SÉRIE ACADÊMICA

BUSINESS ANALYTICS, BIG DATA E INTELIGÊNCIA ARTIFICIAL

FABIANO CASTELLO DE CAMPOS PEREIRA

EDITORA pecege

SÉRIE ACADÊMICA

BUSINESS ANALYTICS, BIG DATA E INTELIGÊNCIA ARTIFICIAL

FABIANO CASTELLO DE CAMPOS PEREIRA

PIRACICABA • SÃO PAULO

EXPEDIENTE EQUIPE

ORGANIZADORES
Daniela Flôres
Francisco Javier Sebastian Mendizabal Alvarez
Marcos Roberto Luppe
Maria Cecília Perantoni Fuchs Ferraz
Ricardo Harbs
Tatiana Rosa Diniz

PROJETO GRÁFICO
Ana Paula Mendes Vidal de Negreiros

REVISÃO
Fernanda Latanze Mendes Rodrigues

DADOS INTERNACIONAIS DE CATALOGAÇÃO NA PUBLICAÇÃO (CIP)
(CÂMARA BRASILEIRA DO LIVRO, SP, BRASIL)

P436b

Pereira, Fabiano Castello de Campos.

Business analytics, big data e inteligência artificial / Fabiano Castello de Campos Pereira. - - Piracicaba, SP : PECEGE Editora, 2022.

Série Acadêmica
ISBN: 978-85-92582-50-0

1. Dados digitais. 2. Tecnologias e infraestruturas. 3. Citizen data scients. 4. Machine learning. 5. Eficiência operacional. I. Autor. II. Título. III. Série.

CDD: 004.65

FICHA CATALOGRÁFICA ELABORADA POR **FELIPE MUSSARELLI CRB 9935/8**

PREZADO(A) ALUNO(A),

Esse material foi desenvolvido no intuito de auxiliá-lo com os estudos nos cursos de **MBA da USP/ESALQ**, servindo como um referencial teórico básico e complementar às aulas oferecidas nos cursos.

Desejamos que esse material, de alguma forma, contribua para acrescentar novos conhecimentos, impulsionar o aprendizado e aprimorar as competências que já possui.

Bons estudos!!!

EQUIPE PECEGE

SOBRE O AUTOR

FABIANO CASTELLO DE CAMPOS PEREIRA

Consultor da cDataLab, empresa com foco em big data, analytics e inteligência artificial. Iniciou sua carreira corporativa em 1994 na Arthur Andersen e atuou como diretor executivo na Deloitte, Electrolux, Ambev e Oi, com forte atuação tanto no Brasil como no exterior. Possui formação em computação pelo Mackenzie, administração pela EAESP-FGV e ciência de dados pela Johns Hopkins University. É mestre pelo programa de pós-graduação stricto senso na FEA/USP - Faculdade de Economia e Administração da Universidade de São Paulo. Obteve as principais certificações internacionais de auditoria - CISA, CISM, CIA, CCSA e CRMA, além de ser conselheiro certificado pelo IBGC no Brasil. É professor da Inova Business School, escola de negócios com foco em futuro, tendências e inovação; dos programas de MBA da ESALQ/USP e da PUC-RS; da FIA/USP, em parceria com o Coursera; e do IBGC no curso de formação de membros de comitês de auditoria.

SUMÁRIO

1. Introdução

Vivemos um momento de digitalização. Os negócios baseados em tecnologia são cada vez mais frequentes e mesmo as empresas tradicionais repensam suas estratégicas, de forma a beneficiarem-se de dados digitais.

De forma geral, o mundo dos negócios tem utilizado os conceitos de fatos e dados como suporte para a tomada de decisões, principalmente pelo fato de os dados estarem cada vez mais disponíveis e, sua armazenagem, cada vez mais barata. É importante notar que fatos e dados não substituem a vivência profissional e o instinto que é fruto de anos de experiência e conhecimento em determinada área. Devem ser vistos, portanto, de forma complementar.

O objetivo desta série acadêmica é proporcionar ao aluno uma visão geral sobre aspectos relacionados a dados digitais (principais definições, diferenças entre termos muitas vezes usados de forma intercambiável, tecnologias e infraestruturas necessárias para sua utilização) de uma forma concisa e, sobretudo, evitando os termos técnicos. As explicações a seguir foram elaboradas visando os alunos que não possuem formação técnica em áreas de tecnologia e afins, para que se beneficiem da aplicação dessas tecnologias, ainda que não tenham conhecimento profundo que permita liderar sua implementação.

2. Analytics e Business Analytics

Analytics é um termo amplamente utilizado, porém poucas vezes definido. Por ter sido criado pelo mercado, sem uma referência seminal, diversas interpretações surgiram ao longo do tempo. É consenso, no entanto, que envolve suporte a decisão de negócios e utiliza dados no formato digital.

Voltando no tempo é possível entender as raízes do que hoje é chamado de Analytics. Nos anos de 1970, criou-se os primeiros sistemas denominados “Decision Support Systems” (DSS), ou Sistemas de Apoio à Decisão. O DSS passou a ser usado como um termo comum para designar aplicações com esse fim no mundo dos negócios, bem como uma linha de pesquisa acadêmica. Ao longo do tempo, outros sistemas mais específicos surgiram, como os painéis de informações executivas, hoje mais conhecidos como “dashboards”. Na década de 1990, o termo “Business Intelligence” (BI) tornou-se amplamente utilizado para definir um vasto conjunto de aplicações, tecnologias e processos de coleta, armazenamento, acesso e análise de dados, cujo objetivo era melhorar a qualidade da tomada de decisão. Assim como o BI evoluiu a partir do DSS, o Analytics pode ser compreendido como uma evolução moderna a partir do BI. Atualmente, quando o termo Analytics é utilizado, subentende-se que o assunto está relacionado a aplicações de análise de dados digitais.

O interesse cada vez maior na utilização de Analytics no mundo dos negócios está centrado no fato de que decisões baseadas em fatos e dados resultam em melhores decisões e, assim, em melhor desempenho

organizacional. Adicionalmente, o uso consistente de dados digitais de qualidade é considerado uma vantagem competitiva.

O termo Business Analytics, definido por Davenport e Harris (2007) como "o uso extensivo de dados, análise estatística e quantitativa, modelos explicativos e preditivos e gerenciamento baseado em fatos para impulsionar decisões e ações", é utilizado de forma intercambiável com o termo mais genérico Analytics.

3. Big Data

Assim como Analytics, Big Data é um termo que surgiu a partir do mercado e passou a ser mais comumente utilizado em meados dos anos 2000. Em um primeiro momento, a maioria das pessoas associa Big Data com uma quantidade muito grande de dados. O conceito não está errado, mas é não possível definir Big Data apenas pela quantidade. Dados digitais, manuseados através de computadores, são utilizados desde os anos 1970, quando os Sistemas Gerenciadores de Bancos de Dados (SGBD) foram incorporados aos computadores da época, que eram chamados de "mainframes". Esses computadores gerenciavam grandes quantidades de dados, proporcionalmente à época, mas o termo Big Data ainda não era utilizado. Nos anos 1970, o mainframe da IBM System/370 tinha oito megabytes de capacidade de armazenamento. Atualmente, os smartphones mais baratos do mercado têm oito gigabytes de capacidade de armazenamento, ou seja, mil vezes mais capacidade que um mainframe há cerca de cinquenta anos. Dessa forma, apenas o volume de dados não é suficiente para definir Big Data, inclusive porque é muito difícil definir uma fronteira numérica.

