UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL

ESCOLA DE ADMINISTRAÇÃO

DEPARTAMENTO DE CIÊNCIAS ADMINISTRATIVAS

Karina Vargas de Moura

Data Science:
Um estudo dos métodos no mercado e na academia

Porto Alegre
2018

KARINA VARGAS DE MOURA

Data Science:
Um estudo dos métodos no mercado e na academia

Trabalho de conclusão de curso de graduação apresentado ao Departamento de Ciências Administrativas da Universidade Federal do Rio Grande do Sul, como requisito parcial para a obtenção do grau de Bacharel em Administração.

Orientadora: Dra. Daniela F. Brauner

Porto Alegre
2018

KARINA VARGAS DE MOURA

Data Science:
Um estudo dos métodos no mercado e na academia

Trabalho de conclusão de curso de graduação apresentado ao Departamento de Ciências Administrativas da Universidade Federal do Rio Grande do Sul, como requisito parcial para a obtenção do grau de Bacharel em Administração.

Trabalho de Conclusão de Curso defendido e aprovado em: 11, dezembro 2018.

Banca examinadora:

_____

Prof. Dra. Daniela Francisco Brauner
Orientadora
UFRGS

_____

Prof. Dra. Raquel Janissek-Muniz
UFRGS

Dedico este trabalho ao meu filho, Fernando, meu farol.

AGRADECIMENTOS

Chegar ao fim de uma graduação na UFRGS foi um caminho árduo, cheio de idas e vindas. Entre trocas de curso e buscas pessoais, fui me deparando com diversas pessoas durante esta caminhada. Umas estiveram ao meu lado, apoiando, ensinando... Outras, disseram que a UFRGS não era o meu lugar. Todas essas pessoas foram importantes no meu caminho, na definição do meu caráter. A todos vocês que acreditaram ou não, obrigada.

Este trabalho é o resultado do conjunto de interações que tive durante toda a minha caminhada e, sem as pessoas que citarei a seguir, não teria sido possível.

A todos os gestores e funcionários das empresas entrevistadas, que me receberam com carinho e dispuseram do seu tempo para compartilhar seu conhecimento, obrigada.

A todos os profissionais da UFRGS que fizeram a diferença e se empenharam em compartilhar e construir com respeito e dedicação, obrigada.

Às pessoas e instituições que oportunizaram o acesso ao ensino superior público e de qualidade, muito obrigada.

À minha orientadora e amiga Daniela F. Brauner, que foi muito além das orientações deste trabalho, foi uma inspiração para a vida, obrigada.

A todo o time da Mconf, que me permitiram o tempo necessário para que este trabalho pudesse ser realizado e mais umas dicas extras, obrigada.

Ao meu antigo time, NetMetric PRAV, pelo caminho trilhado juntos, obrigada.

À minha mãe, Elisabeth Vargas, que sempre sonhou com este momento, e que sempre nos falou sobre a importância de estudar, obrigada.

À minha irmã, Ana Carolina, fonte de inspiração e apoio, parceira de vida e desafios, obrigada.

Às pessoas da minha família que dispuseram de seu tempo para ouvir minhas dificuldades e dar apoio. Em especial minha tia, Fabiane Moura, obrigada.

À família do meu marido que nos apoiou em diversos momentos, obrigada.

Aos bests, meus melhores amigos, amigos da vida, a família que veio com a jornada, pelas noites de conversas e desabafo, pelo caminho compartilhado, obrigada.

Ao meu marido, Felipe Nesello, que aceitou este e outros desafios comigo, que não saiu do meu lado, que me aconselhou, apoiou, cuidou do nosso filho. Sem você eu não estaria aqui, obrigada.

“*You have to be twice as good as them to get half of what they have!”*

Rowan

RESUMO

Para análise de dados são utilizados sistemas de informação e de apoio à decisão, sempre alinhados aos métodos de descoberta de conhecimento. Com o crescimento do volume, variedade e velocidade dos dados, as empresas estão conseguindo informações mais qualificadas e reestruturando seus posicionamentos estratégicos a partir da análise dos dados. Desde 1989, diversos métodos foram propostos, mas, atualmente, com a disseminação e os desafios do *Big Data* e do valor do conhecimento extraído dos dados para as organizações, foi preciso buscar novas formas de se realizar as análises abrangendo, assim, o campo da *Data Science*. A *Data Science* entra, então, como área de estudo capaz de auxiliar as empresas a lidar com essa nova sistemática de análise dos dados, permitindo a extração de conhecimento a partir dos dados de forma mais eficaz. Este trabalho propõe um estudo exploratório, multi-caso sobre o ciclo de vida dos dados nas empresas e na literatura, a fim de identificar, nos métodos utilizados atualmente, possíveis gargalos e dificuldades.

Palavras-Chave: *Big Data*. *Data Science*. Descoberta de conhecimento. Ciclo de vida da *Data Science*. Mineração de dados. Métodos.

ABSTRACT

Information and decision support systems are used for analysis, always aligned with the methods of knowledge discovery. With the growth of volume, variety and speed of data, companies are getting more qualified information and restructuring their strategic positions from the data analysis. Since 1989, several methods have been proposed, but nowadays, with the dissemination and challenges of Big Data and the value of the knowledge extracted from the data for the organizations, it was necessary to look for new ways of performing the analysis, thus covering the field of Data Science. Data Science is then a study area capable of helping companies deal with this new system of data analysis, allowing the extraction of knowledge from the data more effectively. This paper proposes an exploratory, multi-case study on the data life cycle in companies and literature, in order to identify possible bottlenecks and difficulties in the current methods.

Keywords*: Big Data. Data Science. Knowledge discovery. Data Science lifecycle. Data Mining. Methods.*

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

A humanidade vem armazenando e analisando dados desde muito tempo, porém, foi com o desenvolvimento da Tecnologia da Informação (TI) que esses dados passaram a ser produzidos e armazenados de forma mais constante, com maior qualidade e em um grande volume. Não só a evolução da TI, mas também o aumento do uso da *Internet* foi um dos grandes potencializadores no acúmulo e monitoramento de dados e na transformação deles em conhecimento (ROSSETTI; MORALES, 2007).

