Gestão & Tecnologia
Faculdade Delta
Ano VIII,
V. 1 Edição 28
Jan/Jun 2019

ESTUDO DE DADOS DE NEGÓCIO COM DATA SCIENCE

STUDY OF BUSINESS DATA WITH DATA SCIENCE

Lucilius Barsanulfo de Meneses[1];
Ueslei José Ferreira da Silva[1];
Reinaldo de Souza Júnior[2].

RESUMO

Esta pesquisa tem por objetivo principal, analisar o perfil dos alunos e estudo de comportamento do processo de utilização da biblioteca da Faculdade Delta e desenvolver uma análise para gerar possíveis resultados para as campanhas de marketing, no intuito de fazer melhorias no acesso aos cursos para facilitar a relação entre aluno e instituição. Os instrumentos utilizados para a coleta dos dados foram os acessos ao banco de dados da faculdade, onde, o tratamento desses dados e o desenvolvimento da análise foram feitos por meio de técnicas adotadas usando linguagem Python e a IDE Jupyter. Com os resultados das análises, a Faculdade Delta saberá qual o perfil de seus alunos e terá assertividade em decisões que possam trazer retornos positivos para a institui-ção. Este trabalho conta com uma continuidade, sendo que, futuramente, poderão ser criadas análises de previ-sões para antecipar tomadas de decisões de alunos, que por alguns motivos, abandonaram os cursos nos quais estão matriculados.
Palavras-chave: Análise. Big data. Dados. Data science. Machine learning.

ABSTRACT

This research have the main objective to analyze the profile of the students and behavior process of use of the Delta Faculty library. The intention of this analysis is to generate possible results for marketing, in order to make improvements in access to the courses to facilitate the relation between student and institution. The instru-ments used to collect this data were the access to the faculty database, where the processing of this data and the development of the analysis through techniques adopts were done using Python and IDE Jupyter. With the results of the analysis, Delta College will know the students profile and will have assertiveness in decisions that could bring positive returns to the institution. This work has a continuity, where in the future, it can anticipate student decision-making where in some cases, students left the courses where they are enrolled.
Keywords: analysis, big data, data, data science, machine learning.

[1]Faculdade Delta. Graduando do curso de Sistemas de Informação. E-mail: uesleijf@gmail.com; luciliusmene-ses@gmail.com
[2]Faculdade Delta. Mestre em Planejamento e Computação Aplicada Orientador de TCC no curso de Sistemas de Informação. E-mail: reinaldodesouzajr@gmail.com

ISSN 2176-2449

1 INTRODUÇÃO

A análise do estudo de dados de negócio com Data Science é essencial e indispensável para empresas que desejam um resultado assertivo de análises de negócios como: perfil dos seus clientes, porcentagem de lucros ou quadro de prejuízos. Dessa forma, a tomada de deci-são para a resolução de um problema fica mais objetiva com menos chances de erros.

A quantidade de informação que foi gerada com o avanço da tecnologia aumentou de uma forma rápida e desordenada na última década. Por isso, ter uma maneira eficiente de fil-trar tais informações, aproveitando somente o que é importante e o que envolve retorno finan-ceiro e social para as empresas, é essencial para se manter à frente no mercado e no quadro de competitividade.

Pesquisas mostram que 75% de grandes empresas mundiais tomaram a decisão de inves-tir nesse estudo de negócio a partir do ano de 2015, o que gerou um retorno de aproximada-mente US $ 1 Bilhão em financiamento nesse período. As empresas que tomaram essa decisão ficaram à frente das concorrentes que não tiveram essa visão (GARTNER, 2015).

Com o objetivo de analisar o perfil dos alunos da Faculdade Delta e a movimentação de empréstimos de livros de sua biblioteca no ano de 2017. Por meio de amostra da base de da-dos da Faculdade foi analisado: Qual o perfil de seus alunos? Quantos livros foram locados no período analisado? Não havia o resultado exato e o impacto alcançado nas campanhas da Fa-culdade em relação ao seu público alvo.

Com a assertividade dos resultados das análises, a Faculdade Delta teve o resultado do perfil de seus alunos e pôde ter acesso sobre qual o fluxo de locação da sua biblioteca, tendo assim a tomada correta de decisão em campanhas futuras de divulgação dos seus cursos e quanto a investimentos na biblioteca e nas áreas mais procuradas pelo seu público alvo.

Foram utilizadas na análise de dados, ferramentas computacionais para a criação de gráficos, cálculos, estatísticas e tratamento de informações. Com isso, foram obtidos os resul-tados esperados para a análise necessária nas tomadas de decisões.

O uso de Big Data nas empresas é fator essencial para o sucesso em tomadas de decisões e destaque em relação aos concorrentes no segmento. O investimento nessa análise consiste na extração de informações específicas e importantes na base de dados da empresa. O Big Data já ganhou espaço em empresas como: Google, Microsoft e Facebook. Porém, o crescimento do Big Data ocorre de forma desordenada e é necessário transformar os dados em informações úteis. A partir dessa necessidade, foi criado o conceito de Data Science, que por meio de um Cientista de Dados (Data Scientist), são obtidas informações úteis com o tratamento dessa grande massa de dados.

A capacidade autônoma de tomada de decisões pelos computadores, definida como Machine Learning, é uma das vertentes da Inteligência Artificial. Com o crescimento acelera-do do Big Data, as tomadas de decisões devem ser feitas da forma mais rápida possível, para melhor aproveitamento. Assim, os sistemas de aprendizado de máquina ganharam destaque e investimentos para serem eficientes e acompanhar a necessidade. O Facebook é um exemplo de empresa que investiu nessa tecnologia e usa a Inteligência Artificial para reconhecimento de imagens.

Com o avanço da tecnologia, surgiu um novo conceito de indústria: a Indústria 4.0, que foi construída com base em seis princípios (FAUSTINO, 2016):

tomada de decisão em tempo real a partir de análise de dados; monitoramento remoto por meio de sensores instalados na estrutura física da empresa; decisão autônoma das máquinas eventualmente em caso de neces-sidade; uso de softwares orientados a serviço; flexibilidade e rapidez para mudar o uso das máquinas; e, por último, comunicação de toda a empresa por meio da internet.

