Sumário

Ciência de Dados

Arthur H.A. Farias, Samuel K., Lucas L. Arruda, Gabriel O. Vieira

[1] Faculdade de Computação
Universidade Federal de Mato Grosso do Sul (UFMS)
Campo Grande – Brasil

1. Armazenamento de Dados.

Desde sempre o ser humano é um tanto quanto curioso e perceptivo, buscando meios e formas diferentes para as mais diversas propostas, querendo ser reconhecido posteriormente por isso, fomentando formas de relatos e armazenamento de ”dados”e informações. A escrita embrionária em pedras, e mais adiante em papel são grandiosissimos exemplos dessa necessidade de resguardar e transmitir conteúdos para as gerações seguintes.

Mas a evolução tecnológica exigiu que essas formas fossem superadas e evoluíssem, permitindo o armazenamento em maior escala e múltiplo processamento, sendo desenvolvidos diversos métodos para guardar tais concepções, como por exemplo: Tear de Jacquard (madeira perfurada, que demonstrava certos padrões de tecelagem); Tambor Magnético (percursor do HD moderno, guardando bits 0 e 1); HD (sistema de armazenamento em disco, possibilitando o armazenamento de até 5 megabytes), e por fim os cd's e dvd's (O CD-R, composto por seis camadas, permitindo a gravação de dados num drive comum de CD-R. Possibilitando 650 (ou 700) MB de capacidade de armazenamento de dados e que se for usado para gravação de áudio, possui 74 (ou 80) minutos de capacidade. Os Dvd´s vieram logo em seguida, com capacidade de armazenamento maior que aos dos CD´s (4,7 GB), devido a uma tecnologia óptica superior.), Pen Drives, SSD's, e a mais relevante do momento, a Nuvem, que são os aparelhos e serviços predominantemente utilizados pela comunidade geral na contemporaneidade.

2. Origem e suas definições.

Na atualidade, a quantia de informações que são produzidas e divulgadas de maneiras ”online”, é devastadora, através de um simples ”like”em quaisquer redes socias, ou até mesmo dando um ”upload”em uma foto ou vídeo, produz uma enorme e valoroza base para futuras manipulações. Na mesma medida, com esse substancial aumento na produção e armazenamento de dados, todos os componentes de processamento vem tendo de evoluir de forma exponencial, praticamente dobrando sua eficácia ano após ano. E com esse grande aumento na quantidade de dados e na capacidade de processamento, um novo conceito surgiu, o **Big Data**, reunião generalizada de tais conceitos.

Através dessa relevante necessidade de analisar e extrair numerosos dados e informações consideradas ”úteis”, surgiu a Ciência de Dados, amplamente considerada como a mais recente ”Business Intelligence”. Porém, mesmo com tamanhas semelhanças, as duas possuem manuseios e propósitos um pouco diversos; Já que, enquanto o Business Intelligence busca pesquisar dentro de dados ”descritivos”ou ”passados”, visando responder acontecimentos anteriores, a ciência de dados, permeia conteúdos de análises ”preditivas”, propondo resoluções do que pode vir a acontecer, ou seja, sobre o futuro de determinado serviço.

3. Ciência ou Estatística?

O termo ”Ciência de dados”, vem constantemente tornando-se mais e mais popular, praticamente ”explodindo”em todos ambientes, tanto de negócios quanto de pesquisas, sendo cada vez mais necessários novos ”cientistas de dados”, existindo uma quantia exorbitante de vagas no mercado para tal profissional. No entanto, muitos dos que são contrários ao segmento julgam e predizem que não há quaisquer diferenças entre o ”cientista de dados”, e um ”estatístico”, tentando diminuir sua relevância.