Uma forma mais interessante de pensar em Big Data é sob a perspectiva do formato dos dados. Até a era do Big Data, os dados eram armazenados nos SGBDs de forma estruturada, ou seja, com uma estrutura rígida, fixa e em tabelas com linhas e colunas (muito semelhante, por exemplo, a uma planilha Excel atual). Hoje em dia, porém, é possível armazenar e manipular dados em qualquer formato, sobretudo aqueles chamados não estruturados, que envolvem, por exemplo, o texto de um livro, e-mails, postagens em redes sociais e, também, imagens e arquivos de audio e vídeo. Sob a perspectiva da variedade, pode-se assumir que, realmente, algo mudou a partir do Big Data.

Volume e variedade são parte dos chamados "Vs" do Big Data. Junto a eles, originalmente, existe uma terceira dimensão – ou terceiro "V" –, que diz respeito à velocidade. Esta, por sua vez, relaciona-se à dinâmica de obtenção e utilização dos dados, muitas vezes em tempo real, ao contrário dos tempos anteriores a era do Big Data, em que os momentos de obtenção, armazenamento e posterior utilização dos dados eram claramente distintos.

Finalmente, existe o quarto "V", que tem se mostrado cada vez mais

relevante. Ele representa a veracidade e está diretamente ligado à qualidade dos dados, sendo condição básica para a geração de informações fidedignas para tomadas de decisão. Aqui, aplica-se o conceito válido a todo o mundo da computação: lixo que entra, lixo que sai. Ou seja, jamais serão obtidos bons resultados se os dados de entrada não tiverem qualidade.

4. Inteligência Artificial

4.1 Histórico

Quando se fala em Inteligência Artificial (IA), George Ifrah – historiador e matemático francês – mencionou que é possível voltar até o século XVII com as máquinas desenvolvidas por Pascal e Leibniz que podiam emular a capacidade humana de somar e subtrair. Oficialmente, em 1943, foi publicado o primeiro trabalho reconhecido como inteligência artificial, elaborado por Warren McCulloch e Walter Pitts. Logo depois, em 1950, Alan Turing publicou o artigo "Computing machinery and intelligence", propondo um teste em que uma máquina emula o comportamento humano, que ficou conhecido como Teste de Turing e é estudado até hoje.

O estudo de IA apresentou grandes progressos entre os anos 1950 e 1960. Porém, em função da limitação das técnicas naquele tempo e da falta de capacidade computacional, o assunto ficou adormecido por quase três décadas. A partir dos anos 2000, com a disseminação de conhecimento facilitada pela internet, maior quantidade de dados disponíveis e novas técnicas sendo pesquisadas, a IA passou, novamente, a ser um tema amplamente estudado, ganhando cada vez mais notoriedade pelo fato de ser estudada não apenas pela academia, mas pelo próprio o mercado, que passou a adotá-la como solução tecnológica.

4.2 Tipos de IA

Para a maioria das pessoas, IA soa como ficção científica. Essa interpretação faz certo sentido, uma vez que IA costuma ser bastante comum em filmes desse gênero. Talvez o exemplo mais famoso seja I.A. (2001), dirigido por Steven Spielberg, cujo foco da narrativa é um robô programado para amar, que é adotado por um casal que perdeu o filho recentemente. IA, no entanto, para além dos estereótipos e do senso comum, é um campo muito vasto. Os smartphones usam inteligência artificial, diversos aplicativos são baseados nessa tecnologia, e mesmo os carros autônomos não existiriam sem ela. Entretanto, ainda que seja bastante presente no cotidiano da maioria das pessoas, seu uso raramente é percebido. Além disso, por ser uma "buzzword", muitas empresas dizem usá-la, ainda que isso não seja verdade.

Por ser um campo muito vasto e ter caráter mítico, é importante segregar os tipos de IA. Atualmente, existem muitas definições que emanam, principalmente, do mercado, e é interessante notar que não existem, mesmo na academia, definições que sejam aceitas de forma unânime. Para fins didáticos deste material, é importante apenas separar a inteligência em dois aspectos: a que existe e a que não existe (ainda).

4.3 Super Inteligência Artificial (ASI)

Na Super Inteligência Artificial (ASI)[1] – que é hipotética, uma vez ainda não se chegou lá –, máquinas teriam condição de interpretar o comportamento humano e sua inteligência, tornando-se autoconscientes, chegando a superar a capacidade de inteligência humana. Além disso, seriam capazes de desenvolver habilidades de interpretação impossíveis para o cérebro humano, que tem capacidade restrita, por mais inteligente que uma pessoa possa ser. Essa é a IA distópica que a maioria das pessoas conhece, em que os robôs superam e escravizam os humanos, de modo que a principal discussão sobre ela não é sobre como conseguir obtê-la, mas como fazer isso de forma ética e segura.

4.4 Inteligência Artificial Restrita (NAI)

O termo restrito, em Inteligência Artificial Restrita (NAI)[2], faz referência ao fato de os objetivos, nesse tipo de IA, serem claros e específicos. Trata-se de um tipo de IA em que uma solução é criada para executar uma única tarefa, e qualquer conhecimento obtido por meio dessa tarefa não é aplicado a outras, pelo menos não de forma automática e sem supervisão humana. Ao contrário da ASI, que procura imitar processos de pensamento complexos, a IA restrita é projetada para concluir com sucesso uma única tarefa sem assistência humana. Essa é, na prática, a IA que efetivamente está presente no dia a dia da maioria das pessoas, e sua aplicação pode ser encontrada em tradutores, reconhecimento de imagem, sistemas de recomendação entre outros.

4.4.1 Machine learning

"Machine learning" (ML) é a aplicação prática da inteligência artificial restrita. O termo foi cunhado em 1962, por Arthur Samuel, cientista da computação que fez carreira na International Business Machines Corporation (IBM). Samuel demonstrou que computadores podiam ser programados para jogar damas.

Em português, a tradução do termo ML é aprendizado de máquina, e

[1] Em inglês "Artificial Super Intelligence".

[2] Em inglês "Narrow Artificial Intelligence".

não é por acaso. Porém, para entender o conceito é preciso dar um passo atrás e compreender como funciona um computador.

Em um computador, existe o hardware – a máquina física, tangível – e o software – os programas. Os programas são escritos em linguagens específicas, como C++, Basic e Cobol (mais antigas) e Java e Python (mais recentes). Usar linguagens específicas (de programação) é a forma tradicional de programar um computador. A maioria dos programas utilizados no dia a dia, como os sistemas de gestão integrada, por exemplo, são elaborados de forma tradicional.