Para melhor entendimento do departamento de comunicação e investimentos da Facul-dade Delta, foram desenvolvidos o estudo do perfil dos alunos de cada curso e o estudo do comportamento das movimentações de empréstimos da sua biblioteca de forma geral e por curso. O resultado desses dois estudos tem como objetivos: uma melhoria da assertividade no alcance das publicações; e no investimento de seu acervo de livros na biblioteca.

1.1 Contexto Atual das Empresas no mundo de Big Data

Empresas sejam elas grandes ou pequenas, deveriam começar a inserir o uso de Big Data em suas rotinas de planejamentos.

As maiores empresas de tecnologia como, Google, Microsoft e IBM já investem no futu-ro do Big Data. Em pouco tempo é certo que, não só estas empresas irão investir e melhorar a iteração na troca de dados, mas também centenas de outras empresas entrarão nessa corrida com pequenos ou grandes investimentos e se beneficiarão dos resultados.

Foi previsto que 75% das empresas investiriam (ou planejariam investir) em big data nos pró-ximos dois anos. Big data cresceu 25% apenas em 2015. Em 2016, o setor garantiu mais de US $ 1 bilhão em financiamento, superando largamente o ano anterior. (GARTNER, 2015, s.p.).

É importante frisar que Big Data é um conceito de análise de dados voltado para extra-ir informações específicas de uma base real ou uma fonte de informações confiável para serem tratados e implementados em planejamentos decisivos. Na prática o uso habitual do Big Data pode trazer benefícios bem maiores do que esperado como previsões de alta precisão indepen-dente do mercado onde se é aplicado.

Empresas que investem para obter essas informações podem sair na frente dos seus concorrentes fazendo a diferença com visão mais ampla sobre seu público alvo em busca de fidelização e pontos relevantes. (VERT, s.d.)

1.2 Big Data

O tratamento do dilúvio de dados produzidos pelas ciências e por bilhões de usuários de serviços de Internet se apresenta como um dos grandes desafios para a atual sociedade do conhecimento de forma acelerada (BELL et al., 2009 apud PORTO e ZIVIANI, s.d.). Há pouco tempo, Terabytes era uma medida de armazenamento usada somente em empresas, po-rém hoje é uma realidade em nossas casas. Esse crescente volume de dados é nomeado como “Big Data”. Visto hoje como o grande avanço da tecnologia e ainda em crescimento tem desa-fiado cientistas e vem impulsionando iniciativas nas mais diversas áreas da tecnologia.

Novas tecnologias foram criadas para trabalhar com grandes quantidades de dados ge-rados diariamente por meio de sistemas corporativos, e-mails e mí-

dias sociais. Tais ferramentas são responsáveis por analisar e filtrar os dados em tempo real e criar estatísticas. Essas novas tecnologias podem ser utilizadas em duas áreas: as ferramentas de análises e as ferramentas de infraestruturas que armazenam todos os resultados das análises.

Em um cenário dinâmico, onde são gerados por dia 2,5 quintilhões de bytes (BIG DA-TA, s.d.), saber aproveitar esses dados de forma útil e rápida é essencial para ter sucesso no mercado e conseguir estar sempre um passo a frente com o Big Data.

Uma pesquisa de 2013, feita pela Universidade de Oxford, mostra o resultado positivo gerado pelo Big Data nas empresas atuais. Foram consultadas 1.144 empresas de 95 países, inclusive o Brasil, onde mostrou-se que o Big Data já era usado em 53% das empresas para entender e melhorar o resultado final para o cliente. Analisando o setor de esportes como em-presa, um exemplo desse uso é a liga NBA de Basquete (HEKIMA, 2016). Os Softwares que são usados no Big Data filtram os dados não tratados, soltos e sem ligação até transformá-los em dados valiosos que serão usados para tomar as decisões assertivas e coerentes ao contexto da regra de negócio.

Big Data se baseia em cinco "V's" (IBM, s.d.):

Velocidade – o tráfego de dados na internet em 2018 é estimado em 50.000 GB/Segundos. Uma estimativa mais precisa é que a cada 60 segundos 70 horas de vídeos são postados no Youtube, 216.000 fotos são publicadas no Instagram e 204.000.000 de e-mails são enviados;

Volume – Diariamente são criados 2.5 Quintilhões de dados, onde 90% desses dados foram criados nos 2 últimos anos;

Variedade – A diversidade desses dados gira em torno de vídeos, imagens e documen-tos, e sua grande parte, por volta de 90% são de dados não estruturados;

Veracidade – 1/3 das empresas nos Estados Unidos não tem confiança na veracidade desses dados para tomar decisões e a estimativa de prejuízo passa de 3 milhões de Dólares por causa da má qualidade dos dados; e

Valor – A capacidade de ter tomada de decisões a partir de um resultado preciso obti-dos por análises. Algumas empresas por meio dessas análises podem prever com precisão de 97%, eventos que causariam prejuízos milionários.

1.3 Indústria 4.0

Indústria 4.0 é um novo conceito de indústria que envolve tecnologias como: automa-ção, controle e tecnologia da informação. Considerada uma nova revolução industrial, teve início na Alemanha, onde foi citada pela primeira vez na Feira de Hannover em 2011. Seu conceito principal é de que empresas poderão criar redes inteligentes em todos seus setores, podendo assim controlar suas produções de forma autônoma, executando manutenções, corri-gindo falhas nos processos e tomando decisões não previstas no planejamento. Os principais responsáveis por esse projeto são: Siegfried Dais (Robert Bosch GmbH) e Kagermann (aca-tech) (SILVEIRA, s.d.).

Para o desenvolvimento da Indústria 4.0, é necessário seguir 6 princípios básicos (FAUSTINO, 2016):

- Capacidade de operação em tempo real: tomada de decisão em tempo real através da análise de dados;

- Virtualização: monitoramento remoto de cópias virtuais das fábricas através de sensores es-palhados por ela;

- Descentralização: tomada de decisão autônoma das máquinas devido a alguma necessidade e fornecimento de informações do seu ciclo de trabalho;
- Orientação a serviços: uso de softwares orientados a serviço;
- Modularidade: flexibilidade para mudar o uso das máquinas rapidamente para atender a grande demanda de serviços;
- Interoperabilidade: capacidade de toda a indústria (seres humanos e máquinas) se comunicar através da internet.

São consideradas algumas tecnologias pioneiras na Indústria 4.0: Cloud Computing, In-ternet das Coisas, Big Data Analytics e Tecnologia da Segurança (sistemas de informação).