Em artigo na Forbes, por Gil Press, argumenta que: ”...a ciência de dados é uma buzzword sem uma definição clara e simplesmente substituiu ”analista de negócios”no contextos das programas de graduação.”Na seção de perguntas e respostas de seu principal discurso na Reuniões Estatísticas da American Statistical Association, o notório estatístico aplicado Nate Silver disse: “Eu acho que cientista de dados é um termo sexualizado para um estatístico .... A estatística é um ramo da ciência. O cientista de dados é um pouco redundante de alguma forma e as pessoas não devem repreender o termo estatístico.”. Simultaneamente, no setor de negócios, vários pesquisadores e analistas afirmam que os cientistas de dados, por si só, estão longe de ser suficientes para conceder às empresas uma vantagem competitiva real. e consideram os cientistas de dados como apenas uma das quatro maiores famílias de empregos que as empresas precisam para usar grandes dados com eficiência, a saber: analistas de dados, cientistas de dados, desenvolvedores e engenheiros de dados.

Houve até mesmo fundamentações públicas e internacionais sobre o tema. Em 2015, a American Statistical Association fez uma declaração através de um comunicado de imprensa que procura apaziguar essa questão. Basicamente ela afirma que as ciências são complementares, e a estatística procura fomentar um relacionamento mais próximo à ciência de dados para benefício mútuo, não sendo necessários embates, já que ambas devem e podem trabalhar em conjunto para um ”bem maior”.

4. Modernização do termo.

A modernização e independência da termologia e matéria ”Ciência de Dados”, é por muito atribuída a William S. Cleveland. Em seu artigo de 2001, ele defendeu uma expansão da estatística além da teoria para áreas técnicas; ”...isso mudaria significativamente o campo, justifica um novo nome. A Ciência de Dados”. Tornou-se mais amplamente usada nos anos seguintes: em 2002, o Comitê de Dados para Ciência e Tecnologia lançou o Data Science Journal. Em 2003, a Columbia University lançou o The Journal of Data Science. Em 2014, a Seção de Aprendizagem Estatística e Mineração de Dados da American Statistical Association mudou seu nome para Seção de Aprendizagem Estatística e Ciência de Dados, refletindo a crescente popularidade do meio, e daqueles que fomentam constantemente o Big Data, com seu trabalho árduo de coleta, análise, e manipulação de dados.

5. A importância da Ciência de Dados.

A Ciência de Dados trabalha com a infinidade de dados que a sociedade produz atualmente. É por isso que a Ciência de Dados vem ganhando mais importância a cada ano. À medida que a quantidade e a complexidade dos dados aumentam, é preciso contar com profissionais que saibam processar esse grande volume de informações. Afinal, ter mais

dados não é necessariamente sinônimo de ter mais informação e conhecimento. Eles precisam ser processados para serem utilizados.

Então, o papel da Ciência de Dados nas organizações é o de transformar os dados brutos em informações com sentido e valor, por meio do seu processamento. Nas empresas, essas informações se tornam parte da inteligência de negócio — ou business intelligence, em Inglês — para auxiliar na tomada de decisões e criar melhores estratégias para o mercado.

Para esse processo, utiliza-se um conjunto de ferramentas, aliadas ao conhecimento humano. Então, a Ciência de Dados possibilita que isso aconteça de forma automática e inteligente. Com o processamento de milhões de dados, as máquinas já podem pensar e agir por conta própria, quase como humanos. Dessa maneira, o Data Science atua em favor dos objetivos das organizações, na velocidade que a era digital exige.

6. Relação Entre Ciência de Dados e Aprendizado de Máquinas.

Ciência de Dados e **Aprendizado de Máquinas** são duas áreas de conhecimento que se relacionam, entretanto apresentam características muito diferentes.

O Aprendizado de Máquinas, também conhecido como **Machine Learning**, refere-se à capacidade de uma máquina compreender informações, tomar decisões independentes e aprender constantemente, sendo também uma das áreas de IA, mais próxima de como os humanos pensam e agem. Porém, para que esses processos aconteçam, os dados precisam ser fornecidos para a inteligência da máquina, e é nesse momento que ocorre a relação entre Ciência de Dados e Aprendizado de Máquinas.

