

O QUE É DATA SCIENCE: ESTUDO DE CASO E COMO APLICÁ-LA

Francinildo Melo Rodrigues¹

Patrick Reinecke de Alverga²

RESUMO

Atualmente, nesta época em que uma grande gama de dados é produzida todos os dias, foi necessária uma nova abordagem para interpretá-los. O Data Science é um campo interdisciplinar, uma intersecção entre ciência da computação, engenharias, matemática, estatística com áreas de negócio, para investigação de dados que tratam de problemas reais de negócios. Vários benefícios podem ser alcançados dependendo do negócio analisado. Personalização do atendimento, aumento do retorno sobre investimentos, conhecimento do mercado alvo, agilidade nos processos de tomada de decisão, melhoria na prestação de serviços são alguns dos benefícios que o Data Science pode trazer para sua empresa. Neste Artigo será apresentado algumas das tecnologias mais utilizadas na área, bem como um estudo de caso envolvendo o Enem 2017.

Palavras-chave: Data Science, Análise de dados, Aprendizado de máquina, aprendizado não supervisionado.

¹ Acadêmico do Curso de Bacharelado em Sistemas de Informação do Centro Universitário do Rio Grande do Norte (UNI-RN). E-mail: francis584@gmail.com

² Professor Orientador do Curso de Bacharelado em Sistemas de Informação do Centro Universitário do Rio Grande do Norte (UNI-RN) E-mail: reinecke.patrick@gmail.com

WHAT IS DATA SCIENCE: CASE STUDY AND HOW IT IS APPLICABLE

ABSTRACT

Nowadays, in this time in which a wide range of data is produced every day, a new approach was required to interpret them. Data Science is an interdisciplinary field, an intersection between Computer Science, Engineering, Mathematics and Statistics, for data research that addresses real problems in the Business Area. Several benefits can be achieved depending on the business analyzed. Personalization of service, increased return on investments, knowledge of the target market, agility in decision-making processes and improvement in service delivery are some of the benefits that Data Science can bring to a company. This article will present some of the most used technologies in this area, as well as a case study involving Enem 2017.

Keywords: Data Science, Data Analysis, Machine Learning, Unsupervised Learning.

1 INTRODUÇÃO

O crescimento da quantidade de dados, como a maioria dos fenômenos relacionados à TI, também está em conformidade com a Lei de Moore. Semelhante ao número de transistores em chips, a capacidade dos discos rígidos e o poder computacional dos computadores, o universo digital está crescendo exponencialmente e aproximadamente dobrando a cada dois anos. Embora isso não seja um fenômeno novo, de repente muitas organizações perceberam que quantidades crescentes de dados, precisam ser usadas de maneira inteligente para competir com outras organizações em termos de eficiência, velocidade e serviço. No entanto, o objetivo não é coletar o máximo de dados possível. O verdadeiro desafio é transformar dados em informações valiosas. Somente técnicas de Data Science podem relacionar diretamente dados a processos de negócios.

Neste Artigo além de entendermos o que é Data Science, teremos um Estudo de Caso demonstrando o uso destas técnicas nos dados do Exame Nacional do Ensino Médio de 2017.

2 REFERENCIAL TEÓRICO

Nesta sessão será apresentado o referencial teórico necessário para um melhor entendimento do artigo, sempre apresentando as ideias centrais dos temas abordados e citando os autores e fontes de onde foram extraídas tais informações.

2.1 DATA SCIENCE

A Ciência de Dados (Data Science) é a reunião de princípios essenciais que suportam e guiam a extração de informações e conhecimento a partir dos dados. Neste significado, a Ciência de Dados envolve preceitos, processos e técnicas para compreender fenômenos através da análise automatizada de dados (PROVOST e FAWCETT, 2013).

Os termos atuais mais comuns usados para definir Data Science têm sido utilizados ligados a outros campos do conhecimento, como Análise de Dados, Processamento de Dados, Estatística, Descoberta de Conhecimento em Banco de Dados (KDD), Data Mining, Big Data, entre outros. Logo, é possível perceber se tratar de um campo interdisciplinar, que divide significados e ciências de atuação com outras áreas. Uma ilustração da interdisciplinaridade envolvida na Ciência de Dados pode ser vista no “Diagrama Venn de Ciência de Dados” (Conway, 2010).