O Brasil tem alguns desafios na indústria e na economia para se adequar ao perfil de In-dústria 4.0, tais como: ocupa o 69º lugar no Índice Global de Inovação; menos de 10% da Indústria representa o PIB; Queda de 7% na sua produtividade entre 2006/2016; e está em 29º lugar no Índice Global de Competitividade da Manufatura. Apesar dos números não muito favoráveis, o Brasil está no perfil de países que possuem grande potencial para melhorar sua posição em um futuro breve. (ABDI, s.d.)

Apesar das dificuldades, no Brasil a Indústria 4.0 já tem seu horizonte traçado e uma mineradora no Pará já começou a percorrer esse caminho. A mineradora pode ser operada a dois mil Km de distância (de São Paulo à Parauapebas no Pará), poupando assim tempo de viagem, locomoção e gastos de uma equipe especializada para algum tipo de manutenção ou processo diário (OLHAR DIGITAL, 2018).

A estrutura a mineradora conta com a conexão com fibra ótica e a segurança é do mesmo nível que é usado em bancos, garantindo assim a integridade dos dados enviados de São Paulo até o Pará, o que faz uso de alguns dos seis princípios da Indústria 4.0: Virtualização (que reproduz o funcionamento da fábrica de forma exata no ambiente digital) e Interoperabilidade (capacidade de comunicação entre seres humanos e máquinas através da internet).

1.4 Machine Learning

Com a rápida mecanização trazida pela revolução tecnológica, como o Big Data veio e movimentou a indústria de tecnologia, o Machine Learning está ganhando importância e tem lidado de maneira robusta com uma enorme quantidade de dados fazendo previsões precisas.

Os sistemas de aprendizado de máquina existem desde os anos 50, então por que exis-tem avanços em tantas áreas diferentes? Três fatores estão em jogo: dados enormemente au-mentados, algoritmos significativamente aprimorados e hardwares substancialmente mais po-derosos. Com o aumento da implementação e uso de tecnologias outrora revolucionárias, como Big Data, a Internet das Coisas (IoT), Machine Learning (ML) e agora Deep Learning (DL) estão gradualmente se movendo para o caminho de negócios tradicionais.

O artigo da Harvard Business Review escrito por Hilary Mason em Jul 2017 e intitula-do How AI Fits in Your Data Science Team, afirma que a Inteligência Artificial (IA) e o Aprendizado de Máquina em breve assumirão o status do mecanismo, trazendo mudanças radicais para nossas vidas cotidianas. O poder de transformação da IA e da ML já foi percebi-do no atendimento ao cliente (assistentes digitais), na telemedicina (assistência assistida ao paciente), no setor bancário e financeiro (representantes de vendas e robôs) ou na fabricação (trabalhadores da linha de montagem de robôs).

Outro fator importante são as aplicações que utilizam IA e que já estão ativas em nosso cotidiano. Aplicativos de reconhecimento de voz como Siri (Apple) e Cortana (Microsoft) empregam inteligência artificial. O Facebook usa IA com reconhecimento de imagem para identificar rostos familiares da sua lista de contatos. Os sites de comércio eletrônico, como a Amazon, usam-no para personalizar recomendações baseadas em compras ou atividades anteriores.

O Google usa isso para melhorar as pesquisas. O Google Maps sugere as rotas mais rá-pidas analisando a velocidade do tráfego obtida a partir de dados anônimos de localização do smartphone. Empresas financeiras, incluindo o PayPal, usam algoritmos de inteligência artifi-cial para combater fraudes. Serviços de música como o Spotify e outros personalizam suas ofertas através da IA.

Erik Brynjolfsson e Andrew McAfee argumentam que a IA e o aprendizado de máquina logo se tornarão tecnologias de propósitos gerais tão significativas quanto à eletricidade ou o motor de combustão. Elas representam uma mudança marcante em nossas capacidades técni-cas de impulsionar a próxima onda de crescimento econômico. (BRYNJOLFSSON; MCA-FEE, 2017).

1.4.1. Definição

Machine Learning é uma ramificação da Inteligência Artificial cujos sistemas computa-cionais podem adquirir a capacidade de tomar decisões sozinhos. Essa capacidade vem por meio do aprendizado do sistema, em que vários tipos de dados são estudados pelo sistema. Dessa forma, são criados padrões e métodos de análises de dados automatizados sem a neces-sidade de intervenção humana. (SAS, s.d.).

Os métodos de aprendizado do Machine Learning são três:

-Aprendizado Supervisionado: o sistema tem parâmetros para comparar com a sua deci-são tomada, assim pode aprender o que é certo ou o que é errado através de comparações;

-Aprendizado Semi Supervisionado: Semelhante ao aprendizado supervisionado, porém fazendo treinamento com e sem parâmetros de comparação;

-Aprendizado Não Supervisionado: O sistema não tem nenhum parâmetro para comparar com a sua resposta, tendo que descobrir alguma estrutura dentro dos dados fornecidos.

1.5 Data Science

Transformar dados gerados em informações úteis é sem dúvida muito importante, por meio dessa necessidade foi criado o conceito de Data Science, que, consiste na extração de conhecimento e informações para tomada de decisão através de uma grande base de dados: Big Data (TIME MJV, 2015). Os cientistas de dados são os profissionais responsáveis por extrair os dados úteis para auxiliar a tomada de decisão.

Por meio do Data Science, são obtidos resultados como: análises assertivas; retorno so-bre investimentos; criação de estratégias, agilidade na tomada de decisão. Data Science está diretamente relacionada ao Big Data, pois necessita de diferentes tecnologias para a análise precisa e rápida de seus dados. (CETAX, s.d.).

1.5.1 Definição

A profissão de Data Scientist se resume em cinco tipos de tarefas: filtragem de dados; perguntas objetivas e precisas; análise a partir do uso de dados estatís-

ticos e desenvolvimento de Machine Learning; resultados visualizados e a melhoria de modelos e algoritmos para me-lhores rendimentos; resultados e execução. Apesar da evolução rápida dos computadores com Machine Learning, ainda é importante a presença de um cientista de dados com experiência e domínio de toda essa tecnologia para identificar pontos em comum por meio de desafios di-versos. Para um resultado preciso e satisfatório homem e máquina tornam-se uma equipe unida e ambos são essenciais para o sucesso do trabalho.