Com o auxílio de sistemas qualificados e uma maior capacidade de processamento, as informações puderam ser analisadas e transformadas com mais rapidez, tornando o processo de tomada de decisão mais preciso e confiável. Neste cenário, as empresas puderam obter melhorias em seus processos, diminuindo custos, descobrindo gargalos e gerando maior valor na entrega de seus produtos.

A velocidade com que os dados começaram a ser acumulados, bem como sua variedade de formatos, fez com que as tarefas de armazenar, processar e por fim transformá-los em informação ficassem mais complexas (NIST, 2015). Chamamos de *Big Data* essa gama de dados armazenados e de difícil processamento pelos sistemas convencionais (SMITH, 2016).

> "*The sheer volume and variety of Big Data often poses problems for companies because they are unsure of how to put the data to work for them. You see, most businesses understand that there is a large amount of data and they understand that it can create a lot of value, but they are not sure where they are supposed to start*." (SMITH, 2016, p. 12)

A *Data Science* surge, então, como área de estudo responsável em lidar com essa enorme quantidade de dados, objetivando extrair informações e conhecimentos futuros que auxiliem as empresas a entender melhor o seu universo, bem como a se posicionar de forma estratégica perante o mercado. Para facilitar esse processo, as equipes de cientistas de dados utilizam e desenvolvem métodos que servem como guias durante o desenvolvimento de projetos de *Big Data* (NIST, 2015).

No entanto, as empresas e seus talentos ainda não estão completamente habilitados para desenvolver esse processo de maneira eficiente. Provost e Fawcett (2016, p. 1) dizem que "no passado, as empresas podiam contratar equipes de estatísticos, modeladores e analistas para explorar manualmente os conjuntos de dados, mas seu volume e variedade superam muito a capacidade da análise manual".

É preciso trabalhar a *Data Science* em todos os seus aspectos. Trata-se de uma transformação na forma de olhar, processar e gerir um negócio. Dados nunca são apenas dados, dependem de como são concebidos e utilizados, podendo variar entre aqueles que capturam, analisam e extraem conclusões deles (KITCHIN, 2014).

Além disso, os desafios tecnológicos ainda continuam, não apenas pela velocidade em que os dados aumentam de volume ao passar do tempo, mas também pela maior complexidade dos dados armazenados, exigindo cada vez mais dos sistemas (NIST, 2015).

Em busca da compreensão de como as empresas vêm superando esses desafios de análises do *Big Data* para a extração de informações relevantes para o negócio, este trabalho faz a seguinte pergunta: **quais os principais métodos utilizados pelas equipes de *Data Science* para extração do conhecimento a partir de um grande volume de dados?**

1.1 Objetivos

Este trabalho tem por objetivo identificar os métodos mais utilizadas para extração de conhecimento do *Big Data* pelas empresas, como elas se relacionam com os métodos descritos na literatura e identificar possíveis gargalos e/ou dificuldades encontradas durante o ciclo de vida da *Data* Science. Os objetivos específicos, são:

1. Compreender o processo de geração de conhecimento a partir de um grande volume de dados;
2. Identificar os métodos utilizados para descoberta de conhecimento na literatura;
3. Identificar os métodos utilizados para descoberta de conhecimento nas empresas;
4. Identificar dificuldades encontradas no desenvolvimento dos métodos existentes;
5. Estabelecer relações entre os métodos utilizados pelas empresas e os descritos na literatura.

1.2 Justificativa

Nos últimos anos, o aumento da disponibilidade de uma grande quantidade e variedade de dados advindos de diversas fontes diferentes, internas e externas às empresas, consequentemente levou ao aumento do interesse em métodos para extrair informações e conhecimento a partir dos dados (PROVOST; FAWCETT, 2016).

Segundo Provost e Fawcett (2016, p. 2), “muitas empresas têm se diferenciado estrategicamente com a *Data Science*, às vezes, ao ponto de evoluírem para empresas de mineração de dados”. Como, por exemplo, o momento em que as compras começaram a serem feitas *online*, permitindo aos varejistas uma maior compreensão de seus clientes. Um *e-commerce* tem acesso não apenas ao que os clientes compram, mas também ao que mais eles se interessaram, como eles navegaram no site e como foram influenciados pelos anúncios. Em um período muito curto de tempo, essas empresas estavam desenvolvendo algoritmos que conseguiam prever os gostos de seus clientes (BRYNJOLFSSON; MCAFEE, 2012).

Devido à relevância que a *Data Science* vem demonstrando na geração de conhecimento e entendimento do mundo e dos negócios, este trabalho busca a compreensão dos métodos utilizados pelas empresas para gerar valor através da investigação e da análise de uma grande quantidade de dados, como elas se relacionam com os métodos descritas na literatura, e identificar possíveis gargalos e/ou dificuldades encontradas durante o processo de descoberta do conhecimento.