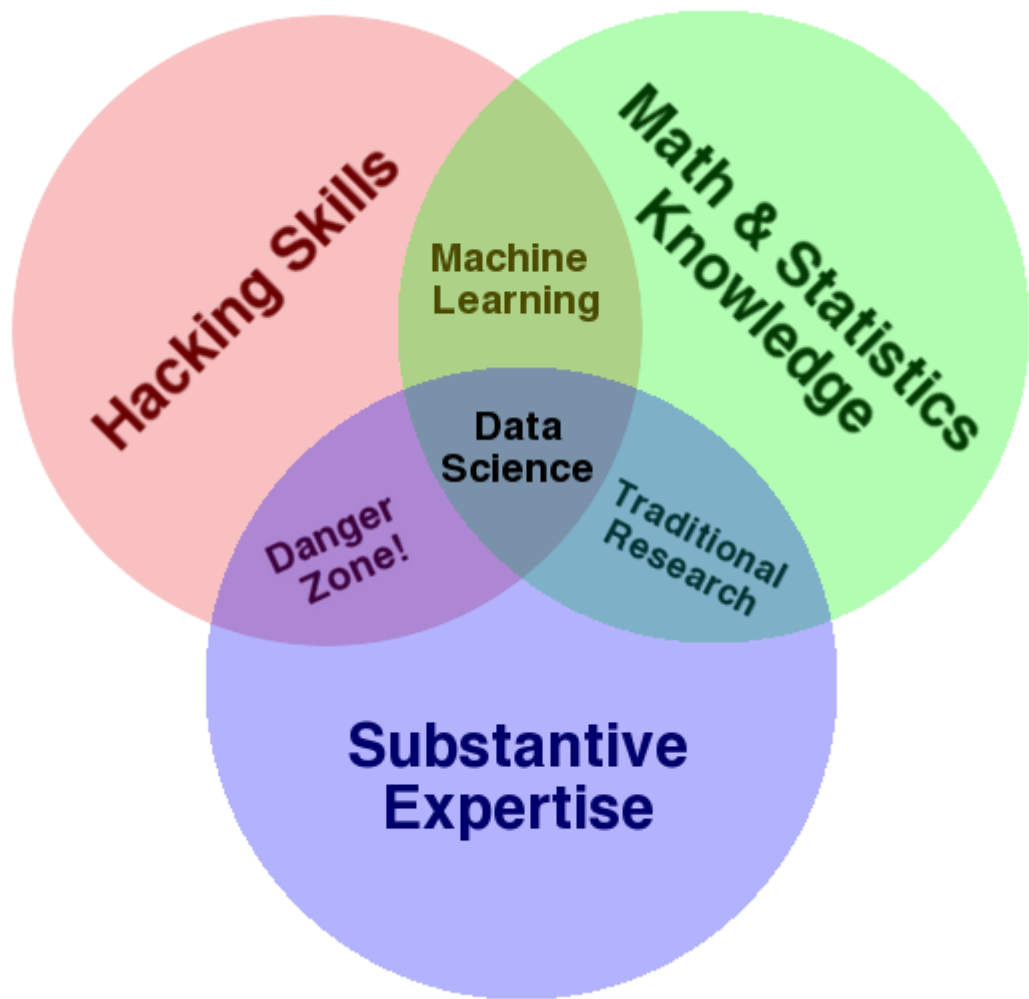


Figura 1: Diagrama Venn de Ciência de Dados (Fonte: CONWAY, 2010)

A aplicação em conjunto das habilidades descritas no Diagrama Venn – Ciência da Computação (Hacking Skills), conhecimentos estatísticos e matemática, e expertise no domínio/negócio – definem, na prática, a Ciência de Dados. (CONWAY, 2010)

2.2 ESTUDO DE CASO

De acordo com Jennifer Platt (1992), o método de estudo de caso não é apenas trabalho de campo, mas lógica de planejamento. É uma investigação empírica que examina extensivamente um fenômeno contemporâneo que está relacionado ao contexto da vida real.

A pesquisa do estudo de caso inclui, portanto, uma metodologia abrangente com projeto de projeto lógico, coleta de dados e abordagens específicas de análise de dados. Isto é, não se limita a uma tática isolada de coleta de dados. (YIN, 2010, p.40)

De acordo com YIN (2010, p.41) existem pelo menos quatro aplicações diferentes nas quais estudos de caso podem ser usados, a saber: Explicar relações causais em intervenções da vida real, descrever o contexto real, ilustrar tópicos específicos em uma avaliação e finalmente examinar as situações em que a intervenção não fornece resultados claros.

Como método de pesquisa o estudo de caso é utilizado em situações em que são analisados fenômenos individuais, grupais, organizacionais, sociais, políticos e relacionados. (YIN, 2010, p.25)

Segundo o autor YIN (2010, p.32) o estudo de caso é o método preferido no exame dos eventos contemporâneos, mas quando os comportamentos relevantes não podem ser manipulados. O estudo de caso se assemelha com o método histórico, mas adiciona duas fontes de evidência, a observação direta dos eventos estudados e entrevistas com pessoas envolvidas nas pesquisas.

2.3 PYTHON

Python é uma linguagem de programação de alto nível, interpretada, de script, imperativa, orientada a objetos, funcional, de tipagem dinâmica e forte. Foi lançada em 1991. Atualmente possui um modelo de desenvolvimento comunitário, aberto e gerenciado pela organização sem fins lucrativos Python Software Foundation. (MCKINNEY, 2013)

2.4 NUMPY

NumPy (abreviação de Numerical Python) fornece uma interface eficiente para armazenar e operar em buffers de dados densos. De certa forma, os arrays NumPy são como uma implementação do tipo lista do Python, mas fornecem operações de armazenamento de dados muito mais eficientes à medida que os arrays aumentam de tamanho. Os nd-arrays formam o núcleo de quase todo o ecossistema de ferramentas de ciência de dados no Python. (VANDERPLAS, 2017).