Fica claro que os resultados satisfatórios de qualquer negócio não dependem exclusi-vamente da quantidade de dados que uma empresa tem, mas sim sobre a forma como serão usadas essas informações e é esse o maior destaque da Data Science e do Data Scientist. (MATOS, s.d.)

Algumas principais vantagens de usar Big Data (CORDEIRO, 2017):

-Serviços financeiros: por meio de análise de dados muitas instituições financeiras acompanham as manifestações emocionais dos clientes pelas mídias sociais, diagnosticando com antecedência as insatisfações e ganhando tempo para neutralizá-las antes da migração entre as instituições, ou fechamento de contas, no caso dos bancos.

-Varejo: por meio da coleta e análise de dados, empresas de varejo costumam identificar os hábitos e preferências de consumo de clientes e informações sociais e demográficas. Com isso, aumentam o número de vendas e elaboram programas de fidelidades mais atraentes. Ou-tro exemplo é o levantamento de dados de antigos clientes e o cruzamento com dados de pro-dutos preferidos por eles: a partir disso, geram-se descontos em produtos específicos atraindo novamente o cliente.

1.5.2 Data Scientist

Não existe formação específica e acadêmica para se tornar um Data Scientist, logo vá-rios profissionais de algumas áreas podem ter o perfil desejado para ingressar nessa carreira:

-O princípio básico é a matemática e estatística, uma vez que algoritmos de

Machine Learning são na sua maior parte baseados em conceitos matemáticos.

-Conhecimento de programação em linguagens como R, Python, Java e em banco de dados, uma vez que interações com bancos de dados relacionais e não relacionais fazem parte do processo de análise.

-Conhecimento de ecossistemas que armazenam os dados que serão usados para a aná-li-se, Hadoop, Streaming de dados com Spark são exemplos dessas estruturas. (DATA SCIEN-CE ACADEMY, 2018a). Os cientistas de dados usam ferramentas de algoritmos para enten-der e prever o desempenho de um negócio. Entender que o desempenho requer um conjunto de habilidades mais técnicas com base em estatísticas, aprendizado de máquina e programa-ção.

1.5.3 Processos de Análise de Dados

No processo de análise de dados, foram levantados todos os requisitos necessários a par-tir das seguintes etapas: Question, aprofundamento no conhecimento de negócio, qual o pro-blema a ser resolvido? Como será resolvido o problema?; Wrangle, aquisição e filtragem dos dados medidos de forma correta para a sua análise; Explore: entendimento das informações medidas e definimos padrões para todas as informações; Conclusions: realização de conclu-sões sobre os dados avaliados; Communicate: comunicação de forma clara e simples ao cliente

sobre os resultados das análises, por meio de slides, gráficos e planilhas.

1.5.4 Ferramentas

Python

Lançada no ano de 1991, por Guido Van Rossum, com um fácil entendimento, e com o paradigma orientado a objetos, tem uma curva de aprendizagem curta até mesmo para inician-tes na área. A sua usabilidade abrange os segmentos desktop, projetos web e também mobile, apesar de ser mais popular entre usuários do Linux, a linguagem é multi-plataforma existindo IDE's e interpretadores para Windows e Mac OS.

Devido a essa baixa curva de aprendizagem, a produtividade para o desenvolvimento em Python é alta (PYSCIENCE-BRASIL, s.d.). Uma grande vantagem é de não necessitar de licença para sua utilização, tanto para estudo ou uso profissional. (BRUNO, 2010)

Com todas essas características, Python se tornou uma das linguagens mais usadas no processo de tratamento de dados científicos, tais como Machine Learning e Data Science, e estudo de dados estatísticos.

Matplotlib

Baseado no Matlab, o Matplotlib é uma biblioteca de códigos que boa parte é feita em Python, usado na criação de gráficos 2D para estudo de análise de Data Science. Foi desen-volvido para ser usado com poucas linhas de código para uma saída clara e com informações precisas através de seus gráficos.

> As plotagens devem ter ótima qualidade de publicação. Um requisito importante para mim é que o texto pareça bom (antialiased, etc.); Saída Postscript para inclu-são com documentos do Tex; Implantável em uma interface gráfica de usuário pa-ra desenvolvimento de aplicativos; Código deve ser fácil o suficiente para que eu possa entendê-lo e estendê-lo; Fazer plotagens deve ser fácil. (MATPLOTLIB, s.d.)

Com um conjunto de funções (matplotlib.pylab) que permite a criação de gráficos bas-tante semelhante ao Matlab, um poderoso gerenciamento de figuras, textos e gráficos com ótima qualidade. O Matplotlib é usado em muitos cenários diferentes.

Jupyter Notebook

Criado em 2014, por Fernando Perez (DADOS E DECISÕES, 2018), Jupyter é uma plataforma web, que consiste na criação de códigos para análise de Data Science e transforma-ção de dados, em que, de forma amigável, pode-se mesclar códigos com textos e ter uma saída simples e objetiva, graças a sua interface gráfica. Apesar de a plataforma rodar em um navega-dor na porta 8888 por padrão, após a sua instalação não é necessário conexão à internet (NO-VELLO, 2017).

A maior parte dos códigos desenvolvidos é em Python, porém a ferramenta tem suporte para várias linguagens. As primeiras linguagens usadas foram: Python, R, Julia e Scala, equa-ções numéricas visualizações de gráficos, estatísticas e textos. (WILLEMS, 2016).

O lado do serviço é executado via kernel para o código Python, ou outra linguagem; e o lado do cliente acessado no browser. O cliente tem, além da criação dos códigos, o controle do kernel, onde pode desligá-lo ou reiniciar o serviço.

Estatística

A estatística é uma área da Matemática com um amplo campo e aplicações em muitos se-tores. Define-se o estudo da coleta, análise, interpretação, apresentação e organização de da-dos. A estatística pode ser usada dentro da ciência de dados utilizando métodos e experimen-tos para medir determinadas situações e testar inúmeras funcionalidades de uma análise. (DA-TA SCIENCE ACADEMY, 2018b).

Mediante os resultados das análises, foi possível medir quantidade, comparar cursos, tempo e frequência de locação de livros por exemplo.

