Uso da ciência de dados como ferramenta para planejamento de atingimento das metas da Agenda 2030 para o desenvolvimento sustentável

Fabio Correa Xavier

Mestre em Ciência da Computação (IME-USP), MBA em Gestão Executiva de Negócios (IBMEC-RJ). Diretor Técnico do Departamento de Tecnologia da Informação, fabio@tce.sp.gov.br

Andrey Fernando da Silva Ribeiro

Pós-graduado em análise, projeto e gerência de sistemas (PUC-RJ), Bacharel em TI pela UERJ. Chefe Técnico da Fiscalização, andrey@tce.sp.gov.br

RESUMO

Este artigo apresenta uma aplicação da Ciência de Dados para obter e cruzar dados com o objetivo de subsidiar um planejamento adequado para se atingir as metas previstas na Agenda 2030 para o Desenvolvimento Sustentável. Como caso prático, mostraremos como um gestor municipal poder fazer uso de dados públicos de população e educação para conseguir atingir a meta 4.2, atendendo 100% da demanda por creches em seu município.

Palavras-chave: Ciência de dados. ODS. Agenda 2030. Desenvolvimento sustentável. Educação infantil. Creche.

INTRODUÇÃO

E se fosse possível prever a ocorrência de um surto de uma doença antes mesmo que este acontecesse? Ou prever a direção para a qual um vírus se espalhará? Ou ainda prever quando uma moléstia chegará à sua cidade?

Certamente estas informações seriam de grande valia nas mãos hábeis de gestores proativos e comprometidos com a saúde e o bem-estar da população local.

Mas o que pode parecer uma utopia ao olhar do leitor, na verdade, já é uma realidade. Basta uma rápida pesquisa na Internet para constatar que existem ferramentas capazes de realizar este feito. Mas como isso é possível?

Segundo Pereira (2015), pesquisadores desenvolveram uma ferramenta chamada ARGO (*AutoRegression with GOogle search data*), capaz de usar dados captados nas pesquisas do Google para realizar tais previsões. Ela é baseada num novo esquema de organização de informações que, unido a uma calibração contínua, gerou uma forma de acompanhar surtos de gripe.

Entretanto, se engana quem acha que esta é uma realidade distante de terras brasileiras. Graças a uma parceria entre uma empresa *startup* malaia e uma ONG carioca, foi possível criar uma tecnologia que usa a inteligência artificial para prever surtos de dengue em determinadas regiões, facilitando a sua prevenção e combate (PRADO, 2015).

Estas duas ferramentas possuem algo em comum além do seu objetivo principal, que é o de prevenir e combater doenças: o uso da Ciência de Dados (ou *Data Science*, em inglês). Trata-se de uma área interdisciplinar voltada para o estudo e a análise de dados, estruturados ou não, que visa à extração de conhecimento. Comumente utiliza aprendizagem de

máquina[1], *Big Data*[2], além de técnicas de outras áreas interdisciplinares como estatística, economia, engenharia e outros campos da computação.

Sendo assim, os dois exemplos acima ilustram de maneira definitiva que se trata de ferramentas valiosas que podem e devem ser utilizados pela Administração Pública no desempenho das suas funções.

A CIÊNCIA DE DADOS

A Ciência de Dados é um campo emergente de pesquisa que se concentra em extrair informações relevantes de um grande e complexo conjunto de dados (DANIEL, 2018). A Ciência de Dados pode envolver o uso de aprendizado de máquina e outras técnicas automatizadas para coleta, processamento, consolidação e visualização de dados, muitas vezes heterogêneos e complexos. Como um campo de investigação multidisciplinar, a Ciência de Dados acaba por envolver conhecimentos das áreas de estatística, pesquisa operacional, matemática, humanidades e ciências sociais, além da ciência da computação (DANIEL, 2018). Por essas características, o uso dessa técnica pode se ser um grande desafio, se levarmos em conta a variedade de técnicas de descoberta de conhecimento e o colossal volume de dados existentes e à disposição.