Além da capacidade de processamento rápido dos arrays que o NumPy adiciona ao Python, um de seus principais objetivos em relação à análise de dados é como o recipiente principal para os dados a serem transmitidos entre os algoritmos. Para dados numéricos, os arrays NumPy são uma maneira muito mais eficiente de armazenar e manipular dados do que as outras estruturas de dados do Python. (MCKINNEY, 2013)

2.5 PANDAS

O Pandas é um pacote mais novo, construído sobre o NumPy, e fornece uma implementação eficiente de um DataFrame. Os DataFrames são essencialmente matrizes multidimensionais com rótulos de linha e coluna anexados e geralmente com tipos heterogêneos e/ou dados ausentes. Além de oferecer uma interface de armazenamento conveniente para dados rotulados, o ele implementa várias operações de dados poderosas, familiares aos usuários de estruturas de bancos de dados e programas de planilhas. (VANDERPLAS, 2017) Ele fornece a funcionalidade de indexação sofisticada para facilitar a reformulação, fatiamento de dados, execução de agregações e seleção de subconjuntos de dados. (MCKINNEY, 2013)

2.6 DASK

O Dask é uma especificação para codificar algoritmos paralelos, usando dicionários, tuplas e callables primitivos do Python. Usamos o dask para criar o `dask.array`, uma biblioteca de matriz N-dimensional paralela que copia a interface

NumPy, usa todos os núcleos de um processador moderno e gerencia bem os dados no disco. Dask.array serve tanto como uma biblioteca geral para ndarrays paralelos fora do núcleo e também como uma demonstração que podemos paralelizar bases de código complexas como o NumPy de maneira direta usando algoritmos bloqueados e agendamento de tarefas. (ROCKLIN, 2015)

2.7 MATPLOTLIB

O matplotlib é a biblioteca Python mais popular para produzir gráficos e outras visualizações de dados 2D. Foi originalmente criado por John D. Hunter (JDH) e agora é mantido por uma grande equipe de desenvolvedores. É bem adequado para criar gráficos apropriados para publicação (MCKINNEY, 2013). Um dos recursos mais importantes dele é sua capacidade de funcionar bem com varios sistemas operacionais e backends gráficos. O Matplotlib suporta dezenas de backends e tipos de saída, o que significa que pode-se contar com ele independentemente do sistema operacional que estiver usando ou do formato de saída desejado. Essa abordagem multiplataforma, tudo para todos, tem sido uma das grandes forças do Matplotlib. (VANDERPLAS, 2017)

2.8 SEABORN

O Matplotlib antecedeu o Pandas em mais de uma década e, portanto, não foi projetado para uso com o Pandas DataFrame. Para visualizar os dados de um Pandas DataFrame, você deve extrair cada série e, muitas vezes, concatena-los para o formato certo. Seria melhor ter uma biblioteca de plotagem que pudesse usar inteligentemente os rótulos do DataFrame em um gráfico. Uma resposta para esses problemas é o Seaborn. Ele fornece uma API em cima do Matplotlib que oferece escolhas sensatas para estilos de plotagem e cores padrões, define funções simples de alto nível para tipos comuns de enredo estatísticos, e se integra com a funcionalidade fornecida pelo Pandas DataFrame. (VANDERPLAS, 2017)

2.9 SCIKIT-LEARN

Existem várias bibliotecas Python que fornecem implementações sólidas de uma variedade de algoritmos de aprendizado de máquina. Uma das mais conhecidas é o Scikit-Learn, um pacote que fornece versões eficientes de um grande número de algoritmos comuns. Ele é caracterizado por uma API limpa, uniforme e simplificada, bem como por uma documentação on-line muito útil e completa. Um benefício dessa uniformidade é que, uma vez que você entende o uso básico e a sintaxe do Scikit-Learn para um tipo de modelo, mudar para um novo modelo ou algoritmo é muito simples. (VANDERPLAS, 2017)

2.10 ALGORITMO K-MEANS

Este método de análise é um método de agrupamento não-hierárquico por repartição. Consiste em um procedimento onde, dado um número de clusters previamente determinado (o K do nome K-means), calcula-se pontos que representam os “centros” destes clusters e que são espalhados homogeneamente no conjunto de dados obtidos e movidos, heurísticamente, até alcançar um equilíbrio estático (HAIR et al., 2005). Procede-se, em seguida, a uma divisão de todos os casos obtidos pelos ‘k’ grupos preestabelecidos e a melhor partição dos ‘n’ casos será aquela que otimize o critério escolhido. Em outras palavras, o procedimento inicia-se usando os valores dos primeiros k casos como estimativas temporárias das médias dos clusters.