1.5.4.5 Pandas

Pandas é uma biblioteca construída em Python usada para análise de vários tipos de da-dos: tabelas, tanto Excel, Libre Office quanto tabelas de banco de dados, arquivos CSV, ma-trizes, ou qualquer tipo de dado para análise. Quando usados de maneira correta, esses dados ficam relacionais, de fácil entendimento e intuitivos.

É trabalhado geralmente de duas formas com as estruturas em Pandas: Série, que é um padrão Unidimensional, e DataFrame (GALVÃO, 2016), estrutura Bidimensional. Com essas duas formas, o Pandas consegue tratar grande parte dos casos de uso em finanças, estatísticas, ciências sociais e engenharia. (PANDAS, 2018).

Pode-se destacar algumas usabilidades do Pandas que deixam a análise de dados mais produtiva e intuitiva: flexibilidade na disposição dos dados no DataFrame, junção de conjunto de dados; carregamento rápido de vários tipos de arquivos; tratamento de dados indisponíveis em um certo padrão; alinhamento automático de DataFrames após criação de um padrão; In-dexação e subconjunto de grande massa de dados.

1.5.4.6 Numpy

Numpy tem um papel fundamental na análise de dados, pois ele é uma biblioteca Python que oferece objeto de matriz multidimensional de alto desempenho e ferramentas para traba-lhar com essas matrizes entre outros derivados tais como uma variedade de rotinas de opera-ções mais ágeis, cálculos matemáticos, lógicas, classificação, seleção, álgebra linear básica, simulação aleatória e muito mais. (SCIPY.ORG, s.d.).

2 ESTUDO DE CASOS

O Centro Tecnológico Delta Ltda. constitui-se numa Instituição de Ensino Superior, de caráter privado, em 25 de outubro de 2005, no município de Goiânia, com o objetivo de elevar o nível cultural e profissional da região, ministrando o ensino superior em seus variados níveis do conhecimento, investindo em pesquisa e extensão por meio de sua unidade, a Faculdade Delta. Seu projeto educacional conta com o apoio do Colégio Delta, que atua há 25 anos em Goiânia no Ensino Fundamental e Médio. Os cursos ministrados na Instituição são: Adminis-tração, Ciências Contábeis, Sistemas de Informação, Tecnologia em Gestão Ambiental, Tecno-logia em Gestão de Recursos Humanos e Pedagogia. A Faculdade Delta foi credenciada e autorizada pelo MEC através da PORTARIA-MEC 1.082, de 21 de novembro de 2007.

2.1 Estudo de Caso I – Perfil dos Alunos do curso

O objetivo é fazer um estudo do perfil dos alunos de cada curso, como forma de forne-cer maiores informações, para o departamento de comunicação da faculdade, como forma de melhorar a assertividade do alcance das publicações.

Nesse estudo de caso, são mostrados os resultados específicos do curso de Sistemas de Informação e o resultado geral da Faculdade Delta.

2.1.1 Question

A Faculdade Delta, buscando conhecer seu público, está utilizando as ferramentas como redes sociais e outras mídias para atraí-los, um método que pode trazer resultados positivos. Sabendo disso, foi proposta essa pesquisa de análise de dados para respostas mais assertivas de acordo com o proposto.

2.1.2 Wrangle

A aquisição dos dados do ano de 2017 deu-se por meio e amostra na base de dados da faculdade e da biblioteca os quais foram selecionados, agrupados e extraídos no formato CSV (Comma-Separated Values).

2.1.3 Explore

Foram explorados os dados gerais da Faculdade Delta dos alunos de cada curso sobre a faixa etária (Figura 1), totalizando 512 alunos e, o sexo dos alunos do curso de Sistema de Informação (Figura 2).

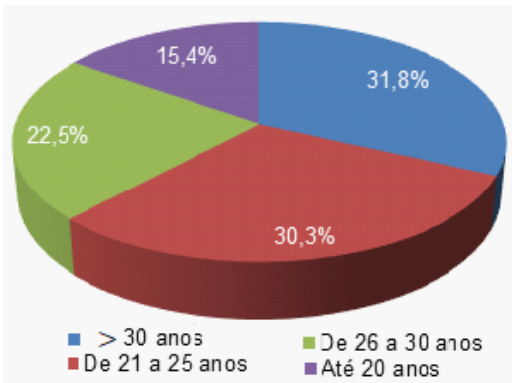

Figura 1. Faixa de idade dos alunos da Faculdade Delta, 2017/1.
Fonte: Elaborado pelos autores.

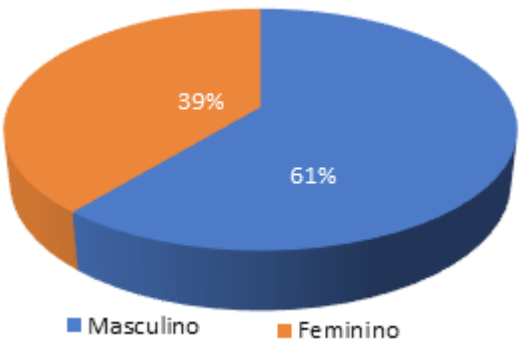

Figura 2. Sexo dos alunos de Sistemas de Informação, 20171.
Fonte: Elaborado pelos autores

O curso de Sistemas de Informação tem 64 homens e 41 mulheres, desta forma, existe uma predominância de homens no curso.

2.1.4 Conclusion

Foram analisadas cada etapa dos dados colhidos levantando assim suas estatísticas e ge-rando gráficos em formato de pizza ou tabela, para cada uma delas, onde foram observados resultados, como mostrado na figura 2 e figura 3.

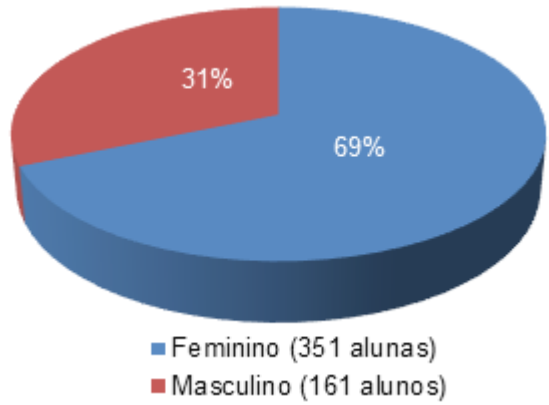

Figura 3. Sexo dos alunos dos cursos da Faculdade Delta, 2017/1
Fonte: elaborado pelos autores.