Para vencer desafios como a exploração de grande quantidade de dados, a procura por padrões consistentes, regras de associação, sequências temporais e detecção de relacionamentos entre variáveis, existem várias técnicas, processos e modelos de extração do conhecimento. Para ilustrar a diversidade de técnicas e áreas de conhecimento, descrevemos algumas delas, sucintamente, a seguir:

I - Mineração de Dados (*Data Mining*): que, segundo Tan (2006), consiste na aplicação de algoritmos sobre bases de dados a fim de se extrair conhecimento útil não trivial dessas. Ainda de acordo com Larose (2005), pode-se definir mineração de dados como sendo *um processo de descoberta de novas correlações, padrões e tendências significativas através do exame de grandes quantidades de dados armazenados em repositórios através do uso de tecnologias de reconhecimento de padrões, de estatística e de matemática.*

II – Mineração de Textos (*Text Mining*): um processo computacional no qual se busca analisar um ou mais documentos na busca de informações relevantes, como, por exemplo, conceitos recorrentes, relacionamentos entre conceitos, dentre outros (FELDMAN; SANGER, 2006).

III – Clusterização: que visam à formação de grupos de observações homogêneos dentro de um mesmo grupo e significativamente distintas das observações inseridas em outros grupos (VICTOR LEONARDO CERVO, 2015)

IV - Lógica Fuzzy: técnica desenvolvida pela necessidade de criação de um método capaz de expressar de uma maneira sistemática quantidades imprecisas, vagas e mal definidas (MORE et al., 2010).

V - Web Crawler: um script ou programa que pode navegar em páginas web de maneira automática, para se atingir alguma finalidade específica (GARG; GUPTA; SINGH, 2017).

Neste momento, podem surgir vários pensamentos desanimadores na mente do leitor, como "isso é muito complicado", "vou ter que remar muito pra chegar nesse nível", ou ainda "não temos recursos para realizar um trabalho de análise deste porte". Por desmistificar essa imagem, demonstraremos na próxima seção uma aplicação simples da Ciência de Dados, tendo como premissa o atingimento de uma das metas dos Objetivos de Desenvolvimento Sustentável - ODS da ONU.

CASO PRÁTICO

A adoção de modelos e técnicas complexas e avançadas de análise de dados pode ser um fator desanimador para a maioria dos profissionais que iniciam esta jornada. Portanto, acreditamos que, no início, simplicidade é essencial. A utilização de modelos simples, porém corretos e precisos, que reflitam a realidade da sua localidade, é o melhor caminho a ser tomado. Uma análise simples e

1. Aprendizagem de máquina é considerada um subcampo da Inteligência Artificial, que trabalha com a ideia de que as máquinas podem aprender sozinhas ao terem acesso a grandes volumes de dados. Esse aprendizado se dá pela detecção de padrões e criação de conexões entre dados, por meio de *Big Data e algoritmos sofisticados.*

2. Grandes conjuntos de dados, caracterizados pelos 3 Vs: Velocidade, Volume e Variedade.

correta certamente alcançará resultados mais efetivos e rápidos.

Do mesmo modo, a utilização de uma abordagem iterativa e incremental, que aumente paulatinamente a complexidade e fidedignidade do modelo, é uma excelente prática. Para isso, a reutilização do conhecimento adquirido nas iterações anteriores é um passo fundamental. Assim, será possível acompanhar a evolução do trabalho, além de facilitar a reanálise periódica dos dados, parte importante que visa entender as mudanças ocorridas ao longo do tempo.

Para exemplificar essa abordagem simples, iterativa e incremental, vamos utilizar uma meta de um dos objetivos de destaque no âmbito dos ODSs, no ano de 2018, no TCESP: a educação inclusiva, equitativa e de qualidade. Para tal, escolhemos uma de suas metas para nortear o exemplo proposto neste artigo, a meta 4.2, enunciada a seguir:

> *Meta 4.2 – Até 2030, garantir que todas as meninas e meninos tenham acesso a um desenvolvimento de qualidade na primeira infância, cuidados e educação pré-escolar, de modo que eles estejam prontos para o ensino primário (NAÇÕES UNIDAS NO BRASIL, 2015).*

Os cuidados pré-escolares certamente passam pelo oferecimento de vagas de creches à população, oferecendo cuidados que possam contribuir para um desenvolvimento de qualidade da primeira infância. O processo de desenvolvimento infantil é influenciado pela interação de diversos ambientes, incluindo berçários, creche e pré-escola, nos quais a criança tem a oportunidade de interagir com os pares e outros adultos, aprendendo novas habilidades cognitivas e socioemocionais (MACANA; COMIM; TAI, 2014).