Na programação tradicional existe um componente chamado condicional. Na prática, é como um computador é comandado a fazer o que é preciso que seja feito. Em um "website", por exemplo, quando não é possível que o usuário passe para o próximo passo sem antes digitar o CPF, isso é feito através de um condicional: "se o CPF não foi digitado, então o usuário não deve passar para a próxima tela". Independente da linguagem utilizada, utiliza-se os condicionais, e sempre nesta forma: "se... então...". Basicamente, os condicionais são regras escritas por programadores.

ML, por outro lado, utiliza um paradigma diferente da programação tradicional: os dados são apresentados, o resultado desejado é declarado e o trabalho do computador é escrever um programa. No ML, o próprio computador é que define os condicionais, a partir dos dados. Uma vez que o computador aprende a partir dos dados, esse processo é chamado de aprendizado de máquina. Hoje, embora existam técnicas supermodernas, o princípio é o mesmo apresentado nos anos 1950. Um bom exemplo para explicar a diferença entre esses dois paradigmas é a solução criada para resolver um problema que afeta a todos: os spams, ou e-mails não solicitados.

A princípio, usava-se a programação tradicional para separar os spams de outros e-mails, por meio de regras como: se o imperativo "compre agora" está presente; se existe a palavra desconto; se a quantidade de letras maiúsculas é maior que 20% do volume total de letras da mensagem; se o remente não está na lista de contatos; se etc., então a mensagem é classificada como spam. Caso contrário, a mensagem é alocada na caixa de entrada.

Nos anos 2000, essa abordagem foi alterada. Imagine-se, por exemplo, que foram criados dois grupos de mensagens, cada um com 10 mil mensagens. Em um dos grupos, com certeza não há spam, e, no outro, todas as mensagens são spam. Usando-se um programa especialista em ML, os dados são apresentados com suas respectivas etiquetas, ou seja, spam ou não spam. O resultado é a criação de um programa capaz de fazer previsões: sempre que um novo e-mail chegar na caixa postal, ele será analisado pelo programa, que dará uma resposta como "a probabilidade deste e-mail ser um spam é de 80%". Com base nessa probabilidade, o serviço de e-mails poderá separar as mensagens que são spam daquelas que não são, e direcioná-las a caixa postal correspondente – caixa de entrada ou spam.

O mesmo princípio aplica-se a qualquer tipo de dado: dados estruturados, imagens, áudios, vídeos, textos, postagens de redes sociais entre outros. Vale ressaltar que, atualmente, quando IA é mencionada no ambiente de negócios, o que está em discussão é o uso de ML, ou seja, de uma ferramenta capaz de fazer previsões.

5. Relacionamento entre Analytics, Big Data e IA

"Information is the oil of the 21st century, and analytics is the combustion engine"[3]. Essa frase foi dita por Peter Sondergaard, vice-presidente sênior do Instituto Gartner, em 2011, em uma palestra na Flórida, e tornou-se bastante popular. É ideal, também, para introduzir o assunto sobre o relacionamento entre Analytics, Big Data e IA.

Dados – e não apenas os digitais – não têm valor algum. Quando se fala em Big Data, a afirmação também é válida. Trata-se de uma afirmação forte, sem dúvida, e refere-se a "information" na frase de abertura desta seção. O sentido da afirmação é que os dados precisam ser manipulados, visando algum tipo de resultado, e esse resultado, então, é que tem valor. É aí que entra o termo analytics. Em suma, o que Peter Sondergaard quiz dizer é que a informação é necessária, mas não é suficiente. Fazendo-se uma analogia com um carro, a informação seria o combustível, ou seja, sem ela o carro não anda. No entanto, apenas combustível não basta, é preciso que o carro tenha um motor que gere movimento.

Dessa forma, pode-se entender que existe a seguinte relação entre Big Data e Analytics: são termos relacionados e interdependentes, ou seja, um depende do outro para que dados gerem valor. Isso porque os dados em si não geram valor, mas as análises geradas a partir deles sim.

5.1 Onde entra, então, a IA?

Pode-se entender Analytics como um grande guarda-chuva que envolve o tratamento de dados digitais. Dentro dele cabem inúmeras formas de tratamentos, muitos dos quais não são novos. Técnicas de visualização de dados, regressão, simulação e machine learning, por exemplo, estão disponíveis há muitos anos. IA, particularmente a IA estreita, é uma forma de tratamento que se enquadra dentro desse grande guarda-chuva que é chamado de Analytics.

5.2 Cientistas de dados

Segundo Davenport e Patil (2012), cientista de dados é a profissão mais

[3] Em tradução literal "A informação é o petróleo do século 21, e a análise é o motor de combustão".

"sexy" do século XXI. Essa afirmação, título de artigo publicado no periódico Harvard Business Review, é amplamente conhecida no mercado e bastante citada na academia.

Cientistas de dados estão entre os atores que exploram o potencial do Big Data para gerar conhecimento, bem como criar novas formas de valor que transformam as organizações e a sociedade. Trata-se de uma profissão nova e com características extremamente técnicas, como pode ser visto no modelo conceitual a seguir (Figura 1), considerando definição, formação, habilidades e ferramentas que os cientistas de dados utilizam.

Figura 1. Modelo conceitual da profissão de cientista de dados
Fonte: Castello (2021).

Do ponto de vista desta série acadêmica, o aspecto de maior relevância desse tema – cientista de dados – é entender que existem profissionais especializados para trabalhar com dados. Por ser uma profissão nova, esses profissionais ainda são raros e contratá-los não é simples, porque suas habilidades, bem como seu nível de senioridade nessas habilidades, podem variar bastante.

5.3 Citizen data scientist

Pela dificuldade na contratação de cientistas de dados, muitas empresas estão se adaptando e descobrindo que não precisam ter um cientista de dados

altamente qualificado para muitas tarefas que envolvem a utilização de dados digitais. Essa adaptação, por sua vez, tem dado origem a um novo tipo de profissional: o citizen data scientist (ou cientista de dados cidadão).

O citizen data scientist é um profissional que tem habilidade de manipular dados digitais, elaborar visualizações de dados e até mesmo criar modelos preditivos baseados em machine learning, mas cuja principal função está fora do campo das atividades de um cientista de dados.

Ao contrário do cientista de dados, que é um profissional dedicado, o citizen data scientist é um profissional de contabilidade, da área comercial ou, ainda, de recursos humanos, mas que possui certas habilidades no tratamento de dados digitais. Conhece, por exemplo, estatística, mas não na profundidade de um cientista de dados. O citizen data scientist nasceu para preencher a lacuna que existem entre os analistas tradicionais (que utilizam ferramentas simples do dia a dia) e os cientistas de dados (que fazem análises avançadas).

Ser um citizen data scientist é uma oportunidade para qualquer profissional, de qualquer área de negócio, em qualquer mercado. Significa acrescentar uma habilidade que pode promover destaque junto aos pares, alavancado promoções e melhorando a empregabilidade. Por analogia, é o mesmo princípio de conhecer uma língua estrangeira, como inglês, espanhol ou mandarim: não é uma profissão em si, mas ser fluente em uma delas traz maior chance de destaque e abre novas possibilidades de empregabilidade.