Por meio de uma rápida análise da figura 3, chegou-se à conclusão de que a Faculdade Delta tem no seu quadro de alunos, uma predominância do sexo feminino, chegando à margem de 69%.

A maior parte dos alunos do curso de Sistemas de Informação mora na cidade de Goiânia. Somando um total de 74 alunos (Tabela 1).

Tabela 1. Cidades dos alunos da Faculdade Delta, 2017/1

MUNICÍPIOS	NÚMERO (PERCENTUAL)
Goiânia	390 (76,2)
Aparecida de Goiânia	97 (18,9)
Goianira	6 (1,2)
Trindade	5 (1,0)
Goiás Velho	2 (0,4)
Senador Canedo	1 (0,2)
Indefinido	8 (3,1)

Fonte: elaborado pelos autores

Por meio de uma breve análise da tabela 1, em relação às cidades onde residiam os alu-nos da Faculdade Delta até o ano de 2017, que se somada à porcentagem das cidades de Goiânia e Aparecida de Goiânia, há 95% do quadro de alunos, com destaque para a cidade de Goiânia, com 76,2% de toda a Faculdade.

Foram criados quatro grupos com perfis diferentes para a análise de faixa etária dos alu-nos do curso de Sistema de Informação da Faculdade Delta. Os grupos foram divididos em: idade até 20 anos; idades entre 21 e 25 anos; idades entre 26 e 30 anos; e por fim, alunos com mais de 30 anos de idade. (Figura 4).

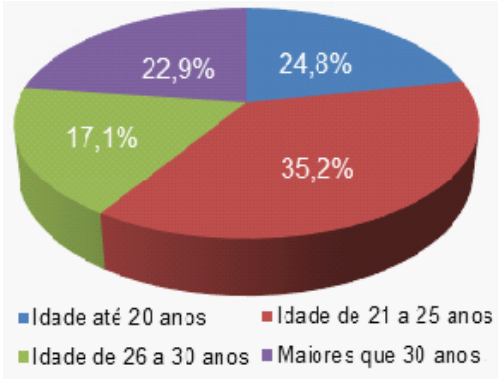

Figura 4. Faixa etária dos alunos do curso de Sistemas de Informação, 2017/1.
Fonte: elaborado pelos autores.

O Curso de Sistema de Informação tem a maior porcentagem no grupo de alunos com a idade entre 21 a 25 anos, com 35,2%.

Na análise dos cursos da Faculdade Delta, a maior porcentagem da faixa etária dos alu-nos está no grupo com idade maior que 30 anos.

2.1.5 Communicate

Os resultados das análises serão entregues para a Faculdade por meio de um relatório com todos os gráficos e seus esclarecimentos e através de uma apresentação para esclarecimen-tos de cada resultado e da análise da aplicabilidade desses dados.

2.2 Estudo de Caso II – Comportamento do Processo de Locação de Livros na Biblioteca

O objetivo dessa análise é fazer um estudo do comportamento das movimentações de empréstimos de livros da biblioteca da Faculdade Delta de forma geral e por curso.

2.2.1 Question

A Faculdade Delta busca saber qual o perfil da movimentação dos livros locados pelos alunos na sua Biblioteca. Por esse motivo, houve a necessidade de uma análise da qual se pode ter um feedback assertivo de qual curso loca mais livros, quais livros são mais locados e em qual período isto ocorre.

Desta forma, a faculdade poderá tomar melhores decisões para cada curso em relação a investimentos com livros físicos para aqueles cursos que têm uma alta procura na biblioteca, ou investimento em um novo tipo de acervo para os cursos que não têm o hábito de procurar livros na Biblioteca.

2.2.2 Wrangle

A aquisição dos dados do ano de 2017 deu-se por meio da extração na base de dados da faculdade e da biblioteca os quais foram selecionados, agrupados e extraídos no formato CSV (Comma-Separated Values).

2.2.3 Explore

Foram explorados os dados da movimentação dos livros locados pelos alunos da Faculdade Delta, e a figura 5 representa um total de 1974 exemplares locados

durante o ano de 2017. A tabela 2, mostra os livros locados por curso no primeiro semestre de 2017.

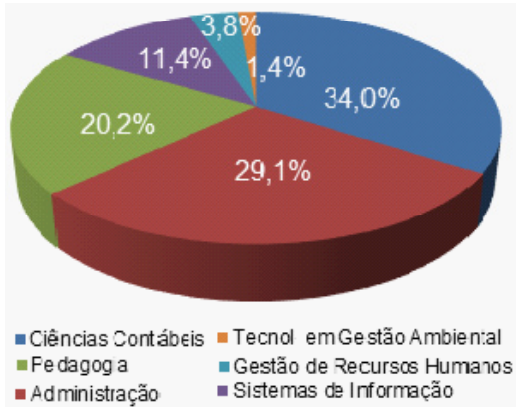

Figura 5. Livros locados por curso no ano letivo 2017
Fonte: elaborado pelos autores

Tabela 2. Livros locados por curso no primeiro semestre de 2017.

PRIMEIRO SEMESTRE DE 2017	
CURSOS	**N (Percentual)**
Administração	306 (37,7)
Ciências Contábeis	292 (36)
Pedagogia	95 (11,7)
Sistema de Informação	57 (7,0)
Gestão de Recursos Humanos	35 (4,3)
Tecnologia da Gestão Ambiental	27 (3,3)

Fonte: elaborado pelos autores.

2.2.4 Conclusion

Na análise de locação de livros da Biblioteca durante o ano letivo de 2017, analisou-se cada etapa dos dados colhidos levantando suas estatísticas e foram observados os seguintes resultados.

Durante todo o ano de 2017, tivemos um total de 1974 livros locados na Biblioteca da Faculdade. O curso de Ciências Contábeis ficou em primeiro lugar no ranking de locações, com 34% desse total, por outro lado o Curso de Tecnologia em Gestão Ambiental ficou com a menor fatia do gráfico de pizza, com apenas 1,4%, totalizando 27 locações durante todo o ano de 2017.

Na análise de locações de livros na biblioteca por quinzena do primeiro semestre de 2017, obteve os seguintes resultados, mostrado na tabela 3.

Na primeira quinzena de fevereiro, o curso de Pedagogia não locou livros e o curso de Administração teve uma porcentagem de 38,9% das locações da segunda quinzena de fevereiro.