Os centroides iniciais são formados através da designação de cada caso ao cluster de centro mais próximo. Com a inclusão de cada caso, a média altera-se, alterando assim o centróide. Um processo iterativo é usado para achar os centróides finais de cada cluster. A cada passo, os casos são agrupados ao cluster de centro mais próximo, e novamente as médias são recalculadas. Este processo continua até que não haja mais alterações nas médias ou que um número pré-determinado de iterações aconteça, encerrando-se o processo (HAIR et al., 2005).

O centróide pode ser a média ou a mediana de um grupo de pontos. Em outras palavras, o objetivo do K-means é minimizar a distância entre cada ponto e o seu respectivo centróide (HAIR et al., 2005).

Ao aplicar um critério de otimização que separe uma amostra em 'k' grupos homogêneos, pretende-se que, dentro de cada grupo, os elementos sejam os mais semelhantes possíveis entre si, ao passo que as semelhanças entre os elementos de grupos distintos sejam as menores possíveis (LOURENÇO & MATIAS, 2000).

A escolha de 'k' pode ser motivada por fatores externos, porém, em geral, esse não é o caso. Uma maneira razoavelmente fácil de encontrar esse valor envolve a plotagem da soma de erros quadrados (entre cada ponto e a média de seu cluster) como uma função de k e a observação de onde o gráfico irá "dobrar". (GRUS, 2015)

3 POPULARIDADE

O termo "ciência de dados" tem surgido em vários contextos nos últimos trinta anos. Em 1974, Naur publicou o Concise Survey of Computer Methods (NAUR, 1974), que utilizou livremente o termo data science em sua pesquisa, porém ganhou mais destaque nos últimos anos devido a alguns fatores como: Diminuição do custo por Gigabyte e crescimento exponencial do poder computacional, conforme é possível verificar nos gráficos abaixo.

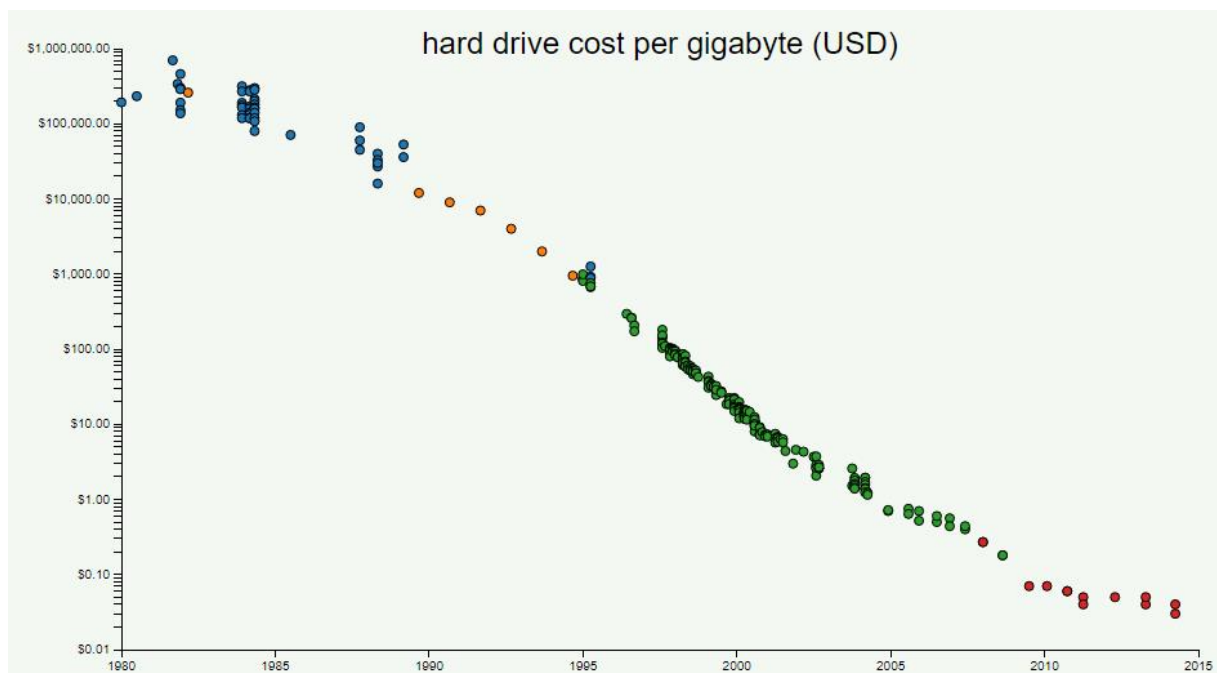


Figura 2: Custo do Gigabyte ao longo dos anos (Fonte: MKOMO, 2014)

No final de 2010 já era praticante irrisório o custo do espaço, menos de 10 centavos por Gigabyte.