Não existe um consenso sobre as habilidades de um citizen data scientist, porém conhecer visualização de dados, conceitos de bancos de dados, e até mesmo uma linguagem de programação, são habilidades que serão apreciadas em qualquer currículo.

6. Infraestrutura para trabalhar com dados: hardware e nuvens

Dados digitais precisam estar armazenados em computadores. E a história dos computadores, por sua vez, é fascinante, e remonta ao século XIX, com as pesquisas e experimentos de Charles Babbage e Ada Lovelace (Isaacson, 2014)[4]. Nas décadas de 60 e 70, a IBM reinava com seus mainframes nas empresas, no governo e junto aos militares. Já entre 1980 e 1990, os microcomputadores tornaram-se populares nas casas e nas empresas, já em uma perspectiva global.

Até os anos 2000, a utilização dos computadores existia apenas "on premises", ou seja, localmente, dentro das instalações das empresas. Na época, eram comuns as instalações chamadas Central de Processamento de Dados

[4] O livro "Os Inovadores", de Walter Isaacson, é uma leitura agradável – com uma pesquisa muito bem-feita – que fala sobre inovações em tecnologia desde o século XIX, quando Ada Lovelace e Charles Babbage escreveram um ensaio sobre como poderia funcionar uma máquina de processar e resolver problemas.

(CPDs), ou, que, em geral, possuíam controle de acesso e potentes unidades de ar-condicionado para manter os computadores refrigerados.

Com a popularização da conectividade, empresas como IBM, Amazon, Google e Microsoft começaram a construir CPDs gigantescos, chamados "Data Centers", onde clientes poderiam manter seus computadores em uma infraestrutura segura, com proteção contra falta de energia e técnicos à disposição. Organizações podiam, então, terceirizar seus CPDs junto a esses prestadores de serviço.

Juntamente com a conectividade, também ganhou impulso o conceito de virtualização que, de forma simples, pode ser entendido como programas que, instalados em potentes computadores, podem criar diversos outros computadores virtuais. A virtualização possibilitou o gerenciamento mais eficiente das necessidades de tecnologia da informação para as organizações.

No passado, caso houvesse, por exemplo, a necessidade de uma máquina com maior capacidade de armazenamento, era necessário comprar uma novo "hard-drive". Esse processo era lento, porque era preciso realizar a compra e aguardar a entrega, além de ter profissionais habilitados para fazer a troca do equipamento e outros ajustes necessários, como a instalação dos programas e a realização de retorno dos dados antigos para o novo equipamento.

Atualmente, a maioria das organizações mantém sua infraestrutura de tecnologia de informação – servidores (virtuais ou não), programas, dados – junto a prestadores de serviços, como os mencionados anteriormente. Nesses casos, considera-se que a infraestrutura está na nuvem.

A maior vantagem do uso da nuvem é a segurança e a flexibilidade. Segurança, porque são locais fisicamente controlados e possuem dispositivos robustos contra falta de energia, e flexibilidade em relação às necessidades, que podem mudar a todo momento. Na nuvem, se uma organização precisa de maior capacidade de armazenamento, ela simplesmente faz a contratação de mais espaço junto ao prestador e, imediatamente, a nova capacidade é disponibilizada.

6.1 O que é melhor: equipamentos locais ou nuvem?

Apesar da utilização de nuvens ser cada vez mais comum na maioria das organizações, nuvens, em geral, têm usos específicos. Quando se trata de dados digitais, particularmente uma grande quantidade de dados, a nuvem é a melhor opção. Não apenas por haver flexibilidade para contratação de mais ou menos espaço, como também pelo fato de os prestadores oferecerem ferramentas específicas para, por exemplo, manipular os dados, gerar visualizações e criar modelos de inteligência artificial.

Porém, muito pode ser feito localmente. Atualmente, notebooks têm potência suficiente para usar analytics em muitas tarefas. Por exemplo, exportar um relatório de um sistema integrado de gestão, como SAP ou Totvs, e gerar visualizações em softwares como PowerBI ou Tableau. Ou, ainda, agregar

informações de clientes e vendas e construir um modelo preditivo de crédito, usando machine learning.

Atualmente, é possível juntar o melhor dos dois mundos: trabalhar localmente com dados no próprio notebook e acessar uma máquina que se encontra na nuvem, inclusive transferindo dados entre o ambiente local e o ambiente na nuvem. A decisão por um ou por outro ambiente dependerá do volume de dados em questão e da capacidade de processamento necessária para manipulá-los.

7. Ferramentas e linguagens para trabalhar com dados

A seguir são apresentadas algumas opções de ferramentas e linguagens de programação comumente utilizadas para tratamentos de dados digitais.

7.1 Gretl

Gretl é um programa gratuito multiplataforma (funciona em Windows, Mac e diversos tipos de Unix/Linux) e disponível em mais de dez idiomas, inclusive em português. Foi desenvolvido originalmente para o público de ciências econômicas, mas é aplicável para diversas necessidades atuais de tratamento de dados digitais. Possui diversas funcionalidades gráficas e ferramentas estatísticas, inclusive regressão linear e logística, que hoje são consideradas técnicas de machine learning.

O Gretl é mais limitado que linguagens de programação como Python e [R], mas tem uma vantagem importante: não necessita de conhecimentos de programação para ser utilizado, já que todas as funcionalidades são acessadas por meio de um menu. Para quem não tem habilidades em programação, mas tem interesse em começar a tratar dados, e deles extrair valor, Gretl é a melhor opção.

7.2 Visualização de dados

Visualização de dados, em si, não é um assunto novo. É utilizada há muitos anos pelas organizações para acompanhamento de indicadores, também chamados de KPI's ("key performance indicators").

No entanto, a partir de 2010, surgiu uma nova categoria de programas de visualização, inaugurando um conceito denominado "self-service dataviz" ("dataviz" é comumente encontrado como sinônimo de "data visualization").

A diferença dessa nova categoria é que, enquanto os softwares mais antigos precisam de consultores especialmente treinados para construir as visualizações, ela é voltada para os usuários finais, de modo que construir visualizações passou a ser uma tarefa tão simples quanto usar planilhas eletrônicas. Os principais softwares de self-service dataviz, de acordo com o Gartner (2021), são o PowerBI (Microsoft), Tableau e Qlik (Figura 2).