Na primeira quinzena de março, houve um equilíbrio entre os cursos de Administração (43,1%) e Ciências Contábeis (45,9%). O curso de Ciências Contábeis obteve destaque nas locações na primeira (48,4%) e segunda quinzena de abril (57%).

Na primeira quinzena de maio, os cursos de Tecnologia em Gestão Ambiental e Pedagogia estão ausentes, pois ambos não locaram livros nesse período.

Enquanto, o curso de Administração obteve maior porcentagem na locação na segunda quinzena de maio com 47,2%.

Não houve locação do curso de Pedagogia na primeira e segunda quinzena de junho.

Tabela 3. Livros locados por curso no primeiro entre fevereiro a junho de 2017.

Livros locados por curso entre fevereiro a junho de 2017 Números de exemplares (Percentual)							
Meses/ 1ª ou 2ª quinzena	Ciências Contábeis	Administração	Sistema de Informação	Gestão de Recursos Humanos	Tecnologia da Gestão Ambiental	Pedagogia	Total por mês
Fev/1ª	19 (39,6)	13 (27,1)	10 (20,8)	04 (8,3)	02 (4,2)	00 (0,0)	48
Fev/2ª	26 (36,1)	28 (38,9)	05 (6,9)	03 (4,2)	03 (4,2)	07 (9,7)	72
Mar/1ª	50 (45,9)	47 (43,1)	01 (0,9)	02 (1,8)	03 (2,8)	06 (5,5)	109
Mar/2ª	54 (24,3)	38 (27,1)	05 (3,6)	09 (6,4)	00 (0,0)	54 (38,6)	160
Abr/1ª	30 (48,4)	23 (37,1)	03 (4,80)	00 (0,0)	00 (0,0)	06 (9,7)	62
Abr/2ª	55 (57,0)	21 (21,6)	12 (12,4)	06 (6,2)	03 (3,1)	00 (0,0)	97
Maio/1ª	38 (31,7)	69 (57,5)	10 (18,3)	04 (2,5)	00 (0,0)	00 (0,0)	121
Maio/2ª	27(22,0)	58 (47,2)	05 (4,1)	03 (2,4)	08 (6,5)	22 (17,9)	123
Jun/1ª	11 (33,3)	06 (18,2)	04 (12,1)	04 (12,1)	08 (24,2)	00 (0,0)	33
Jun/2ª	02 (28,6)	03 (42,9)	02 (28,6)	00 (0,0)	00 (0,0)	00 (0,0)	7
TOTAL	**292**	**306**	**57**	**35**	**27**	**95**	**832**

Fonte: elaborado pelos autores.

O curso de Administração locou mais livros que os demais cursos.

De forma mais detalhada, são mostrados na tabela 3, os resultados da análise por curso e distribuídos quinzenalmente.

Na análise da tabela 4 sobre a locação dos livros da Faculdade Delta, obtive-se os cinco livros mais locados ao longo do ano de 2017. (Tabela 4).

Tabela 4. Títulos de livros mais locados na biblioteca da Faculdade Delta, 2017.

LIVROS	N (%)
Ética geral e profissional em contabilidade	19 (22,4)
Planejamento estratégico: conceito, metodologia e prática	19 (22,4)
Introdução à teoria geral da administração	16 (18,9)
Manual de contabilidade tributária	16 (18,9)
Planejamento estratégico: conceitos	15 (17,6)

Fonte: elaborado pelos autores.

6.2.5 Communicate

Os resultados das análises serão entregues para a Faculdade por meio de um relatório com todos os gráficos e seus esclarecimentos, além de uma apresentação para esclarecimentos de cada resultado e de análise da aplicabilidade desses dados.

3 CONCLUSÕES

O objetivo desse trabalho foi fazer uma análise permitindo extrair ativos a partir de uma base de dados fornecida em arquivo CSV. Nesse sentido, foi realizada a estruturação dos dados recebidos, organizados visando suportar a sua proposição.

O trabalho desenvolvido atendeu às expectativas e gerou resultados satisfatórios ao longo do prazo devido permitindo a produção e desenvolvimento de

análises que foram apresentadas por meio de gráficos e histogramas. O modelo desenvolvido da análise foi construído de forma a simplificar a pesquisa gerando flexibilidade e escalabilidade deixando o processo minimizado de acordo com as necessidades futuras podendo expandir se assim necessário for.

Existem alguns pontos importantes a se considerar, novas possibilidades foram apresentadas podendo tornar esse processo mais longo e mais completo, porém essa pesquisa foi feita superficialmente sem usar métodos avançados levando em consideração o tamanho dos dados recebidos e o tempo sugerido.

As possibilidades de implementação desse trabalho são enormes e passíveis de adaptação. A associação de elementos para a pesquisa pode gerar diversas vertentes e resultados; mas apesar destes conceitos não terem sido abordados no projeto o modelo desenvolvido atendeu às expectativas, conseguindo mostrar os resultados desejados, evidenciando em questão a situação do perfil dos alunos da Faculdade Delta.

A princípio a dificuldade no desenvolvimento desse trabalho foi a necessidade de aprender uma nova linguagem, Python, juntamente como uma nova forma de analisar um cenário para o levantamento de requisitos. Após definido o objetivo do trabalho, surgiu uma nova dificuldade: saber como criar o cenário para ser analisado, quais os requisitos a serem pesquisados com base nos levantamentos dos dados e que medidas estratégicas poderiam ser tomadas para gerar o resultado satisfatório, saber filtrar em uma base de dados sem um padrão definido e com falta de dados, somente as informações úteis para cada resultado esperado, onde foi mais necessário ponto de vista crítico e domínio das ferramentas do que o desenvolvimento de algoritmos em si. Aprendemos que na análise de dados e estatísticas com DataScience, é necessário o conhecimento de programação, porém, esse não é a vertente principal.

A continuidade dessa primeira fase do trabalho será programar algoritmos de previsões para várias áreas da Faculdade Delta, em que, a partir de análises comportamentais, a tomada de decisão será feita de forma antecipada a possíveis resultados e cenários indesejáveis, evitando dessa forma prejuízos para a faculdade. Podemos citar o caso de uso onde existe um quadro de alunos que têm chances de entrar para o perfil de desistências na faculdade. Por meio de um perfil gerado por algoritmos de previsões, conseguiremos obter todas as informações para identificar e antecipar a decisão de desistência desses alunos para mantê-los na faculdade.