Entretanto, sabemos que, até 2030, a quantidade de vagas em creches necessárias ao atendimento da população pode (e deve) variar. Sendo assim, resta-nos responder, mesmo que de maneira aproximada, a seguinte pergunta:

Quantas vagas de creche serão necessárias em 2030 para que o meu Município atenda a toda a demanda da população nesta data?

A primeira etapa deste processo consiste em extrair e transformar dados úteis da imensidão de informações públicas disponíveis em várias mídias, especialmente na Internet. Para este caso, utilizamos dados do Índice de Efetividade da Gestão Municipal - IEG-M (Exercício 2017), extraindo os dados referentes à demanda por creche, do município[3] em análise. Esses dados são mostrados na Tabela 1.

DEMANDA POR VAGAS DE CRECHE DO MUNICÍPIO	
Vagas de creche oferecidas	5.636
Demanda municipal por vagas em creche	7.944
Demanda não atendida	2.308 (40,95% das vagas existentes)

Tabela 1 - Demanda por vagas de creche do Município em análise.

Fonte: IEG-M TCESP

Do Portal de Estatísticas do Estado de São Paulo da Fundação SEADE, extraímos os dados de projeção de crescimento populacional do mesmo Município, exibidos na Tabela2.

PROJEÇÃO DO CRESCIMENTO POPULACIONAL DO MUNICÍPIO	
2014	273.854
2015	276.852
2016	279.626
2017	282.428
2018	285.257

Tabela 2 - Dados de projeção do crescimento populacional do Município em análise

Fonte: Fundação Seade

Os dados da Tabela 2 permitem o cálculo da taxa de crescimento populacional anual, por meio da seguinte fórmula:

$$\text{onde: } \left[\left(\frac{f}{s}\right)^{1/y} - 1\right] \times 100$$

3. Omitimos o nome do Município de forma proposital.

f → População final (ano de 2018)

s → População final (ano de 2014)

y → Número de anos em análise
(no nosso caso, 2018-2014=4)

Assim, temos:

$$\left[\left(\frac{285.257}{273.854}\right)^{\frac{1}{2018-2014}} - 1\right] \times 100 = 1{,}03\%\ a.a.$$

Figura 1 – Fórmula para cálculo de taxa de crescimento anual

Como demonstrado anteriomente, vimos que a taxa de crescimento anual da população é de **1,03%** ao ano. Em um primeiro modelo, poderíamos aplicar a mesma taxa na demanda de vagas por creches, o que nos daria, no ano de 2030, uma necessidade de 9.070 vagas em creches.

Entretanto, sabemos que o número de vagas existentes hoje é menor que a demanda atual. Sendo assim, para atingir 100% de atendimento à demanda de vagas, a taxa de crescimento da oferta de vagas deve ser maior que a taxa de crescimento populacional.

Aplicando a mesma fórmula da Figura 1, calculamos que a taxa de crescimento necessária para igualar a quantidade de vagas à demanda no ano de 2030, é de **3,73%** ao ano.

Para ilustrar graficamente os cálculos exibidos até aqui, plotamos no Gráfico 1 as duas taxas de crescimento apuradas.

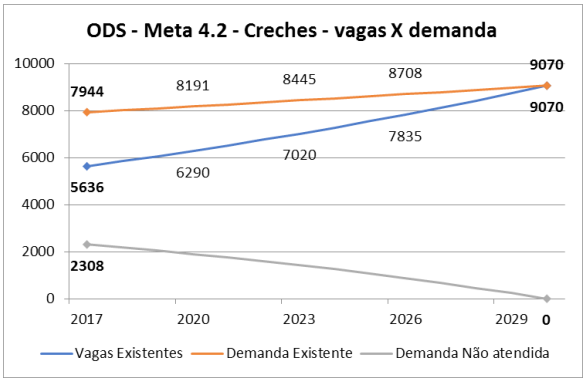

Gráfico 1 - Vagas x Demanda - Iteração 1

APRIMORANDO O MODELO COM DADOS MAIS PRECISOS

Uma das características mais importantes de um cientista de dados é a inquietude frente aos resultados alcançados, é ter aquela sensação de que alguma coisa sempre pode ser melhorada.