2 REVISÃO TEÓRICA

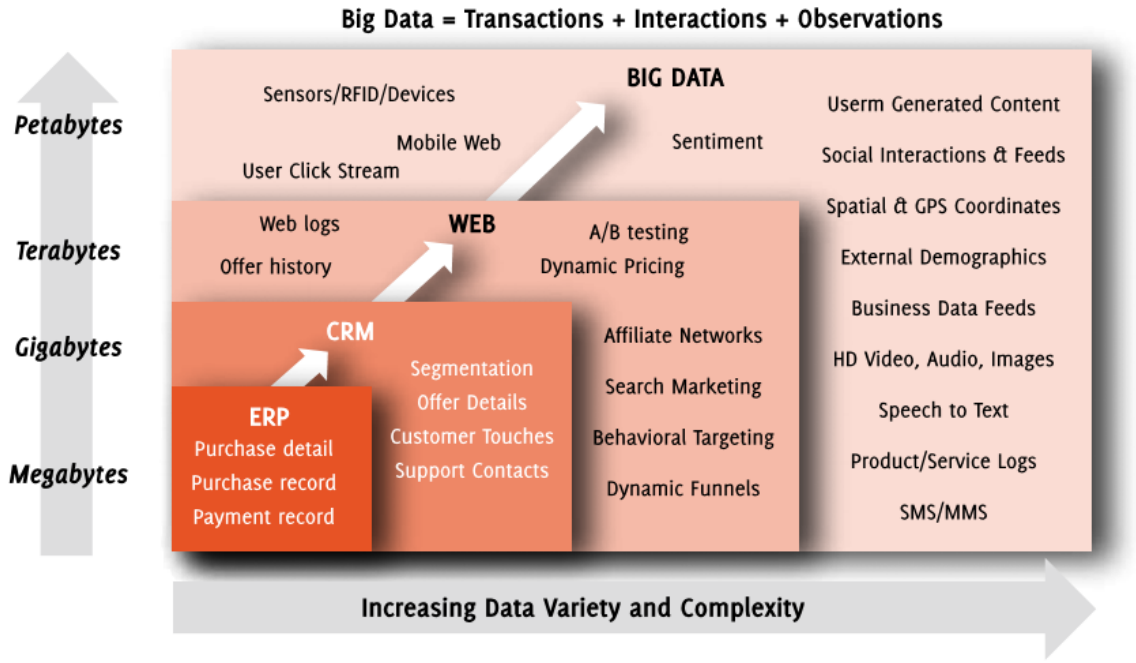
Os dados, juntamente com a informação e o conhecimento, formam um conjunto de elementos importantes para a tomada de decisão nas empresas (ANGELONI, 2003). Para a compreensão da importância dos dados na geração de valor e definição de estratégias, é preciso antes compreender o que são dados, como eles se relacionam com a informação e como o *Big Data* exigiu novas formas de transformar os dados em informação, trazendo consigo uma série de atividades desenvolvidas pela *Data Science*. Este capítulo está dividido em duas seções: *Big Data* e *Data Science*.

2.1 *Big Data*

Para iniciar a compreensão sobre o *Big Data*, uma breve definição sobre o termo que dá vida ao assunto, dados: dados são um conjunto de elementos ou ocorrências em estado bruto, ou seja, não possuem significado quando não vinculados com a realidade (ANGELONI, 2003).

O armazenamento de dados em massa teve seu início com o processo de digitalização (MAURO et al., 2016), porém, foi com o avanço do uso de tecnologias como *Enterprise Resource Planning* (ERP), *Customer Relationship Management* (CRM) e *Web* que o termo *Big Data* e seus desafios começaram a ser foco de discussões e de desenvolvimento empresarial (BLOEM et al., 2012). A Figura 1 ilustra a interação entre os dados acumulados e os sistemas diversos.

Figura 1- Interação entre os dados acumulados em sistemas diversos

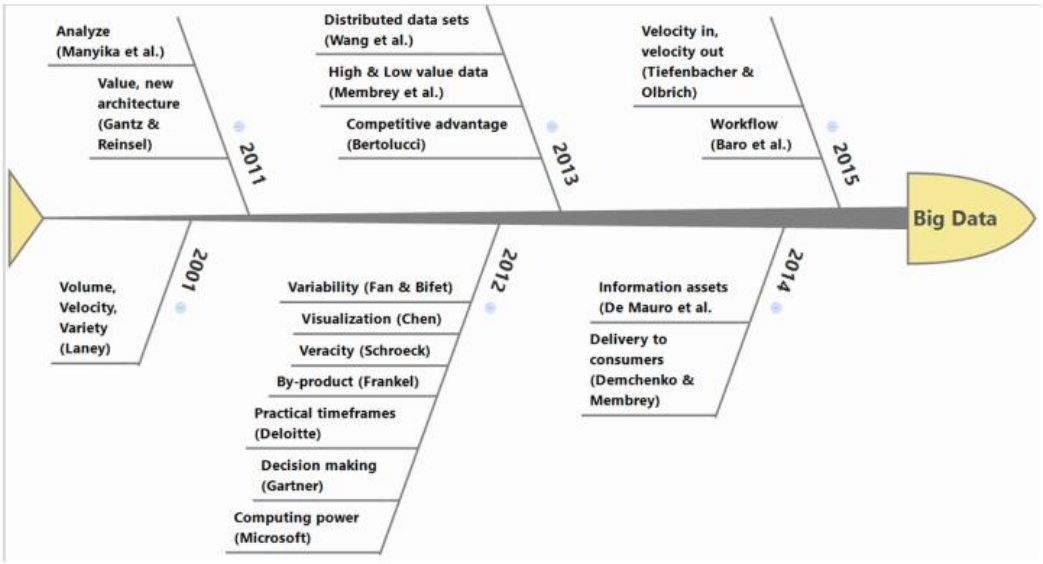


Fonte: Bloem et al. (2012, p. 5)

Em 2001, Laney deu um primeiro indício do que seria enfrentado com o acúmulo dos dados ao passar do tempo. Em seu relatório "*3D Data Management: Controlling Data Volume, Velocity and Variety"*, feito para a empresa META Group, hoje Gartner, ele toma como tema central os dados acumulados por *e-commerces*. Nele, Laney (2001) alerta sobre os desafios do gerenciamento de dados com o crescimento de três magnitudes, conhecidas como 3V's (volume, variedade e velocidade).

Apesar de o conceito de *Big Data* estar mais frequentemente relacionado à definição dos 3V's de Laney, o termo só foi atribuído a essas magnitudes alguns anos depois (MAURO et al., 2016), e continua em constante evolução (PORRAS; YLIJOKI, 2016). Em seus estudos, Porras e Ylijoki (2016) descobriram 17 definições diferentes para *Big Data*, que lhes deram a seguinte perspectiva com relação ao seu avanço (Figura 2).

Figura 2 - Evolução da definição de *Big Data* em forma de espinhaço



Fonte: Porras e Ylijoki (2016, p. 74)

Tanto Mauro, Greco e Grimaldi (2016) quanto Porras e Ylijoki (2016) chegaram à conclusão de que muitas das definições propostas para o termo *Big Data* são logicamente inconsistentes, porém ambos concordam que os 3V's são elementos importantes para o desenvolvimento do *Big Data*. Ainda sobre o conceito, os autores afirmam que a definição do *Big Data* deveria estar relacionada à sua natureza, não incluindo o uso pretendido com a sua manipulação.