REFERÊNCIAS

ABDI. Agenda brasileira para a Indústria 4.0. s.d. Disponível em: <http://www.industria40.gov.br/>. Acesso em: 05 ago. 2018.

BIG DATA. Big Data: tudo que você sempre quis saber sobre o tema! s.d. Disponível em: <http://www.bigdatabusiness.com.br/tudo-sobre-big-data/>. Acesso em: 30 mar. 2018.

BRUNO. Aprendendo a programar em Python -Introdução. 2010. Disponível em: <https://www.devmedia.com.br/aprendendo-a-programar-em-python-introducao/17093>. Acesso em: 22 ago. 2018.

BRYNJOLFSSON, Erik; MCAFEE, Andrew. What's Driving the Machine Learning Explosion?.2017. Disponível em: <https://hbr.org/2017/07/whats-driving-the-machine-learning-explosion>. Acesso em: 08 maio 2018

CETAX. Data Science: O que é, conceito e definição. s.d. Disponível em: <https://www.cetax.com.br/blog/data-science>. Acesso em: 01 abr. 2018.

CORDEIRO, Cristiano. Vantagens gerais e específicas do Big Data: mostramos tudo aqui! 2017. Disponível em: <http://www.neomind.com.br:81/blog/big-data-quais-as-vantagens-gerais-e-especificas>. Acesso em: 29 set. 2018.

DADOS E DECISÕES, O Jupyter Notebook: o que é? 2018. Disponível em: <https://dadose-decisoes.com.br/o-jupyter-notebook-o-que-e/>. Acesso em: 15 ago. 2018.

DATA SCIENCE ACADEMY. 7 Maneiras que os Cientistas de Dados usam estatística. 2018b. Disponível em: <https://datascienceacademy.com.br/,blog/7-maneiras-que-os-cientistas-de-dados-usam-estatistica/>. Acesso em: 29 set. 2018.

DATA SCIENCE ACADEMY. Cientista de Dados –Por Onde Começar em 8 Passos. 2018a. Disponível em: <http://datascienceacademy.com.br/blog/cientista-de-dados-por-onde-comecar-em-8-passos/>. Acesso em: 18 ago. 2018.

FAUSTINO, Bruno. Seis princípios básicos da Indústria 4.0 para os CIOs. 2016. Disponível em: <http://cio.com.br/noticias/2016/05/02/seis-principios-basicos-da-industria-4-0-para-os-cios/>. Acesso em: 05 ago. 2018.

GALVÃO, Felipe. Ciência de Dados com Python: Básico do Pandas – Leitura de

DataFrames. 2016. Disponível em: <http://felipegalvao.com.br/blog/2016/02/18/ciencia-de-dados-com-python-basico-do-pandas-leitura-de-dataframes/>. Acesso em: 29 set. 2018.

GARTNER. Gartner Survey Shows More Than 75 Percent of Companies Are Investing or Planning to Invest in Big Data in the Next Two Years.2015. Disponível em: <https://www.gartner.com/newsroom/id/3130817>. Acesso em: 26 set. 2018.

HEKIMA. Big Data: tudo que você sempre quis saber sobre o tema! 2016. Disponível em: <http://www.bigdatabusiness.com.br/tudo-sobre-big-data/>. Acesso em: 30 mar. 2018.

IBM. The Four V's of Big Data. s.d. Disponível em: <https://www.ibmbigdatahub.com/infographic/four-vs-big-data>. Acesso em: 15 set. 2018.

MATOS, David. O que é Data Science? s.d. Disponível em: <http://www.cienciaedados.com/o-que-e-data-science>. Acesso em: 30 abr. 2018.

MATPLOTLIB. Matplotlib. s.d. Disponível em: < https://matplotlib.org/users/history.html>. Acesso em: 15 ago. 2018.

NOVELLO, Rafael. Jupyter Notebook na nuvem para análises com muitos dados. 2017. Disponível em: <https://imasters.com.br/cloud/jupyter-notebook-na-nuvem-para-analises-com-muitos-dados>. Acesso em: 15 ago. 2018.

OLHAR DIGITAL. Indústria 4.0: mineradora no Pará pode ser operada a 2.000km de distância. 2018. Disponível em: <https://olhardigital.com.br/video/industria-4-0-mineradora-no-para-pode-ser-operada-a-2-000km-de-distancia/77923>. Acesso em: 06 ago. 2018.

PANDAS. Pandas: powerful Python data analysis toolkit. 2018. Disponível em: < <http://pandas.pydata.org/pandas-docs/stable/>. Acesso em: 20 set. 2018.

PORTO, Fábio; ZIVIANI, Artur. Ciência de Dados. Laboratório Nacional de Computação Científica (LNCC) Petrópolis, RJ, s.d.

PYSCIENCE BRASIL. Python: O que é? Por que usar? s.d. Disponível em: <http://pyscience-brasil.wikidot.com/python:python-oq-e-pq>. Acesso em: 29 set. 2018.

SAS. Machine Learning. s.d. Disponível em: <https://www.sas.com/pt_br/insights/analytics/machine-learning.html>. Acesso em: 20 set. 2018.

SCIPY.ORG. What is NumPy? s.d. Disponível em: <https://docs.scipy.org/doc/numpy/user/whatisnumpy.html>. Acesso em: 10 set. 2018.

SILVEIRA, Cristiano Bertulucci. O Que é Indústria 4.0 e Como Ela Vai Impactar o Mundo. s.d. Disponível em: <https://www.citisystems.com.br/industria-4-0/>. Acesso em: 04 ago. 2018.

TIME MJV. O que é Data Science? 2015. Disponível em: <http://blog.mjv.com.br/ideias/o-que-e-data-science>. Acesso em: 30 set. 2018.

VERT. Por que o Big Data é tão importante para as empresas? s.d. Disponível em: <http://www.vert.com.br/blog-vert/por-que-o-big-data-e-tao-importante-para-as-empresas/>. Acesso em: 15 set. 2018.

WILLEMS, Karlijn. Jupyter Notebook Tutorial: The Definitive Guide. 2016. Disponível em: <https://www.datacamp.com/community/tutorials/tutorial-jupyter-notebook>. Acesso em: 15 ago. 2018.