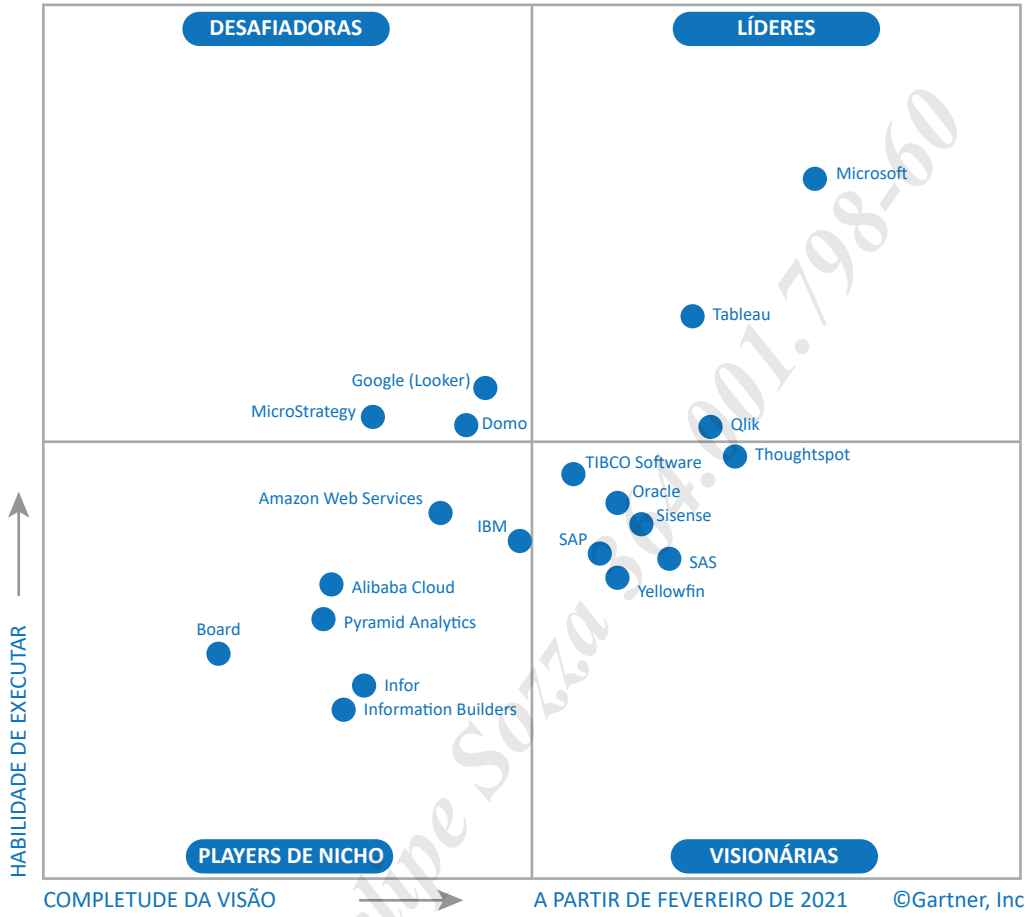

Figura 2. Principais softwares de self-service dataviz

Fonte: Gartner (2021).

Atualmente, aprender a utilizar tais ferramentas é um diferencial, mas no futuro será uma habilidade tão comum quanto conhecer os pacotes de escritório (edição de texto, apresentações e planilhas eletrônicas).

A grande vantagem dessas ferramentas é a facilidade de uso. Existem versões gratuitas em português, além de tutoriais que tornam mais fácil o aprendizado dessas ferramentas, principalmente por meio de autoestudo. A possibilidade de filtrar e interagir com os dados torna as análises não apenas mais fáceis, mas também mais profundas.

7.3 Linguagens de programação

Apenas com linguagens de programação é possível extrair todo o potencial das possibilidades de extração de valor dos dados. Atualmente, no ambiente da ciência de dados, as linguagens Python e [R] são as mais utilizadas.

[R] é uma linguagem que surgiu com o objetivo de tratar dados com fins estatísticos. Por ser uma linguagem gratuita e de código livre, diversas melhorias foram feitas por sua comunidade, de forma que hoje os principais modelos de "machine learning" podem ser elaborados a partir dela. Python também é uma linguagem gratuita e, na comunidade de ciência de dados, considerada a principal linguagem para inteligência artificial.

Python é a melhor opção para quem não tem conhecimento de programação, e deseja aprender uma linguagem. Trata-se de uma linguagem fácil, com possibilidades que vão muito além do [R], além de possuir uma comunidade global enorme.

8. Introdução às metodologias de trabalho com dados

Atualmente, o termo "Data Mining" (DM), ou mineração de dados, é utilizado pelo mercado, de forma geral, para nomear o processo de extração de valor dos dados. A origem do termo, no entanto, vem de uma metodologia denominada "Knowledge Discovery in Databases" (KDD), ou extração de conhecimento, sendo DM apenas uma das fases dessa metodologia. Na prática, KDD e DM são, muitas vezes, utilizadas de forma intercambiável, embora sejam vistas pela maior parte das pessoas como atividade, e não como metodologia.

Há diversas metodologias no mercado, sendo as mais consagradas para analisar dados de forma estruturada: KDD, Sample, Explore, Modify, Model, Assess (SEMMA)) e Cross Industry Standard Process for Data Mining (CRISP-DM). KDD é a mais antiga e talvez mais conhecida, e SEMMA é uma metodologia com maior foco nos aspectos técnicos. No entanto, neste material serão exploradas características do CRISP-DM, atualmente o mais utilizado, além de ser considerado o padrão "de-facto" de mercado.

8.1 Cross Industry Standard Process for Data Mining (CRISP-DM)

CRISP-DM foi criado no final dos anos 1990 por um consórcio formado por três organizações: DaimlerChrysler, SPSS e NCR. A ideia por trás do CRISP-DM era prover uma metodologia padrão, gratuita e amplamente disponível para os profissionais envolvidos nas atividades de extração de valor de dados digitais. A primeira versão foi lançada em 2000.

CRISP-DM é uma forma estruturada de planejar um projeto que utiliza dados digitais. Tem caráter bastante pragmático, é flexível e útil para resolver problemas de negócios complexos. Ainda que seja criticado por não ter

elementos de gerenciamento de projetos, o fato de ser uma metodologia genérica para uso em várias indústrias conta a seu favor, assim como o fato de não estar vinculada a uma ferramenta específica.

Originalmente apresentada como um modelo de processo, relacionando os estágios do ciclo de vida dos dados, CRISP-DM hoje é conhecida como uma metodologia (Figura 3).