O que nos leva à seguinte pergunta: estariam disponíveis dados mais precisos para o cálculo da taxa de crescimento populacional?

Garimpando o banco de dados do Portal de Estatística do Estado de São Paulo da Fundação Seade, constatamos que há dados mais precisos. A referida entidade também projeta a população por faixa etária. Sendo assim, podemos utilizar as projeções populacionais da faixa de 0 a 4 anos, objeto do nosso estudo, como mostrado na Tabela 3.

PROJEÇÃO DE CRESCIMENTO POPULACIONAL	
PERÍODOS	POPULAÇÃO DE 0 A 4 ANOS
2014	**20.358**
2015	20.750
2016	20.783
2017	20.807
2018	**20.819**

Tabela 3 - Projeção de crescimento populacional para a faixa etária de 0 a 4 anos.

Fonte: Fundação Seade

Aplicando a fórmula da Figura 1, temos:

$$\left[\left(\frac{20.819}{20.358}\right)^{\frac{1}{2018-2014}} - 1\right] \times 100 = 0{,}56\%\ a.a.$$

Percebemos no cálculo acima que a taxa de crescimento da população de 0 a 4 anos é muito menor que a da população como um todo. Sendo assim, a quantidade de vagas necessárias até 2030 cai de 9.070 para 8.544, uma redução de 526 vagas (5,8%), que seriam criadas sem necessidade. Ilustramos, novamente, essas projeções de crescimento no Gráfico 2.

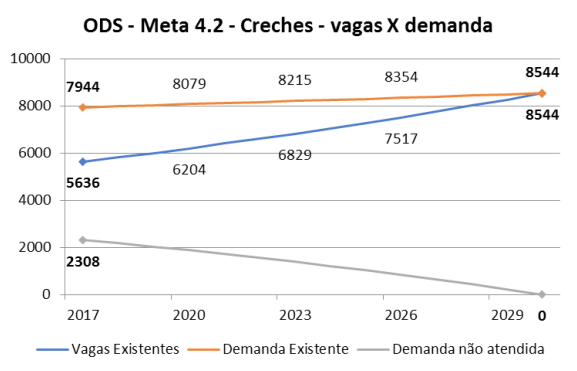

Gráfico 2 - Vagas x Demanda - Iteração 2

Poderíamos, ainda, fazer novas iterações, refinamentos e melhorias no modelo, como estratificar por regiões da cidade ou por renda, aumentando a precisão e complexidade da projeção.

CONCLUSÃO

Este artigo procurou apresentar de maneira introdutória o que é a Ciência de Dados e como utilizá-la para se planejar o atingimento de uma das metas propostas na Agenda 2030 dos Objetivos de Desenvolvimento Sustentável – ODS. Vimos que a abordagem iterativa e incremental, aumentando paulatinamente a complexidade, quantidade de dados e precisão do modelo é um caminho menos doloroso para se iniciar nessa técnica. Por fim, separamos algumas dicas que podem ser muito valiosas para iniciantes neste assunto:

Dica #1: Revisitar seus dados e suas análises periodicamente.

Deve-se buscar erros, incoerências, dados incorretos, imprecisos ou inadequados à sua necessidade. É uma boa prática verificar se o seu modelo espelha a realidade, certificando-se de que os resultados projetados são compatíveis com os alcançados.

Dica #2: Aprenda com os dados dos outros.

Se alguém já gastou tempo e dinheiro para desenvolver um estudo que se aplica à sua realidade, não gaste tudo de novo. Use-o!

Dica #3: Nenhum dado estatístico substitui o trabalho de campo.

Os dois trabalhos são complementares e importantes. Testar na prática os modelos, colhendo os dados da sua localidade em campo, é uma boa prática. Com isso, a precisão do modelo será sempre melhor.