O profissional de Ciência de Dados desenvolve modelos e algoritmos que estruturam a lógica de pensamento da máquina, para que os dados sejam processados e ela realize ações a partir da sua compreensão. Devido a isso, o Aprendizado de Máquinas é uma das áreas tecnológicas que a Ciência de Dados se aplica e permite desenvolver diversas soluções.

7. Aplicações no Meio Tecnológico:

A Ciência de Dados é aplicada em diversas linguagens de programação, atualmente, as linguagens mais recomendadas para ela, são: Python, Linguagem R, Scala, SAS, Java, JavaScript, Matlab, Julia, C e Swift. Essa ciência também pode ser utilizada de inúmeras formas pelas empresas, em diferentes áreas como: produção, marketing, vendas, financeiro, RH e jurídico. Independente do ramo da empresa, a Ciência de Dados possui informações fundamentais para a gestão interna, otimização e direcionamento de estratégias, ela também contribui para a compreensão de tendências do cenário econômico, bem como o comportamento dos consumidores. Para que se tenha um melhor entendimento sobre o assunto, serão mostradas aplicações da Ciência de Dados em diferentes áreas no mercado.

7.1. Algoritmos de Pesquisa.

Alguns **mecanismos de pesquisa**, como o Google, gerenciam incontáveis dados a cada segundo, pois são responsáveis por organizar o conteúdo na internet e entender o que os usuários estão pesquisando para que assim possam apresentar os melhores resultados.

Para isso, os algoritmos dos usuários utilizam modelos de **Processamento de Linguagem Natural (Natural Language Processing)**. O NLP é um tipo de aplicação de Aprendizado de Máquina dedicada a compreender a linguagem humana e permitir uma comunicação eficiente entre máquinas e humanos. Por conta disso, a cada interação, os algoritmos aprimoram sua compreensão de linguagem, podendo assim entregar melhores resultados de pesquisa aos usuários.

7.2. Personalização do Marketing.

No marketing, a Ciência de Dados auxilia na criação de experiências personalizadas, utilizando ferramentas que coletam informações sobre os usuários, como o histórico de navegação, compras e interações com o produto, por exemplo, e assim identificam quais são os interesses e comportamentos de cada um. Dessa forma, é possível personalizar estratégias.

No **E-Commerce**, é comum utilizar vitrines com recomendações de produtos personalizadas, baseando-se em itens e categorias vistas pelo usuário. Já no **E-mail Marketing**, as mensagens possuem conteúdos personalizados com base nos interesses do consumidor.

7.3. Medicina.

Na Medicina, a Ciência de Dados traz importantes transformações. Nela os dados podem, por exemplo, monitorar sinais vitais de uma pessoa e alertar sobre a necessidade de procurar um médico ou se medicar.

Nos hospitais, a análise preditiva desses dados pode identificar a possibilidade de eventos futuros, como o risco de morte ou as chances de sucesso de um tratamento, por exemplo. Além disso, a aplicação desses dados no Aprendizado de Máquina possibilita o desenvolvimento de soluções automatizadas em equipamentos cirúrgicos, por exemplo, tornando assim os procedimentos mais eficientes.