Segundo o instituto norte-americano National Institute of Standards and Technology (NIST, 2015, p. 5), "*Big Data consists of extensive datasets - primarily in the characteristics of volume, variety, velocity, and/or variability - that require a scalable architecture for efficient storage, manipulation, and analysis*". Tal definição corrobora com os apontamentos levantados por Mauro, Greco e Grimaldi (2016) e Porras e Ylijoki (2016). O termo variabilidade foi incluído por (FAN; BIFET, 2012) conforme referenciado na Figura 2.

Volume: é a característica mais relacionada ao *Big Data*, sinalizando o tamanho dos dados armazenados em recursos computacionais (NIST, 2015, p. 13). Para dar uma visão da magnitude desse armazenamento, segundo um estudo realizado pela empresa *EMC Digital Universe Study*, de 2013 a 2020, o universo digital crescerá por um fator de 10x, de 4,4 trilhões de *gigabytes* para 44 trilhões (SCHMARZO , 2014).

Variedade: são os diferentes tipos de dados disponíveis para análise. Eles podem ser classificados como: estruturados (armazenados em banco de dados

relacionais), não-estruturados (texto livre) ou semiestruturados (páginas na *Internet*) (ALMEIDA, 2002).

> "*Previously, most of the data in business systems was structured data, where each record was consistently structured and could be described efficiently in a relational model. Records are conceptualized as the rows in a table where data elements are in the cells. Unstructured data types, such as text, image, video, and relationship data, have been increasing in both volume and prominence. (...) The need to analyze unstructured or semi-structured data has been present for many years. However, the Big Data paradigm shift has increased the emphasis on extracting the value from unstructured or relationship data*." (NIST, 2015, p. 12)

Velocidade: é a taxa com que os dados são criados, armazenados, analisados e visualizados. Isso significa uma enorme quantidade de dados sendo processados em um curto período de tempo (NIST, 2015).

Variabilidade: qualquer alteração nos dados ao longo do tempo (NIST, 2015).

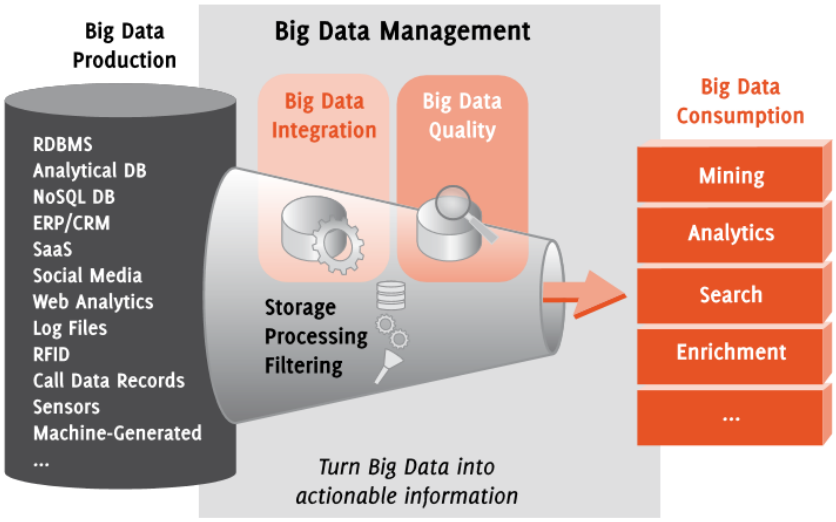
Outras características como: veracidade (fidelidade dos dados), valor (valor da informação gerada para a empresa), volatilidade (mudança na estrutura dos dados ao longo do tempo) e validade (relevância da informação para um determinado momento) foram atribuídas ao conceito de *Big Data* no decorrer do tempo, conforme podemos verificar na Figura 2. No entanto, NIST (2015) defende que essas características se referem à etapa de análise do processo da *Data Science*. São características que afetam a qualidade da informação bem como o conhecimento gerado e entregue para a empresa no processo de extração de valor a partir dos dados.

A preocupação com a veracidade dos dados não é uma novidade: *"Veracity refers to the completeness and accuracy of the data and relates to the vernacular 'garbage-in, garbage-out' description for data quality issues in existence for a long time*." (NIST, 2015, p.17), porém, com a mudança de paradigma na análise dos dados, ainda que o conjunto tenha dados negativos, em uma gama muito grande de dados pode não afetar o resultado.

Ao mesmo tempo em que essa explosão de informações vem criando oportunidades para novas maneiras de combinar e analisar os dados com o objetivo de extrair valor para as organizações, ela também está criando um desafio significativo em relação ao seu tamanho, gestão e análise (NIST, 2015). Esses desafios vêm sendo assimilados pelas empresas, que estão compreendendo a importância de se ter uma nova arquitetura e gestão do *Big Data* projetada para que possam dar prioridade às tarefas de análise e descoberta de conhecimento (DAVENPORT, 2014). A Figura 3

ilustra um esquema desenvolvido por (BLOEM et al., 2012) sobre a produção, gestão e consumo do *Big Data*.

Figura 3 - Gestão do *Big Data*



Fonte: Bloem et al. (2012, p. 13)

O conhecimento gerado pela análise de um grande volume de dados vem transformando os modelos de negócios e a maneira como olhamos para o mundo. No entanto, as características do *Big Data* estão muito mais relacionadas à *Big Science*:

> *"(...) the Big Data concept is most closely related to what we call Big Science. Tere, the Volume, Variety and Velocity aspects, in combination with state-of-the-art hardware and software, are most obviously present, although some people may contest scientific Relevance and Value, certainly in times of crisis*." (BLOEM et al., 2012, p. 23).

Como trazer o *Big Data* da *Big Science* para o *Big Business*? Essa foi a pergunta formulada em "*Creating clarity with Big Data*" e a resposta para ela foi: o coração da resposta é a *Data Science*, a arte de transformar os dados existentes em novos *insights* através dos quais uma organização pode ou vai agir. (BLOEM et al., 2012, p. 24).

2.2 *Data Science*

Observa-se o início da *Data Science* na mesma época em que as primeiras empresas digitais nasceram. Com a quantidade e variedade de dados aos quais elas tinham acesso, foi possível começar um processo de mudanças no atendimento aos seus clientes em cima da análise desses dados. Tais processos acabaram por atingir os modelos de negócio das empresas, iniciando uma nova forma de olhar para as suas estratégias (BRYNJOLFSSON; MCAFEE, 2012).