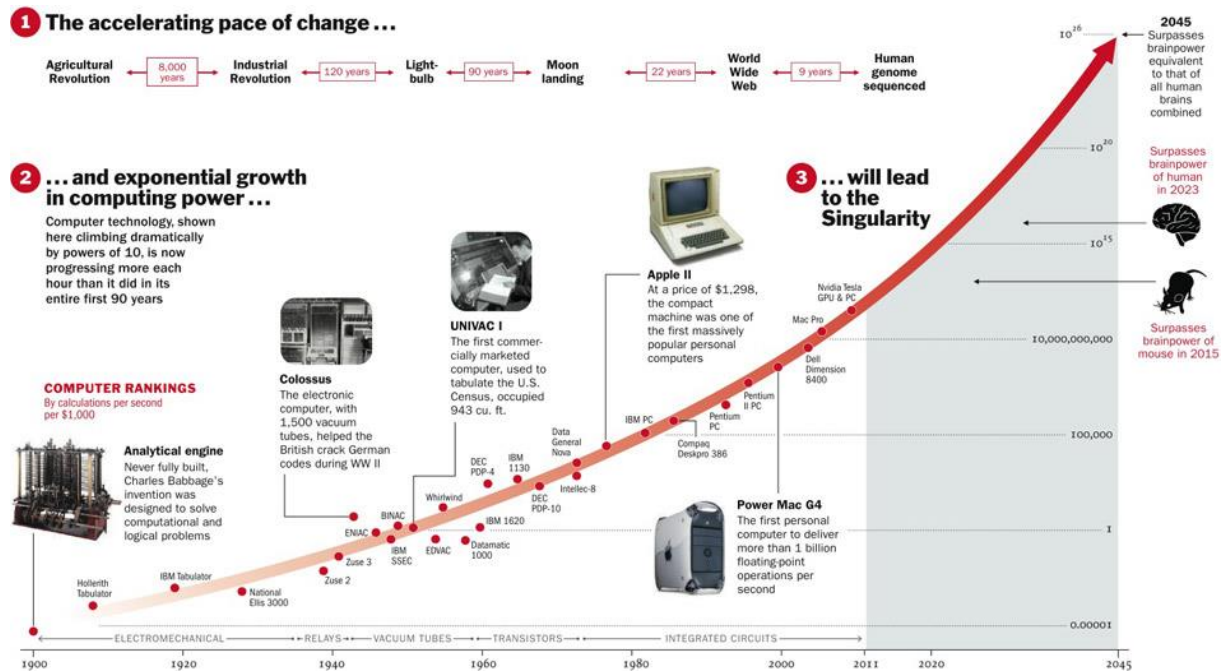


Figura 3: Crescimento do poder computacional (Fonte: TIME, 2012)

Em 2012 já tínhamos processadores com a capacidade de um cérebro de um rato funcionando em nossas casas. Estimasse que em 2045 passará de todos os cérebros humanos juntos.

Hoje coletar, analisar e comunicar informações é mais fácil, barato e rápido que no passado.

4 ESTUDO DE CASO

Foi utilizado como objeto de estudo o microdado do Enem 2017, disponibilizado pelo Instituto Nacional de Estudos e Pesquisas. Os microdados se constituem no menor nível de desagregação de dados recolhidos por pesquisas, avaliações e exames realizados pelo INEP.

Em um primeiro momento foi realizado um estudo sobre como os dados contidos no microdado eram organizados. Percebeu-se que eles estavam divididos entre dados do participante, dados da escola, dados dos pedidos de atendimento especializado, dados dos pedidos de atendimento específico, dados dos pedidos de recursos específicos para realização das provas, dados do local de aplicação da prova, dados da prova objetiva, dados da redação e dados do questionário socioeconômico. Decidido que trabalharíamos somente com os dados do Rio Grande do Norte, prosseguimos para o próximo passo.

Iniciamos o processo de extração dos dados utilizando a linguagem Python e as bibliotecas Numpy, Pandas e Dask. Com a biblioteca Dask, foi possível abrir o microdado sem carregar os dados para a memória da máquina utilizada, mas somente sua estrutura, sem comprometer assim tempo, processamento e memória. Em seguida, extraímos todos os dados referentes ao Rio Grande do Norte.

Posteriormente demos entrada ao processo de limpeza e tratamento dos dados, removendo inconsistências e/ou tratando dados faltosos conforme as necessidades do estudo de caso.

Terminado o tratamento dos dados, prosseguimos para a exploração. Percebemos que muitos participantes se inscreveram, responderam ao questionário socioeconômico, porém não foram fazer a prova. Retiramos estes faltosos para termos dados mais precisos.

Neste primeiro momento, decidiu-se que trabalharíamos somente com as notas das avaliações. Se utilizando do algoritmo K-means, mais precisamente da biblioteca de aprendizado de máquina Scikit-Learn, efetuamos um estudo dos dados. Começamos com o método Elbow, já que não sabíamos em quantos grupos poderíamos agrupar os participantes. Basicamente o que o método faz é testar a

variância dos dados em relação ao número de clusters(grupos). É considerado um valor ideal de clusters quando o aumento no número de clusters não representa um valor significativo de ganho.

