classe majoritária.

Geração de Dados Sintéticos: além do SMOTE, outras técnicas como ADASYN (*Adaptive Synthetic Sampling*) podem ser usadas para gerar dados sintéticos, mantendo a estrutura dos dados originais e criando novas amostras para a classe minoritária.

A escolha da técnica de balanceamento depende do contexto do problema, do tamanho do conjunto de dados e das características das classes. No entanto, é importante considerar as implicações dessas técnicas, pois *oversampling* excessivo pode levar ao *overfitting* e undersampling pode resultar em perda de informações. O objetivo é encontrar um equilíbrio que melhore o desempenho do modelo sem introduzir viés ou distorções nos dados.

Métricas de avaliação Aprendizado Supervisionado

A avaliação de modelos de aprendizado de máquina supervisionado é essencial para determinar o quão bem um modelo está performando em tarefas como classificação e regressão. Existem diversas métricas que fornecem insights sobre diferentes aspectos do desempenho do modelo. Aqui estão algumas das principais métricas de avaliação:

Acurácia (*Accuracy*): a acurácia é uma métrica simples e amplamente usada que mede a proporção de previsões corretas em relação ao total de previsões. É apropriada quando as classes estão balanceadas, mas pode ser enganosa quando as classes têm desequilíbrio.

Precisão (*Precision*): a precisão mede a proporção de verdadeiros positivos (amostras corretamente classificadas como positivas) em relação ao total de amostras classificadas como positivas. É relevante quando o foco está em minimizar os falsos positivos.

Revocação (*Recall*): a revocação, também conhecida como taxa de verdadeiros positivos ou sensibilidade, mede a proporção de verdadeiros

positivos em relação ao total de amostras positivas. É importante quando o objetivo é minimizar os falsos negativos.

F1-Score: o F1-Score é a média harmônica da precisão e revocação, o que a torna uma métrica balanceada que leva em consideração tanto os falsos positivos quanto os falsos negativos.

Matriz de Confusão: a matriz de confusão é uma representação tabular das previsões do modelo em relação às classes reais. Ela fornece informações sobre verdadeiros positivos, falsos positivos, verdadeiros negativos e falsos negativos, permitindo a análise detalhada do desempenho.

A escolha da métrica de avaliação depende do tipo de problema, do desequilíbrio das classes e dos objetivos específicos. É importante compreender as métricas relevantes para a tarefa em questão e interpretar seus resultados de forma adequada para tomar decisões informadas sobre a performance do modelo

Referências

ABITEBOUL, S. Querying Semi-Structured Data. Inria-Rocquencourt, 1970.

AWS – AMAZON WEB SERVICES. O que é ciência de dados?. Disponível em: <https://aws.amazon.com/pt/what-is/data-science/>. Acesso em: 17 de ago, 2023.

CAIRO, A. The Truthful Art: Data, Charts, and Maps for Communication. New Riders Publishing, 2016.

CIELEN, D.; MEYSMAN, A.; ALI, M. Introducing Data Science: Big Data, Machine Learning, and More, Using Python Tools. Manning Publications, 2016.

DAVENPORT, T. H.; KIM, J. Keeping Up with the Quants: Your Guide to Understanding and Using Analytics. Harvard Business Review Press, 2013.

DEGROOT, M. H.; SCHERVISH, M. J. Probability and Statistics. 4. ed. Pearson, 2011.

ELMASRI, R.; SHAMKANT, N. B. Sistemas de Bancos de Dados. Pearson Universidades, 2019. p. 1152.

EMC EDUCATION SERVICES. Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data. 1. ed. Wiley, 2015. p. 432.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. Knowledge discovery and datamining: Towards a unifying framework. In: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, KDD'96, p. 82–88, AAAI Press, 1996.

GLASSDOOR. Salários de Cientista De Dados em Brasil. Disponível em: https://www.glassdoor.com.br/Sal%C3%A1rios/cientista-de-dados-sal%C3%A1rio-SRC. Acesso em: 14 de ago, 2023.

GOODFELLOW, I., BENGIO, Y.; COURVILLE, A. Deep Learning. MIT Press, 2016.

HAN, J.; HAIHONG, E.; LE, G. Data Mining: Concepts and Techniques. Morgan Kaufmann, 2011.

KANUNGO, T. et al. An Efficient K-Means Clustering Algorithm Analysis and Implementation. In: IEEE Transactions on Pattern Analysis and Machine Intelligence. 24. v. 7. n. p. 881-892. 2002.

KELLEHER, J.; MAC NAMEE, B.; D'ARCY, A. Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies. MIT Press, 2015.

MCKINNEY, W. Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython. O'Reilly Media, 2018.

METHANIAS, C. J. Projetando sistemas de apoio à decisão baseados em Data Warehouse. Axcel Books, 2004.

MUELLER, J. P.; MASSARON, L. Machine Learning: For Dummies. 1. ed. John Wiley & Sons, 2016. p. 435 p.

VANDERPLAS, J. Python Data Science Handbook. O'Reilly Media, 2016.