A Amazon, por exemplo, salva as suas buscas, correlaciona o que você procura com a busca de outros usuários e utiliza esse resultado para criar recomendações mais apropriadas para o seu perfil. Essas recomendações são "produtos dos dados" que ajudam a impulsionar o negócio de varejo mais tradicional da empresa. Eles surgem porque a Amazon entende que um livro não é apenas um livro, e um cliente não é apenas um cliente; os clientes geram uma trilha de "escape de dados" que podem ser minerados e colocados em uso (LOUKIDES, 2010).

Fazendo uma comparação no *Google Trends* (FARIAS, 2017) utilizando as palavras-chave “*Big data*” e “*Data science*”, percebe-se o aumento significativo do interesse pelo assunto por volta do final de 2011 - o intervalo de tempo pesquisado foi de 10 anos (janeiro/2008 a janeiro/2018), conforme gráfico da Figura 4. Um marco considerado significativo para o aumento do interesse por esses termos no ano de 2011 foram os relatórios publicados pela *McKinsey Global Institute* e pela IDC (GANTZ, REINSEL, 2011; MANYIKA et al., 2011).

Figura 4 - Aumento do interesse nas buscas pelos termos "*Big data*" e "*Data science*"

Fonte: *Google Trends* (2018)

A *Data Science* está diretamente relacionada ao *Big Data* pois surge justamente pelas características especiais relacionadas a uma grande quantidade de dados (volume, variedade, velocidade, variabilidade) (DHAR, 2013). Ela se destaca então para ajudar as empresas a lidarem com essa nova sistemática de análise, gerenciamento e extração de informação a partir dos dados (NIST, 2015). "*Data Science is the extraction of actionable knowledgement directly from data through a process of discovery, or hypothesis formulation and hypothesis testing*." (NIST, 2015, p. 5).

Em muitos projetos de *Data Science*, primeiro visualizam-se os dados brutos para que se forme uma hipótese, para então investigar. Como em qualquer ciência experimental, o resultado final pode ser a necessidade de reformular a hipótese. O conceito-chave é que a *Data Science* é uma ciência empírica, que realiza o processo científico diretamente sobre os dados. A hipótese pode ser conduzida por uma necessidade de negócio, ou pode ser a reafirmação de uma necessidade da empresa em termos de uma hipótese técnica (NIST, 2015). O que diferencia a *Data Science* da estatística é a sua abordagem holística (LOUKIDES, 2010).

Por se tratar de um conjunto de técnicas que extraem informações dos dados, a *Data Science* muitas vezes é confundida com o termo *Data Mining*. Em um esforço de desfazer essa ambiguidade, os autores Provost e Fawcett (2016, p. 2) fornecem um esclarecimento: "*Data Science* é um conjunto de princípios fundamentais que norteiam a extração de conhecimento a partir dos dados. *Data Mining* é a extração de

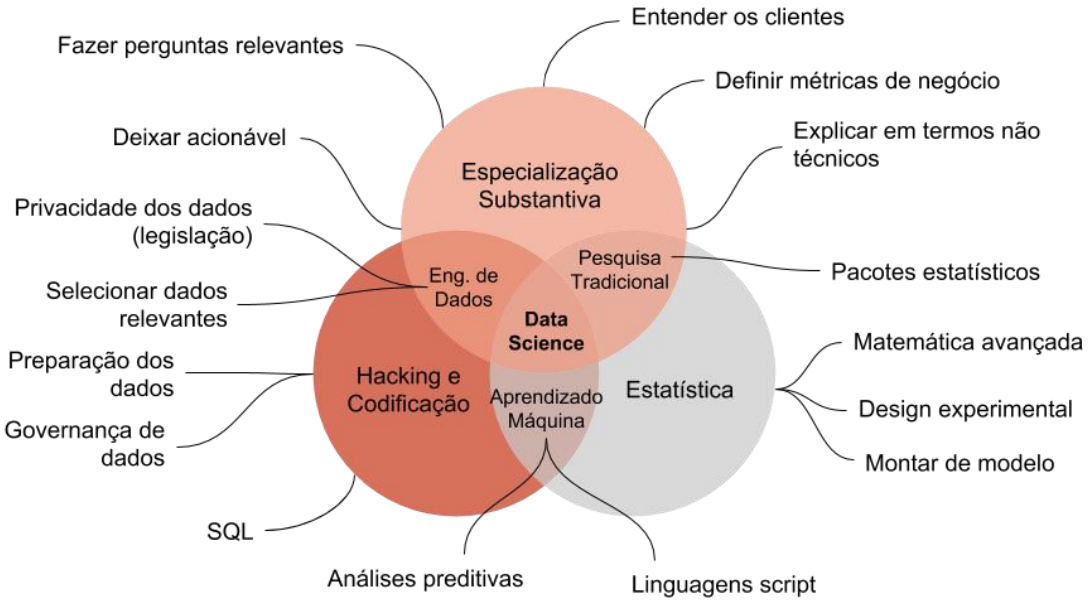
conhecimento a partir deles, por meio de tecnologias que incorporam esses princípios".

Para auxiliar a compreensão das várias áreas de conhecimento que envolvem a *Data Science*, Provost e Fawcett (2016) classificam alguns conceitos fundamentais em três tipos:

1 Conceitos gerais sobre como a *Data Science* pode auxiliar a estratégia da empresa, a organização da equipe e dos processos para projetos que envolvam análise de grandes dados;
2 Conceitos sobre os processos do pensamento analítico que auxiliam a coleta apropriada dos dados e a utilização de métodos adequados. Estes conceitos incluem o processo de *Data Mining*;
3 Conceitos gerais para a extração de conhecimento a partir dos dados.

A equipe de *Data Science* é responsável por determinar o ciclo de vida do processo de descoberta de conhecimento e geração de valor dentro das empresas (NIST, 2015; MICROSOFT, 2017). Por isso a importância desses profissionais, dada a extensa base de conhecimento que envolve a *Data Science* e seus estágios. A seguir, um diagrama de Venn proposto em 2016 pela empresa Gartner (TAYLOR, 2016) que ilustra as áreas de conhecimento que um projeto de dados envolve.