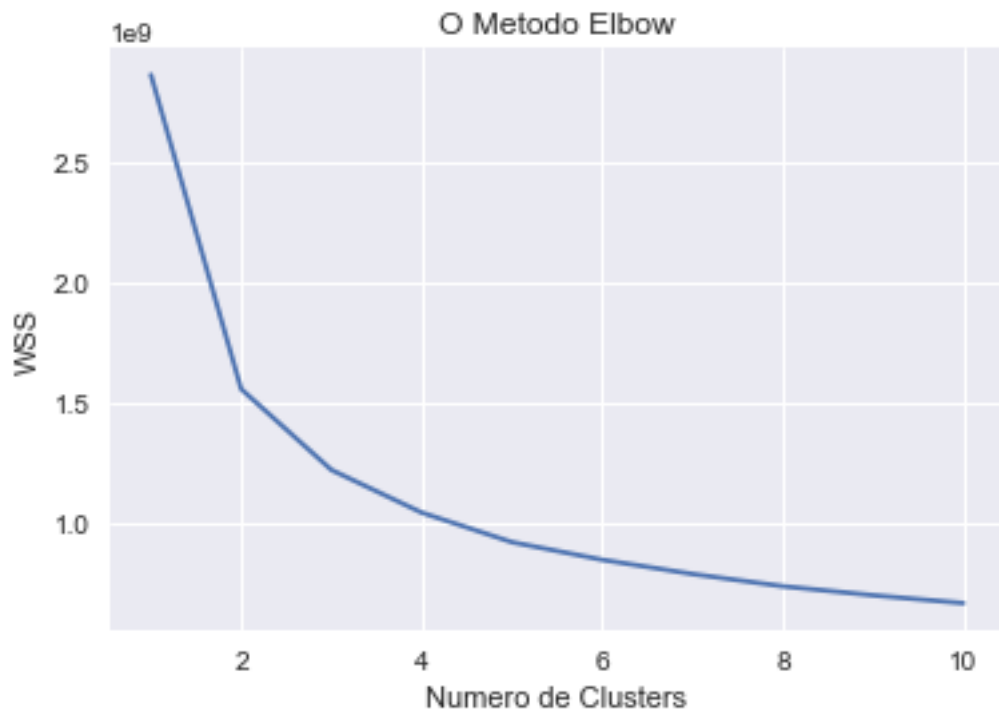


Figura 4: Resultado do Método Elbow (fonte: Autoria Própria. 2018)

Conseguimos identificar três grupos de alunos participantes. Por se tratar de um algoritmo de aprendizado não-supervisionado, não foi necessário passar nenhum dado além das notas dos participantes em todas as avaliações e o número de grupos. Ao final o K-means nos entregou 3 classes de participantes, que logo inserimos na tabela original, na nova coluna intitulada K-Answer.

De posse dos grupos e das notas, iniciamos a fase de visualização dos dados. Utilizando as bibliotecas Matplotlib e Seaborn, começamos criando um gráfico de visualização em pares com o Seaborn chamado pairplot, muito utilizado para

encontrar alguma correlação entre os dados analisados.



Figura 5: Pairplot das notas dos participantes (Fonte: Autoria Própria, 2018)

Percebemos que os grupos se distinguem muito bem entre si quando cruzávamos os dados das notas em diferentes provas e a nova classificação que nos foi dada pelo K-means. A classe 0 foi a que tirou as piores notas, a classe 1 a que se destacou com as melhores notas e a classe 2 a que obteve notas variando de regular a boa.

Continuamos as pesquisas para entender quem eram esses três grupos cruzando os dados deles com as respostas do questionário socioeconômico. A primeira pergunta era a respeito do grau de escolaridade do pai.

0	Nunca estudou.
1	Não completou a 4ª série/5º ano do Ensino Fundamental.
2	Completo a 4ª série/5º ano, mas não completou a 8ª série/9º ano do Ensino Fundamental.
3	Completo a 8ª série/9º ano do Ensino Fundamental, mas não completou o Ensino Médio.
4	Completo o Ensino Médio, mas não completou a Faculdade.
5	Completo a Faculdade, mas não completou a Pós-graduação.
6	Completo a Pós-graduação.
7	Não sei.

Figura 6: Quadro de opções respondidas (Fonte: Autoria Própria, 2018)

A classe 0 tinham em sua maioria pais que não completaram o 5º ano do ensino fundamental. A classe 1 tinham em sua maioria pais que completaram o ensino médio e alguns com ensino superior completo, enquanto que a classe 2 tinham maioria dos pais dividida entre o ensino médio completo e o ensino fundamental incompleto.