REFERÊNCIAS BIBLIOGRÁFICAS

MACANA, Esmeralda Correa; COMIM, Flávio; TAI, Silvio Hong Tiing. **Impactos da Creche na Primeira Infância: efeitos dependendo das características da família e do grau de exposição ao centro de cuidado**. Porto Alegre, 2014. Disponível em: <http://repositorio.pucrs.br/dspace/bitstream/10923/10687/2/Impactos_da_creche_na_primeira_infancia_efeitos_dependendo_das_caracteristicas_da_familia_e_o_grau_de_exposicao_ao.pdf>. Acesso em: 22 nov. 2018.

DANIEL, Ben. **Reimaging Research Methodology as Data Science**. Big Data And Cognitive Computing, [s.l.], v. 2, n. 1, p.4-20, 12 fev. 2018. MDPI AG. http://dx.doi.org/10.3390/bdcc2010004.

FELDMAN, R.; SANGER, J. **Text Mining Handbook**. Cambridge (MA): Cambridge University Press, 2006.

GARG, Abhinav; GUPTA, Kratika; SINGH, Abhijeet. **Survey of Web Crawler Algorithms**. International Journal Of Advanced Research In Computer Science, v. 5, n. 8, maio 2017.

LAROSE, D. "**Discovering knowledge in data: an introduction to data mining**", John Wiley and Sons, Inc., 2005, 723p.

MORE, Jesus et al. **Evaluation of the Globo.Com Portal Efficiency: A Case Study In Light Of The Fuzzy Set Theory**. Jistem Journal Of Information Systems And Technology Management, [s.l.], v. 7, n. 2, p.353-374, 30 ago. 2010. TECSI. http://dx.doi.org/10.4301/s1807-17752010000200006. Disponível em: <http://www.jistem.fea.usp.br/index.php/jistem/article/view/10.4301%252FS1807-17752010000200006/210>. Acesso em: 22 nov. 2018.

NAÇÕES UNIDAS NO BRASIL. **Transformando Nosso Mundo: A Agenda 2030 para o Desenvolvimento Sustentável**. 2015. Disponível em: <https://nacoesunidas.org/pos2015/ods4/>. Acesso em: 20 nov. 2018.

PEREIRA, Leonardo. **Ferramenta usa dados do Google para prever surtos de gripe**. 2015. Disponível em: <https://olhardigital.com.br/noticia/ferramenta-usa-dados-do-google-para-prever-surtos-de-gripe/52855>. Acesso em: 19 nov. 2018.

PRADO, Jean. **Startup usa inteligência artificial para prever surtos de dengue no Brasil**. 2015. Disponível em: <https://tecnoblog.net/189257/inteligencia-artificial-surtos-dengue-brasil/>. Acesso em: 20 nov. 2018.

SÃO PAULO. Fundação Seade. Secretaria de Planejamento e Gestão do Estado de São Paulo. **Perfil dos Municípios Paulistas**. 2018. Disponível em: <http://www.perfil.seade.gov.br/>. Acesso em: 20 nov. 2018.

SÃO PAULO. TRIBUNAL DE CONTAS DO ESTADO DE SÃO PAULO. . **Índice de Efetividade da Gestão Municipal**. 2017. Disponível em: <https://iegm.tce.sp.gov.br/>. Acesso em: 20 nov. 2018.

TAN, P.; STEINBACH, M.; KUMAR, V. "**Introduction to Data Mining**", Addison-Wesley Longman Publishing Co., Inc., 2006, 476p.

VICTOR LEONARDO CERVO. **Seleção de variáveis para clusterização de bateladas produtivas através de ACP e remapeamento kernel**. Production, [s.l.], v. 25, n. 4, p.826-833, 18 ago. 2015. FapUNIFESP (SciELO). http://dx.doi.org/10.1590/0103-6513.143613.

YANG, Shihao; SANTILLANA, Mauricio; KOU, S. C.. **Accurate estimation of influenza epidemics using Google search data via ARGO**. 2015. Disponível em: <http://www.pnas.org/content/early/2015/11/04/1515373112>. Acesso em: 20 nov. 2018.