7.4. Políticas Urbanas.

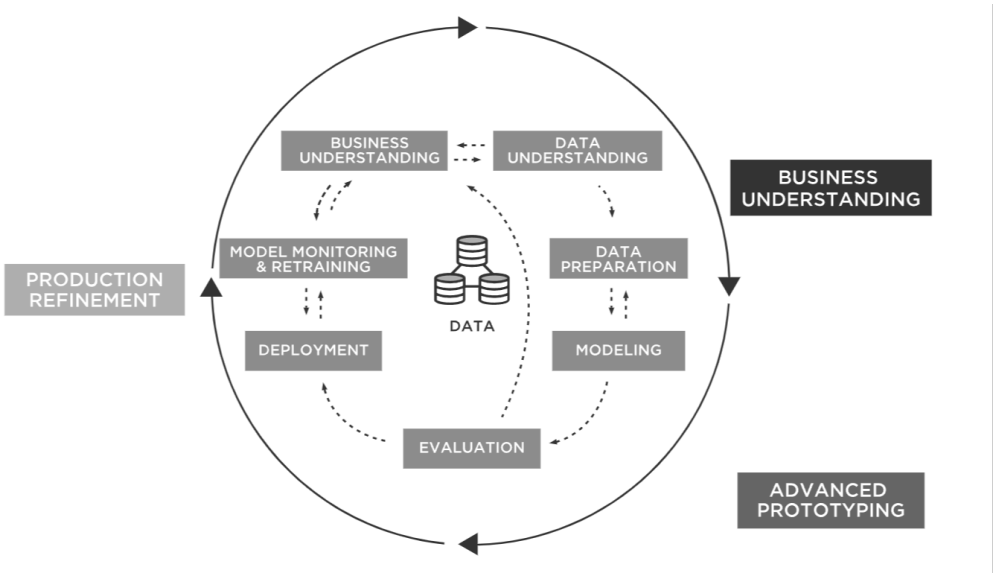
Por fim a Ciência de Dados pode também auxiliar na criação de cidades mais inteligentes. As **Smart Cities** se baseiam na inteligência dos dados para melhorar a qualidade de vida e a experiência dos turistas nas cidades.

Esses dados permitem identificar as melhores rotas de tráfego, assim como também áreas com altos índices de poluição e zonas com uma maior incidência de criminalidade, por exemplo. Deste modo, é possível utilizar a Ciência de Dados para criar políticas urbanas que resolvam as dificuldades das grandes cidades atualmente.

8. Sofware de Formação:

A Ciência de Dados é um subconjunto da **inteligência artificial (IA)** e se refere mais às áreas sobrepostas de estatísticas, métodos científicos e análise de dados, das quais são utilizadas para extrair significado e percepções dos dados. No entanto, para que ela possa ser colocada em prática e as análises aconteçam de forma bem-sucedida, é necessário que um profissional da área saiba capturar, armazenar e processar os dados. Logo após os processos de captura e armazenamento dos dados, inicia-se então a fase de preparação dos conteúdos, onde ocorre a validade e veracidade das informações.

O profissional de Ciência de Dados é responsável por verificar se essa informação é ou não verdadeira. No entanto, a validação e direcionamento adequado desse Big Data é realizado pelos **algoritmos da Ciência de Dados**, que são capazes de processar de forma ágil grandes quantidades de dados. Em geral, o processo de Ciência de Dados é algo complexo de ser compreendido e envolve uma série de etapas, sendo elas: Entender o problema do negócio; Reunir e integrar os dados brutos; Explorar e preparar os dados; Testar, ajustar e implantar os modelos; Monitorar, atualizar e controlar os modelos.



8.1. Entender o problema do negócio.

O processo de Ciência de Dados tem inicio com o entendimento do problema a ser resolvido. Por exemplo, um usuário empresarial que tenha interesse em saber como aumentar seu número de vendas, em muitos casos iria ao momento que buscasse por respostas, pesquisar questões muito amplas e ambíguas que não levam a uma hipótese imediatamente pesquisável. O papel da Ciência de Dados é dividir esses problemas em hipóteses pesquisáveis e testáveis, ou seja, transformar perguntas amplas em perguntas menores e objetivas. Em geral seu trabalho é entender qual decisão precisa ser tomada e trabalhar inversamente a partir dela.