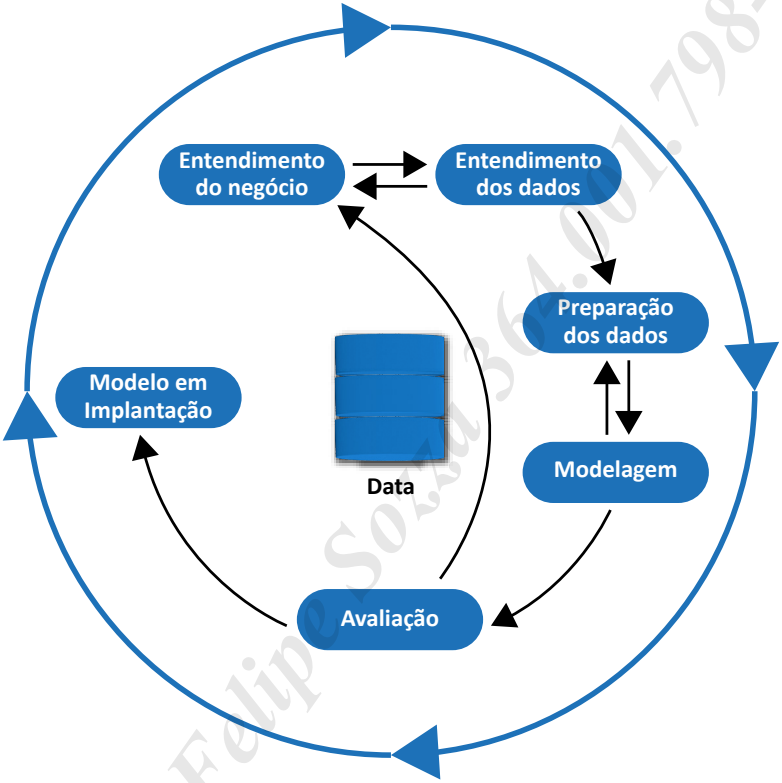

Figura 3. Estágios da Metodologia CRISP-DM
Fonte: Smart Vision Europe (2021).

É importante notar que, muitas vezes, nem todas os estágios do processo são necessários, do mesmo modo que se pode nomes diferentes para a mesma coisa. Por exemplo, há uma preferência no mercado pelo termo análise exploratória, mas, na prática, trata-se do estágio "data understanding" do CRISP-DM. Sendo assim, o mais importante é entender a essência de cada estágio, uma vez que uma das características mais marcantes do CRISP-DM é sua flexibilidade.De forma resumida, a seguir serão descritos os principais estágios do CRISP-DM.

- **Primeiro estágio – Entendimento do negócio**

Há uma frase bastante conhecida que se aplica perfeitamente ao primeiro

estágio, que é o entendimento do negócio: "para quem não sabe aonde quer chegar, qualquer caminho serve". É comum ver projetos de análise de dados não darem resultados satisfatórios simplesmente por não terem clareza quanto onde querem chegar.

O principal objetivo do primeiro estágio do CRISP-DM é entender o que se espera obter na perspectiva de negócio, bem como realizar uma avaliação de alto nível sobre a possibilidade de alcançar tais objetivos. O que interessa neste estágio é saber se o resultado será um aumento na receita, redução de custo, melhora do "market share" ou eficiência logística, entre tantos outros objetivos que podem impactar o negócio. Este estágio compreende, além dos objetivos, a identificação de recursos e restrições, a análise dos riscos e a declaração dos benefícios esperados no projeto.

- **Segundo estágio – Entendimento dos dados**

No segundo estágio do CRISP-DM é preciso coletar os dados que foram identificados na fase anterior, quando os recursos foram planejados, e carregá-los de forma que possam ser acessados pelas ferramentas selecionadas. Esta fase envolve a exploração de dados e, principalmente, a análise sobre sua qualidade.

- **Terceiro estágio – Preparação dos dados**

Neste estágio, efetivamente, inicia-se o processo de montagem da base de dados que será usada no estágio seguinte e envolve seleção, limpeza, transformação e, também, integração entre diversas bases de dados.

- **Quarto estágio – Modelagem**

O CRISP-DM envolve um estágio de modelagem porque a metodologia foi criada com foco de mineração de dados, mas é possível notar que os primeiros três estágios aplicam-se a praticamente qualquer projeto que trabalhe com dados. Na fase de modelagem, utiliza-se uma base de dados única, que já está validada, limpa e transformada, apenas com as informações consideradas relevantes. A essa base de dados única dá-se o nome de dataset[5].

Nesta fase, define-se a técnica de modelagem a ser utilizada, considerando como critérios a precisão, a velocidade de aprendizado e o nível de explicação dos resultados.

- **Quinto estágio – Avaliação dos resultados**

Quando a avaliação do modelo é feita, no estágio anterior, a maior preocupação é com a precisão. Neste estágio, no entanto, o objetivo é avaliar se o modelo atende os objetivos definidos para o projeto.

- **Sexto estágio – Modelo em produção ("deploy")**

Colocar um modelo em produção, termo que é usado para definir que o modelo irá atuar sobre dados reais, é algo tão complexo quanto um sistema de

[5] Dataset é um termo em língua Inglesa que significa conjunto de dados. Está grafada sem aspas porque, mesmo sendo um anglicismo, é uma palavra corriqueira no mundo dos dados.

computação tradicional, e necessita de pessoal técnico especializado.

8.2 Uma nota final sobre metodologias

Conforme mencionado anteriormente, há diversas metodologias no mercado. CRISP-DM tornou-se o padrão de mercado, mas não necessariamente é a melhor opção para todas as necessidades. Se o projeto é simples, muitas vezes SEMMA será mais útil.

Quando às metodologias para trabalhar com dados, o mais importante é que se utilize alguma. Ou seja, é imperativo que uma metodologia seja utilizada. Como a própria palavra diz, usar uma metodologia traz método ao projeto, e o encadeamento de etapas lógicas minimiza o risco de falhas do projeto.

9. Técnicas de machine learning

As técnicas de machine learning podem ser divididas em três grandes grupos: modelos supervisionados, não supervisionados e de aprendizado por reforço. Nesta série acadêmica serão tratados os dois primeiros, uma vez que o aprendizado por reforço envolve um grupo de técnicas que são complexas e tem usos bastante específicos.

9.1 Modelos supervisionados

O grupo de técnicas mais comum em machine learning é o de aprendizado supervisionado, no qual são fornecidos aos programas especialistas em machine learning os dados e as respostas conhecidas, e o aprendizado ocorre com base nas características dos dados.

Os modelos supervisionados são divididos em dois tipos de técnicas: classificação e regressão.

As técnicas de classificação são algoritmos que, como o próprio nome diz, atuam para classificar os dados em categorias. O resultado de modelos preditivos de classificação retorna à probabilidade de determinado item, com determinadas características, pertencer a determinada classe.

O exemplo sobre spam, citado anteriormente, é um problema resolvido por classificação. A partir de um modelo preditivo, um novo e-mail é apresentado e o modelo calcula a probabilidade de a mensagem ser spam. A classificação é realizada com base nessa probabilidade, definindo se o novo e-mail será incluído na caixa de entrada ou na pasta de spam.

Já as técnicas de regressão, são utilizadas na previsão de dados em respostas contínuas, ou seja, quando se está procurando um número absoluto. Por exemplo, pode-se criar um modelo preditivo de regressão para prever o valor de locação de um imóvel a partir de suas características, como a metragem, o número de quartos, a existência ou não de piscina na propriedade, etc. Esse

modelo pode ser utilizado, então, pegando-se as características de qualquer imóvel, de forma a prever qual será o valor esperado do aluguel.

Assim, utilizam-se modelos supervisionados para resolver problemas quando há clareza quanto ao que se está procurando: se um e-mail é spam ou não, qual o valor esperado de um aluguel, etc.