Figura 5 - Diagrama Venn da *Data Science*



Fonte: Taylor (2016) - adaptado pela autora

2.2.1 Ciclo de vida da *Data Science*

A *Data Science*, em todo o ciclo de vida dos dados, incorpora princípios, técnicas e métodos de muitas disciplinas e domínios, incluindo limpeza de dados, gerenciamento de dados, análise, visualização, engenharia. "*The data life cycle is the set of processes in an application that transform raw data into actionable knowledge*." (NIST, 2015, p. 8).

Pode-se pensar o processo da *Data Science* como uma extensão ou variação do método científico O'neil e Schutt (2014, p. 43):

- Faça uma pergunta;
- Faça uma pesquisa de embasamento;
- Construa uma hipótese;
- Teste sua hipótese através de experimentos;
- Analise seus dados e escreva suas conclusões;
- Comunique seus resultados.

Alguns métodos utilizados para o processo de *Data Mining* foram estendidos para a *Data Science* (MAYO, 2016), como as citadas a seguir. Segundo Marbán, Mariscal e Segovia (2009, p. 1) *"CRISP-DM (CRoss-Industry Standard Process for DM) is the most used methodology for developing DM (Data Mining) & KD (Knowledge Discovery) projects. It is actually a 'de facto' standard*". Uma pesquisa com 200 participantes, realizada pela comunidade *KDnuggets*, em 2014, indicou os seguintes métodos como os mais utilizados em projetos de *Data Science*: CRISP-DM (43%); Metodologias próprias (27,5%); SEMMA (8,5%), *KDD Process* (7,5%).

O processo de descoberta de conhecimento a partir dos dados vem evoluindo com o passar dos anos em cima de teorias já pré-estabelecidas. Desde os métodos vinculados à *Data Mining*: "*The Data Science Process and its predecessor CRISP-DM are basically re-workings of the KDD Process.*" (MAYO, 2016), até sua similaridade com o método científico.

A seguir, a descrição dos métodos pesquisados na literatura que abordam o ciclo de vida dos projetos de análise de dados.

2.2.1.1 *KDD Process*

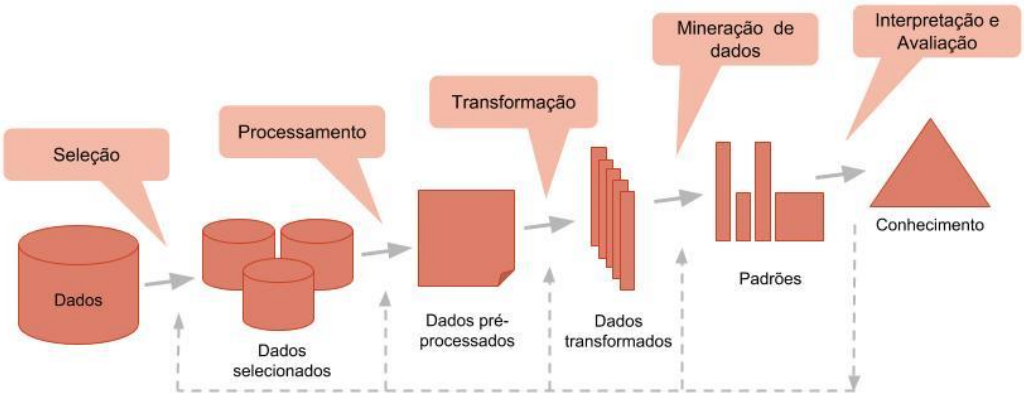
Segundo Fayyad, Piatetsky-Shapiro e Smyth (1996), a frase “*knowledge discovery in databases*” foi cunhada no primeiro *workshop* de KDD, em 1989, por Piatetsky-Shapiro, com o objetivo de enfatizar que o conhecimento é o produto final de uma descoberta baseada em dados.

Em seu artigo “*From Data Mining to Knowledge Discovery in Databases*”, Fayyad, Piatetsky-Shapiro e Smyth (1996) trazem como tema central a diferenciação entre o *KDD Process* e a *Data Mining*. Para os autores, o *KDD Process* se refere ao conjunto de passos e ou processos que visam a descoberta de conhecimento útil a partir dos dados, enquanto que a *Data Mining* seria uma etapa desse processo de descobrimento envolvendo a fase de modelagem dos dados. "*KDD is the nontrivial process of identifying valid, novel, potentially useful, and ultimateandly understandable patterns in data*." (FAYYAD et al.,1996, p. 40).

O termo "processo não trivial" remete a necessidade de execução de várias etapas, de certa forma complexas, para alcançar o objetivo de identificar padrões que sejam "úteis" e de "fácil compreensão" através da análise dos dados. São consideradas cinco etapas do processo KDD (Figura 6).

Ainda aparecem duas características importantes do *KDD Process*, ele é iterativo e interativo. Iterativo porque prevê uma sequência de atividades onde o resultado de uma etapa depende da outra. Interativo porque o analista pode intervir nas atividades. Cada etapa do processo pode ser repetida inúmeras vezes.

Figura 6 - Etapas do *KDD Process*



Fonte: Fayyad et al. (1996, p. 41) – adaptado pela autora

1. Seleção (*Selection*): esta etapa consiste em selecionar um conjunto ou subconjunto de dados que farão parte da análise. As fontes de dados podem ser variadas (planilhas, sistemas gerenciais, *data warehouses*) e possuir dados com formatos diferentes (estruturados, semiestruturados e não-estruturados).
2. Processamento (*Preprocessing*): esta etapa consiste em fazer a verificação da qualidade dos dados armazenados. A base passa por um processo de limpar, corrigir ou remover dados inconsistentes, verificar dados ausentes ou incompletos, identificar anomalias (*outliers*).
3. Transformação (*Transformation*): esta etapa consiste em aplicar técnicas de transformação como: normalização, agregação, criação de novos atributos, redução e sintetização dos dados. Aqui os dados ficam disponíveis agrupados em um mesmo local para a aplicação dos modelos de análise.
4. Mineração de Dados (*Data Mining*): esta etapa consiste em construir modelos ou aplicar técnicas de mineração de dados. Essas técnicas têm por objetivo (1) verificar uma hipótese, (2) descobrir novos padrões de forma autônoma. Além disso, a descoberta pode ser dividida em: preditiva e

descritiva. Esses modelos geralmente são aplicados e refeitos inúmeras vezes dependendo do objetivo do projeto.