8.2. Recolher e integrar os dados brutos.

Após a compreensão do problema do negócio, a próxima etapa envolve a coleta e integração dos dados brutos. Primeiramente será visto quais dados estão disponíveis. Em muitos casos, esses dados podem estar em diferentes formatos e sistemas, fazendo com que as técnicas de **data wrangling e data prepping** sejam utilizadas para converter os dados brutos em um formato adequado para as técnicas utilizadas. Caso os dados não estejam disponíveis, cientistas de dados, engenheiros de dados e TI buscam trazer novos dados a um ambiente de tipo sandbox para teste.

8.3. Explorar e preparar os dados.

Em seguida, o processo de exploração e preparação dos dados se inicia. Na maioria das vezes é empregada uma ferramenta para visualizar os dados organizando-os em gráficos

para auxiliar na compreensão dos padrões gerais e de quaisquer valores discrepantes em potencial. É neste momento que o analista começa a entender quais fatores ajudarão na resolução do problema e então começará a transformar, criar novas variáveis e preparar os dados para modelagem.

8.4. Testar, ajustar e implantar modelos.

Ao chegar nesta etapa, geralmente são utilizados algoritmos para criação de modelos a partir dos dados de entrada, usando diferentes técnicas para a testagem de modelos diversos. Os modelos e algoritmos estatísticos são aplicados ao conjunto de dados com o objetivo de generalizar o comportamento da variável destinada com base nos preditores de entrada.

As saídas geralmente são previsões que podem ser exibidas em relatórios incorporados ou infundidas diretamente nos sistemas de produção para tomar decisões próximas ao ponto de impacto. Portanto, logo após a implantação dos modelos nos sistemas, eles são utilizados para pontuar novos dados de entrada que nunca foram vistos antes.

8.5. Monitorar, atualizar e controlar os modelos.

Após a implantação, os modelos são monitorados para que possam ser atualizados e retreinados à medida que os dados são alterados devido à mudança de comportamento de eventos do mundo real. Por este motivo, a criação de uma estratégia de operações modelo é crucial para gerenciar as mudanças nos modelos de produção.

Em resumo, o processo de data science utiliza os dados e, em seguida, fornece informações preditivas que são usadas para aplicativos do mundo real. Este processo possui a seguinte ordem: Dados de entrada; Dados de preparação; Aplicação de aprendizado de máquina; Implante, pontuação e gerenciamento de modelos; Dados de saída.

9. O trabalho de um cientista de dados.

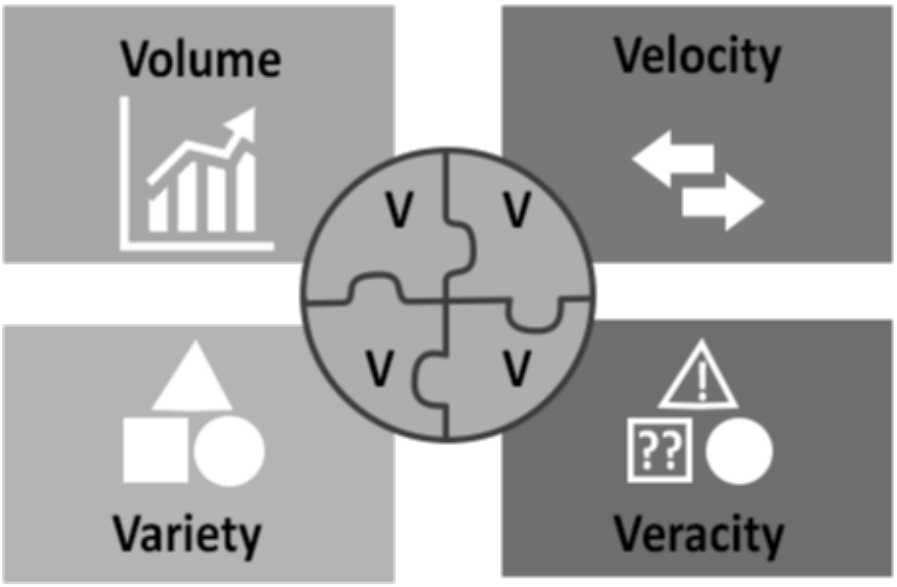
É inegável a importância da ciência de dados no decorrer do cotidiano contemporâneo. Nesse sentido, surge a necessidade de profissionais capazes de interpretar, manipular e redirecionar as informações coletades e armazenadas nas bases de dados. Estes acontecimentos, juntamente com a era Big Data, acarretaram na ascensão da profissão do cientista de dados.