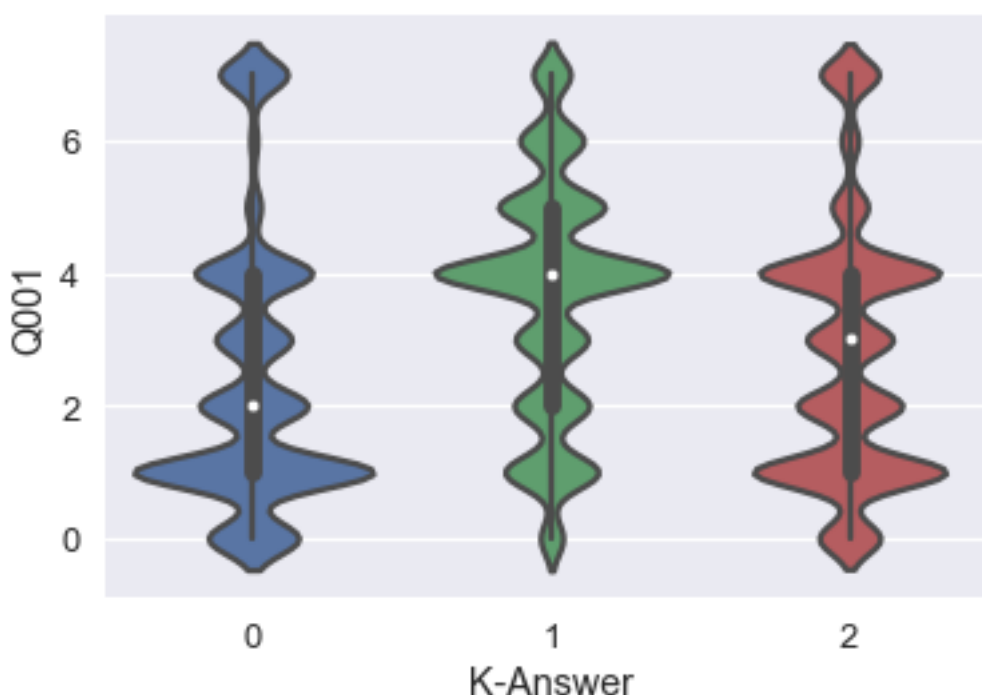


Figura 7: Respostas da Questão 1 (Fonte: Autoria Própria, 2018)

Na segunda questão foi perguntado sobre o grau de escolaridade da mãe. A classe 0 tinha em sua maioria mulheres que não completaram o 5º ano do ensino

fundamental. A classe 1 tinham mães que completaram o ensino médio e o ensino superior, já a classe 2 tinha maioria das mães com ensino médio completo e uma parte com e o ensino fundamental incompleto.

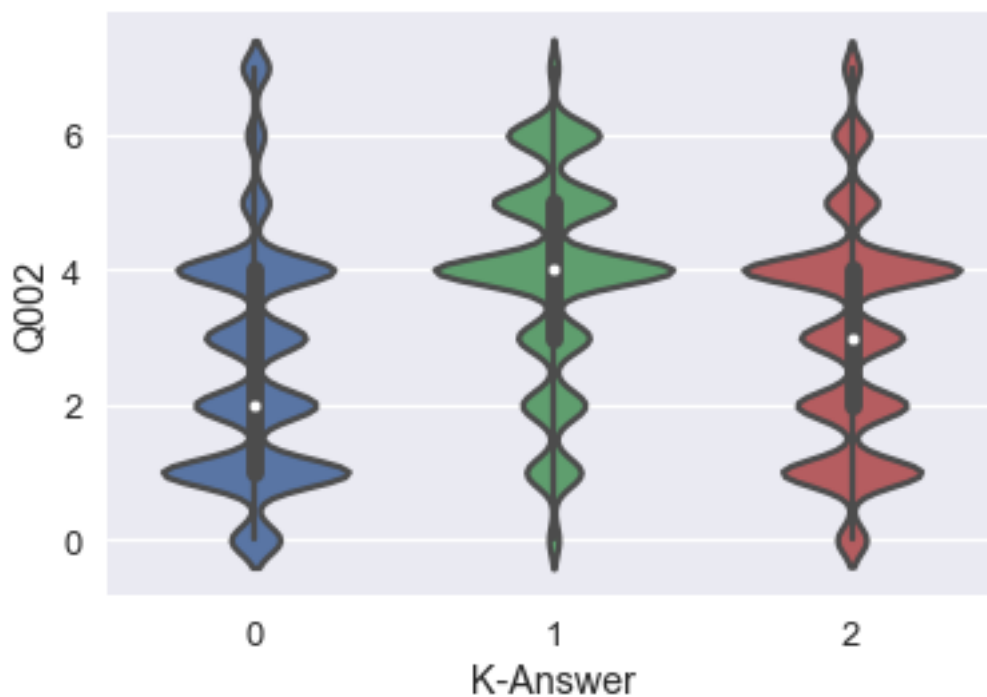


Figura 8: Respostas da Questão 2 (Fonte: Autoria Própria, 2018)

Ao observar os gráficos foi possível perceber que as famílias em que os pais priorizaram os estudos na própria formação, possuem filhos com notas elevadas no Exame.