9.2 Modelos não supervisionados

Ao contrário dos modelos supervisionados, as técnicas não supervisionadas são utilizadas para realizar agrupamentos sem um objetivo previamente definido, procurando nos dados padrões que não são evidentes se analisados manualmente. Entre as diversas técnicas de modelos não supervisionados destacam-se, principalmente, as de "clustering" e a de regras de associação.

Clustering, que em traduzido significa agrupamento, é uma técnica utilizada para fazer segmentações com base nas características dos dados. É muito utilizada, por exemplo, para realizar segmentação de clientes. É importante notar que as técnicas de clustering não fornecem um resultado prático de utilização imediata, mas "insights", ou seja, informações que podem ser utilizadas para definição de estratégias, entre outras aplicações.

As técnicas de regras de associação, como o próprio nome diz, procuram buscar associações em um conjunto de dados. Um dos usos clássicos dessas técnicas ocorre na indústria de varejo.

A partir de um modelo que leve em consideração cestas de produtos (na prática, a compra de um cliente, ou tudo que consta em um cupom fiscal), um resultado possível é a associação entre dois ou mais produtos. Com base em uma forte associação entre os produtos macarrão e molho de tomate, por exemplo, uma loja pode fazer uma promoção de determinada massa de macarrão e, ao lado, colocar uma prateleira com opções de molho de tomate. A promoção do macarrão diminui a margem de lucro, mas atrai o público para a loja, que, possivelmente, levará também o molho de tomate, cuja margem de lucro é mais alta.

10. Exemplos de aplicação no mundo dos negócios

Antes de mais nada, é importante entender que IA é meio, não fim. Por isso, o foco deve estar sempre no problema de negócio a ser resolvido. Assim como utiliza-se um sistema integrado como SAP ou Totvs para melhorar processos e ganhar maior controle e produtividade, usa-se IA como meio para buscar, como resultado, algum tipo de eficiência operacional, que, de forma bem simples, significa reduzir custos ou aumentar a receita. A Figura 4, a seguir, apresenta o que pode ser chamado de mapa do calor em relação à relevância técnica da IA para as indústrias:

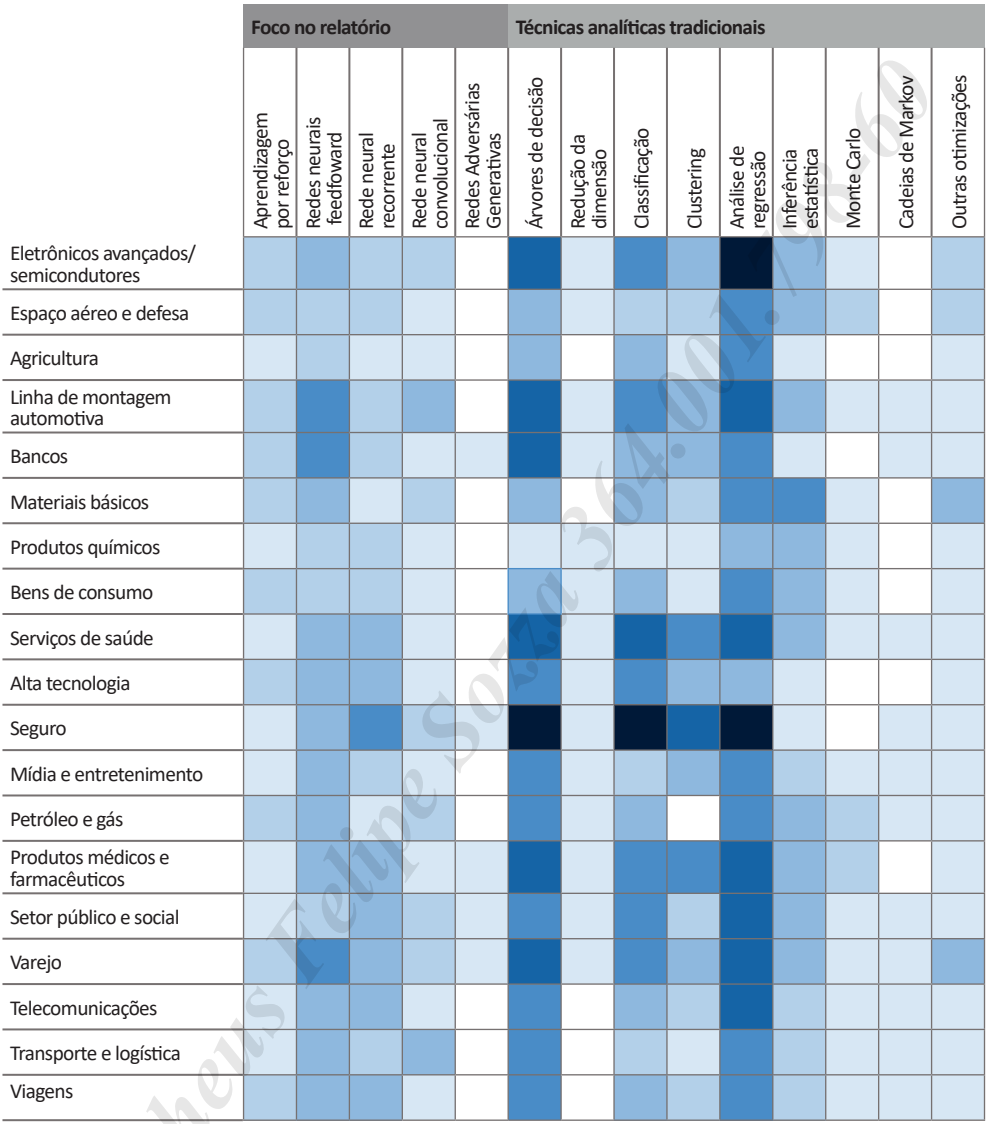

Figura 4. Mapa de calor – relevância técnica da IA para as indústrias
Fonte: Chui et al. (2018).

Pode-se, inclusive, usar IA para aumentar a satisfação do cliente, porém o resultado esperado é o aumento da receita. Clientes satisfeitos compram mais e de forma recorrente. A seguir, são apresentados alguns exemplos.

➢ Previsão de demanda

Pode-se usar IA para prever demandas futuras por produto com base no histórico de vendas. Isso possibilita melhor gerenciamento de produção e de estoque, evitando falta de produtos e menor nível de imobilização de capital.

➢ Roteirização logística

Encontrar as melhores rotas é importante para reduzir custos logísticos, prover insumos para a produção no momento adequado e entregar os produtos corretos para os clientes corretos no menor tempo possível. Por meio da utilização de dados como tráfego em cada região, bem como incidência de assaltos e acidentes, é possível encontrar não apenas a rota mais curta, mas também a mais segura. Não se trata de um modelo de IA simples de ser elaborado, mas existem diversas empresas no mercado que oferecem esse serviço.

➢ Manutenção preventiva

Máquinas, equipamentos e veículos precisam estar sempre em boas condições de uso, uma vez que paradas não planejadas impactam a rotina das organizações. Atualmente, com o advento de internet das coisas (IoT), é possível acompanhar, em tempo real, condições de operação como temperatura, pressão, vibração, consumo de combustível e energia. Um modelo bastante utilizado nesses casos é chamado detecção de anomalias, que é uma das aplicações de IA.