10. Preocupações de um profissional de dados.

É seguro afirmar que, no escopo de un cientista de dados, é essencial a **interdisciplinalidade** a medida que, para manipular devidamente o grande volume de dados é exigido uma maior destreza que engloba compreensão de todo processo de coleta de dados, bem como a transformação dos mesmos com o objetivo final de suprimir as necessidades dos usuários.

De acordo com os estudos da pesquisadora Renata Curty, as capacidades exigidas dos profissionais variam entre conhecimentos matemáticos e também computacionais, explicitados na imagem abaixo:

Não somente o conhecimento destas matérias, como também a formação em uma das principais áreas, sendo elas Computação e Matemática, o que evidencia a forte



presença de noções lógico-matemáticas na capacitação de um profissional de dados, direcionados a resolução de problemas reais em inúmeros contextos.

Muito se discute hodiernamente sobre as funcionalidades da ciência de dados.Um conceito fundamental a ser abordado são **"os quatro v's"**, explicitados na Figura **??**.

Como supracitado, os quatro v's são como linhas guia para um profissional da ciência de dados e significam respectivamente: **volume, velocidade, variedade e veracidade**.

A partir deste conceito, a preocupação do cientista deixou de ser a localização de dados, afinal os mesmos já estão por todo lugar e seu crescimento é exponencial e até de certa forma abundante nos servidores *web*.

11. Conclusão.

Em síntese, o trabalho do cientista de dados compreende a manipulação e direcionamento do enorme volume de dados, com o foco final na qualidade de vida do usuário comum. Para isso, cabe ao profissional encarar as necessidades do mundo contemporâneo através de soluções lógicas e matemáticas. Dessa formas, é possível em favorecimento das necessidades do mundo digital.

12. Referências.

CAVALCANTE, Naje; "Saiba mais sobre a história do armazenamento de dados";http://www.fabricainfo.com/artigos/saiba-mais-sobre-historia-do-armazenamento-de-dados/;Acessado no dia 17 de maio de 2022.

LOPES, Daniel; "Data Science saiba como a ciência de dados pode te ajudar";https://navita.com.br/blog/data-science-saiba-como-a-ciencia-de-dados-pode-te-ajudar/;Acessado no dia 19 de maio de 2022.

GUTIERREZ, Renata; ”A ciência de dados e a cientista de dados”;https://revistas.ufpr.br/atoz/article/view/79882/43418; Acessado no dia 20 de maio de 2022

CURTY, R. G.;SERAFIM, J. S.,A formação em ciência de dados: uma análise preliminar do panorama estadunidense, Inf. Inf. Londrina, v. 21, n. 2 (2016).

CCM; ”Ciência de dados: o que é, como funciona e qual importância”; https://blog.ccmtecnologia.com.br/post/ciencia-de-dados-o-que-e-como-funciona-qual-importancia; Acessado no dia 11 de maio de 2022.

VICTÓRIA, Penélope; ”Linguagem Ciência de Dados: qual a melhor do mercado?”; https://blog.geekhunter.com.br/qual-a-melhor-linguagem-para-ciencia-de-dados/; Acessado no dia 11 de maio de 2022.

TIBCO; ”O que é Ciência de Dados?”; https://www.tibco.com/pt-br/reference-center/what-is-data-science; Acessado no dia 10 de maio de 2022.

INFNET, Instituto; ”O que é Data Science?”; https://blog.infnet.com.br/data-science/o-que-e-data-science-ciencia-de-dados/; Acessado no dia 10 de 2022.