5 CONCLUSÃO

Atualmente o cientista de dados tornou-se indispensável para o desenvolvimento e evolução de um produto ou de um serviço em qualquer empresa. O “novo mundo” é feito de dados, estruturados, sem estrutura alguma, mas recorrente, ou seja, de tempos em tempos um novo bloco de dados com 10000 linhas e 300 colunas chegará e precisará ser coletado, limpo, normalizado, combinado, estruturado e organizado para novas análises. Saber contextualizá-lo será indispensável para a tomada de decisão da empresa.

A revista Harvard Business Review publicou uma matéria sobre dizendo que o trabalho do Cientista de Dados seria o trabalho mais Sexy do século 21(DAVENPORT, 2012). Uma questão levantada por ela foi se algumas empresas seriam sensatas em esperar até que a segunda geração de cientistas de dados surgisse, e os candidatos fossem mais numerosos, menos caros e mais fáceis de examinar e assimilar em um ambiente de negócios. O problema com esse raciocínio é que o avanço do big data não mostra sinais de desaceleração. Se as empresas ficarem de fora nos primeiros dias dessa tendência de falta de talentos, correm o risco de ficar para trás à medida que os concorrentes e os parceiros de canal ganham vantagens quase inatacáveis.

A linguagem Python, usada no estudo de caso, é simples e de fácil aprendizado. Ótima para criação de scripts e automatização de regras para mineração de dados. Se integra facilmente em um fluxo de trabalho de desenvolvimento de software e pode ser usada em diferentes partes nesse processo (backend, frontend). Não é tão completa para análise estatística como R, mas já percorreu um longo caminho nestes últimos anos. A velocidade de Numpy e Pandas para processar grandes volumes de dados é impressionante. Dask permitiu que fosse aberto o microdado e selecionado todo o conteúdo para pesquisa sem que fosse necessário carregar qualquer linha ou coluna, além de paralelizar todo o processamento, fazendo com que todos os núcleos do meu processador fossem usados. Enfim, uso das tecnologias apresentadas mostrou-se muito proveitosa.

Estamos em uma nova Era, a dos Dados. Informação é o Novo Petróleo e se quisermos pega-lo, vamos precisar de pessoas que saibam perfurar.

REFERÊNCIAS

CONWAY, D. **THE DATA SCIENCE VENN DIAGRAM**. Disponível em: <http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram> . Último acesso em nov. 2018.

DAVENPORT, Thomas H. & D. J. PATIL. **DATA SCIENTIST: The Sexiest Job of the 21st Century**. United States of America: Harvard Business Review. 2012

GRUS, Joel. **DATA SCIENCE FROM SCRATCH: First Principles with Python**. First Edition. United States of America: O'Reilly Media, Inc. 2015

HAIR, J.F.Jr.; ANDERSON, R.E.; TATHAM, R.L.; BLACK, W.C. **ANÁLISE MULTIVARIADA DE DADOS**. 5ª edição. Porto Alegre: Bookman, 2005.

LOURENÇO, Alexandre; MATIAS, Rui P. **ESTATÍSTICA MULTIVARIADA**. Instituto Superior de Engenharia do Porto. 2000

MAROCO, J. **ANÁLISE ESTATÍSTICA: Com a utilização do SPSS**. Lisboa: Silabo, 2003.

MCKINNEY, Wes. **PYTHON FOR DATA ANALYSIS**. First Edition. United States of America: O'Reilly Media, Inc. 2013

MKOMO. **A HISTORY OF STORAGE COST**. Disponível em: <http://www.mkomo.com/cost-per-gigabyte-update> . Último acesso em nov. 2018.

NAUR, Peter. **CONCISE SURVEY OF COMPUTER METHODS**. Suécia. Petrocelli Books. 1974

PROVOST, F., FAWCETT, T. **DATA SCIENCE AND ITS RELATIONSHIP TO BIG DATA AND DATA-DRIVEN DECISION MAKING**. Big Data Journal, Vol. 1, pp. 51-59. 2013

ROCKLIN, Matthew. **DASK: Parallel computation with blocked algorithms and task**

scheduling. In: Proceedings of the 14th Python in Science Conference. 130-136. Citeseer. 2015

TIME, Magazine. **EXPONENTIAL GROWTH IN COMPUTING POWER WILL LEAD TO THE SINGULARITY**. Disponível em: <http://content.time.com/time/interactive/0,31813,2048601,00.html> . Último acesso em nov. 2018.

VANDERPLAS, Jake. **PYTHON DATA SCIENCE HANDBOOK**: Essential Tools for Working with Data. First Edition. United States of America: O'Reilly Media, Inc. 2017

YIN, Robert K. **ESTUDO DE CASO**: Planejamento e métodos. 4ª Edição. Porto Alegre. Bookman. 2010