➢ Atendimento a Clientes

Chatbots[6] já são comuns em diversas atividades, particularmente no atendimento a clientes. Embora esses programas não sejam IA, utilizam-se delas em sua sua operação. Uma das coisas que os chatbots permitem é a interpretação do que o cliente deseja, através da técnica de processamento de linguagem natural, ou "Natural Language Processing" (NLP). Com chatbots, pode-se ligar para o banco e dizer "qual o meu saldo", em vez de, como antigamente, ouvir uma longa mensagem com uma das opções dizendo "digite 2 para saldo".

➢ Sistemas de recomendação

Amplamente utilizados, servem não apenas para recomendar filmes e músicas nas plataformas de "streaming", mas também recomendar produtos para clientes. A Amazon utiliza recomendações há anos, com algoritmos sofisticados que identificam ligações entre produtos. Por exemplo, alguém que compra um pacote de lâmina de barbear verá uma recomendação de espuma de barbear. Ainda, tendo acesso a todos os produtos comprados por todos os usuários, o sistema de recomendação consegue identificar produtos que podem ser do interesse do cliente que visita o site.

[6] Chatbots são programas de computador que utilizam IA cada vez mais aperfeiçoada para imitar conversas com usuários de várias plataformas e aplicativos.

- Crédito para clientes

Embota já existam algoritmos para avaliação de crédito há muito tempo, a possibilidade de usar IA permite que as avaliações sejam mais precisas do que nos métodos tradicionais.

Dicas complementares

- Artigo Tutorial: "Big Data Analytics: Concepts, Technologies, and Applications", de Hugh J. Watson (Association for Information Systems). Trata-se de um material bastante denso e técnico, mas fornece uma visão geral sobre big data, bancos de dados, Hadoop e Map Reduce.
  Onde buscar: Communications of the Association for Information Systems 34: 1247-1268. http://aisel.aisnet.org/cais/vol34/iss1/65
- Livro "Os Inovadores", de Walter Isaacson. É uma leitura agradável com uma pesquisa muito bem-feita, e fala sobre inovações em tecnologia desde o século XIX, quando Ada Lovelace e Charles Babbage escreveram um ensaio sobre como poderia funcionar uma máquina de processar e resolver problemas.
  Onde buscar: https://www.amazon.com.br/Os-inovadores-Walter-Isaacson/dp/8535925023
- Livro "Storytelling com Dados. Um Guia Sobre Visualização de Dados Para Profissionais de Negócios", de Cole Nussbaumer Knaflic, um excelente livro sobre storytelling e visualização de dados.
  Onde buscar: https://www.amazon.com.br/Storytelling-com-Dados-Visualiza-C3-A7-C3-A3o-Profissionais-dp-8550804681/dp/8550804681
- Discussão "Notes from the AI frontier: insights from hundreds of use cases", publicada pelo McKinsey Global Institute (MGI). Ainda que não seja recente (abril de 2018), o material apresenta, de forma bastante abrangente, onde os principais tipos de IA podem ser utilizados em diversas indústrias.
  Onde buscar: http://fabianocastello.com.br/ia.repo/2%20CASES%20B%20MGI_Notes-from-AI-Frontier_Discussion-paper.pdf
- Filme "O Jogo da Imitação", biografia de Alan Turing, matemático brilhante que liderou um grupo da inteligência britânica na missão de decifrar os códigos da máquina Enigma, usada pela Alemanha Nazista durante a Segunda Guerra Mundial. Alan Turing é uma referência no mundo da computação e considerado um dos pais da IA. Onde buscar: https://www.primevideo.com/detail/0RE594TD30MV9KITHPFBQ312IA

RECAPITULANDO

Esta série acadêmica tem por objetivo proporcionar ao aluno uma visão geral sobre aspectos relacionados a dados digitais, desde as principais definições e diferenças entre termos até tecnologias e infraestruturas necessárias para sua utilização. Este material foi elaborado com vistas a oferecer aos alunos que não possuem formação técnica em áreas de tecnologia e afins, possibilitando que se beneficiem da aplicação dessas tecnologias, ainda que não possuam conhecimento profundo que permita liderar sua implementação. Foram apresentados definições e conceitos de Analytics, Big Data e Inteligência Artificial, bem como a relação entre esses termos. Além disso, discutiu-se a nova profissão de cientistas de dados e, considerando a falta de oferta desses profissionais, como o mercado está suprindo a demanda através de profissionais que possuem habilidades para trabalhar com dados dentro de suas próprias profissões, por meio dos chamados citizen data scients. Este material apresenta, também, informações sobre infraestrutura de hardware e o conceito de nuvem, exemplos de ferramentas para manipulação e visualização de dados, considerações sobre linguagens de programação e metodologias que suportam projetos com foco em dados digitais. Finalmente, as principais técnicas de machine learning foram introduzidas, assim como a principal aplicação utilizada atualmente em inteligência artificial, segregadas em supervisionadas e não supervisionadas. Por fim, conclui-se a série acadêmica com a tentativa de evidenciar a IA como uma ferramenta, não como um objetivo em si, por meio da citação de alguns exemplos de como a inteligência artificial permite melhorar a eficiência operacional das organizações.

Referências

Castello, F. 2021 Cientistas de dados: proposta de um modelo conceitual considerando sua definição, sua formação, suas habilidades e as ferramentas que utilizam. Dissertação (Mestrado). Universidade de São Paulo, São Paulo, SP, Brasil.

Chui, M.; Manyika, J.; Miremadi, M.; Henke, N.; Chung, R.; Nel, P.; Malhotra, S. 2018. Notes from the AI frontier insights from hundreds of use cases. McKinsey & Company. . Disponível em: <http://fabianocastello.com.br/ia.repo/2%20CASES%20B%20MGI_Notes-from-AI-Frontier_Discussion-paper.pdf>.

Davenport, T.H.; Harris, J.G. 2007. Competing on Analytics: The New Science of Winning. Harvard Business School Press, Boston, MA, EUA.

Davenport, T.H.; Patil, D.J. 2012. Data Scientist: The Sexiest Job of the 21st Century. Harvard Business Review. Disponível em: <https://hbr.org/2012/10/data-scientist-the-

sexiest-job-of-the-21st-century>.

Gartner. 2021. Gartner Research: Gartner Magic Quadrant for Analytics and Business Intelligence Platforms. Disponível em: <https://www.gartner.com/en/documents/3996944>.

Isaacson, W. 2014. Os inovadores. Companhia das Letras, São Paulo, SP, Brasil.

Smart Vision Europe. 2021. Background to SPSS Modeler and overview of the CRISP DM methodology. Disponível em: <https://www.sv-europe.com/courses/introduction-spss-modeler/lessons/background-to-spss-modeler-and-overview-of-the-crisp-dm-methodology/>.

EDITORA pecege

ISBN 978-85-92582-50-0