5. Interpretação e Avaliação (*Interpretation / Evaluation*): esta etapa consiste em avaliar o desempenho do modelo, aplicando em cima de dados que não foram utilizados na fase de treinamento ou mineração. A validação pode ser feita de diversas formas, algumas delas são: utilizar medidas estatísticas, passar pela avaliação dos profissionais de negócio.

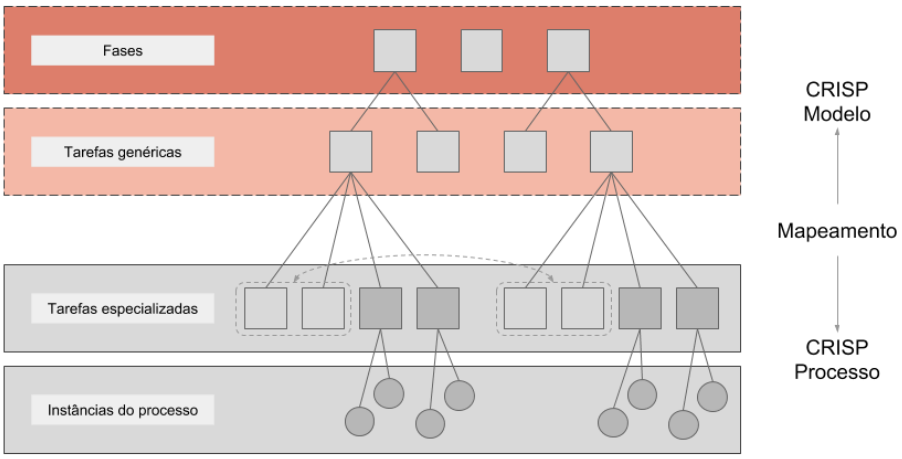
Em seu artigo, Fayyad, Piatetsky-Shapiro e Smyth (1996) ainda citam alguns passos que não estão visíveis na Figura 6. A saber: identificar o objetivo da aplicação do *KDD Process* na perspectiva do cliente; realizar uma análise exploratória; selecionar modelos de hipóteses e, por fim, agir sobre o conhecimento descoberto.

2.2.1.2 CRISP-DM

O CRISP-DM foi concebido em 1996 por um consórcio de três empresas que aplicavam *Data Mining* em seus negócios: Daimler Chrysler, SPSS e NCR4. (CHAPMAN et al., 2000).

A metodologia CRISP-DM é descrita em termos de um modelo de processo hierárquico, consistindo em conjuntos de tarefas descritas em quatro níveis de abstração (do geral ao específico): fase, tarefa genérica, tarefa especializada e instância do processo (Figura 7).

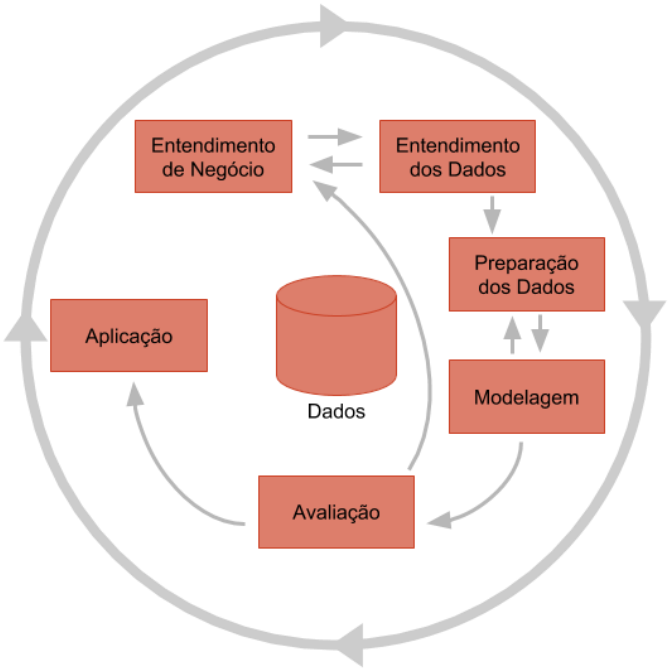
Figura 7 - Quatro níveis da metodologia CRISP-DM



Fonte: Chapman et al. (2000, p. 6) – adaptado pela autora

O modelo CRISP-DM apresenta uma visão geral do ciclo de vida de um projeto de *Data Mining*, contendo as fases e as tarefas relacionadas ao projeto, porém sem um alto nível de identificação dos relacionamentos entre as tarefas.

Figura 8 - Fases do modelo CRISP-DM



Fonte: Chapman et al. (2000, p. 10) – adaptado pela autora

O ciclo de vida apresentado pela metodologia CRISP-DM consiste em seis fases que não precisam ser seguidas rigorosamente, tendo certa flexibilidade para ir e vir retomando as etapas anteriores conforme se faz necessário.

A Figura 8 ilustra com clareza o ciclo de vida de projeto de mineração de dados que assume uma forma contínua, retroalimentando as fases conforme se descobrem novas informações ou aperfeiçoamentos do processo. A seguir, uma breve descrição das etapas:

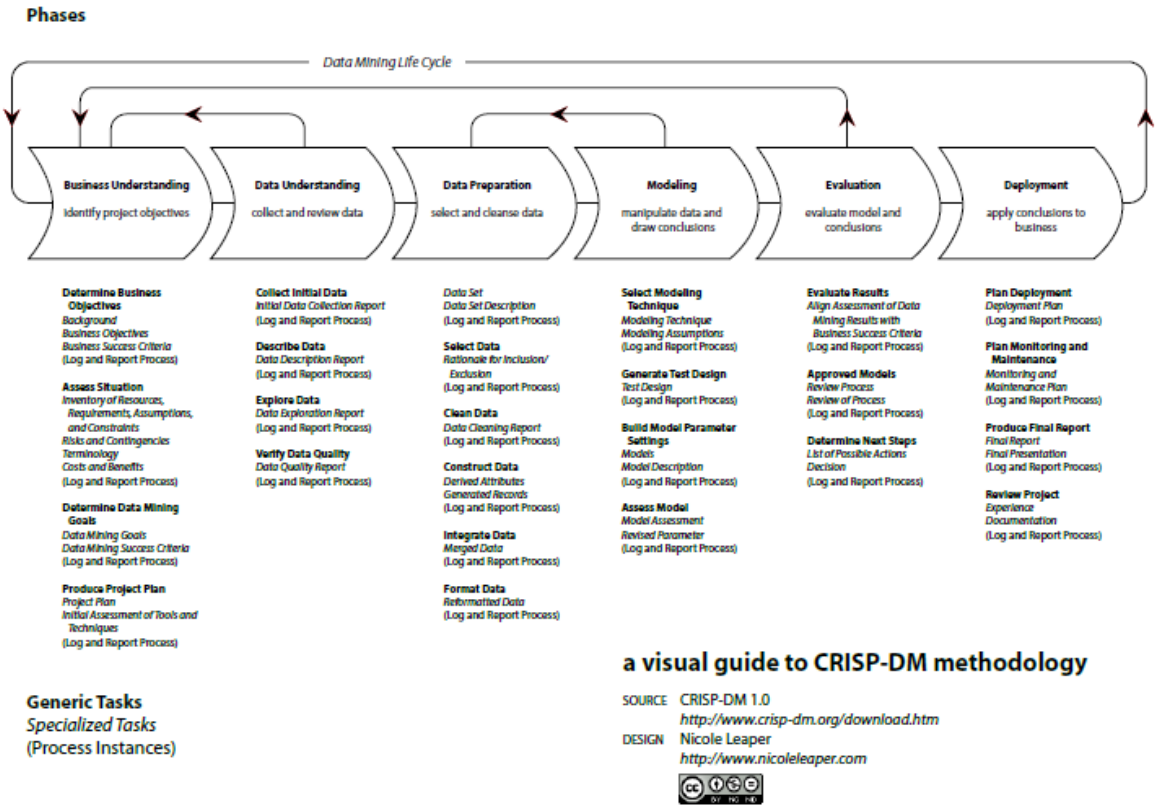
1. Entendimento de Negócio (*Business Understanding*): esta etapa consiste em definir os objetivos do projeto de *Data Mining* com um olhar para os problemas de negócio. É importante salientar que o ciclo de desenvolvimento do projeto só tem efeito se for orientado para resolver um problema de negócio.
2. Entendimento dos Dados (*Data Understanding*): esta etapa consiste desde a captura dos dados até a identificação de problemas relacionados à

qualidade. Nesta fase é também onde se formam hipóteses em cima do que se aprendeu com os dados.

3. Preparação dos Dados (*Data Preparation*): esta etapa consiste em preparar os dados para a modelagem. É a construção de um conjunto de dados obtidos dos dados brutos iniciais, porém que passaram pela limpeza e transformação necessárias para a próxima etapa.
4. Modelagem (*Modeling*): esta etapa consiste em aplicar técnicas de modelagem diferenciadas de acordo com o problema a ser resolvido. Esta etapa pode ser executada várias vezes, inclusive permite voltar a etapas anteriores para ajustar os dados de acordo com os requisitos da técnica a ser utilizada.
5. Avaliação (*Evaluation*): esta etapa consiste em realizar testes com o modelo gerado para validar se atendem às necessidades do negócio. Ainda é possível avaliar se algum objetivo de negócio não tenha sido contemplado com a modelagem proposta.
6. Utilização ou Aplicação (*Deployment*): esta etapa pode ter entregas diferentes, é aqui que a empresa faz uso de toda a análise desenvolvida. Essa análise pode ser desde a apresentação dos resultados da modelagem para a tomada de decisão até a aplicação do modelo em um outro conjunto de dados, incluindo o modelo como parte do processo de geração de informação para tomada de decisão dentro empresa.

A Figura 9 serve como um guia visual o ciclo de vida do CRISP-DM, suas etapas e respectivas tarefas.

Figura 9 - Guia visual do modelo CRISP-DM



Fonte: Gonzáles (2018)

2.2.1.3 SEMMA

SEMMA é um acrônimo para as palavras *Sample, Explore, Modify, Model e Assess*. A SEMMA foi definida pela empresa SAS como uma organização lógica para o uso da ferramenta de mineração de dados desenvolvida por eles, a *SAS Enterprise Miner*.

> "*SEMMA is not a data mining methodology but rather a logical organization of the functional tool set of SAS Enterprise Miner for carrying out the core tasks of data mining. Enterprise Miner can be used as part of any iterative data mining methodology adopted by the client*." SAS (2006)

A seguir, uma breve descrição dos passos lógicos:

1. Amostra (*Sample*): esta etapa consiste na separação de uma amostra significativa o suficiente para extrair a informação necessária em cima da análise destes dados. As amostras podem ser usadas em momentos diferentes do processo: treinamento dos dados, validação dos dados e realização de testes.

2. Explorar (*Explore*): esta etapa consiste na em buscar tendências e anomalias nos dados através de recursos gráficos e estatísticos.
3. Modificar (*Modify*): esta etapa consiste na transformação e preparação dos dados para aplicação dos modelos de extração de conhecimento. Incluindo tarefas como limpeza, agrupamentos, transformação de variáveis etc.
4. Modelar (*Model*): esta etapa consiste em aplicar técnicas de modelagem em mineração de dados. Cada modelo tem um propósito e deverá ser definido de acordo com as necessidades do problema e dos dados disponíveis para a análise.
5. Avaliar (*Assess*): esta etapa consiste em fazer a validação dos resultados obtidos através da aplicação do modelo proposto na etapa anterior. O modelo deverá ser avaliado por sua utilidade, confiabilidade e desempenho. Uma das formas de validar é aplicando o modelo em um conjunto de dados diferente do que foi utilizado na etapa de modelagem para verificar se o algoritmo chega de forma assertiva nas mesmas definições.

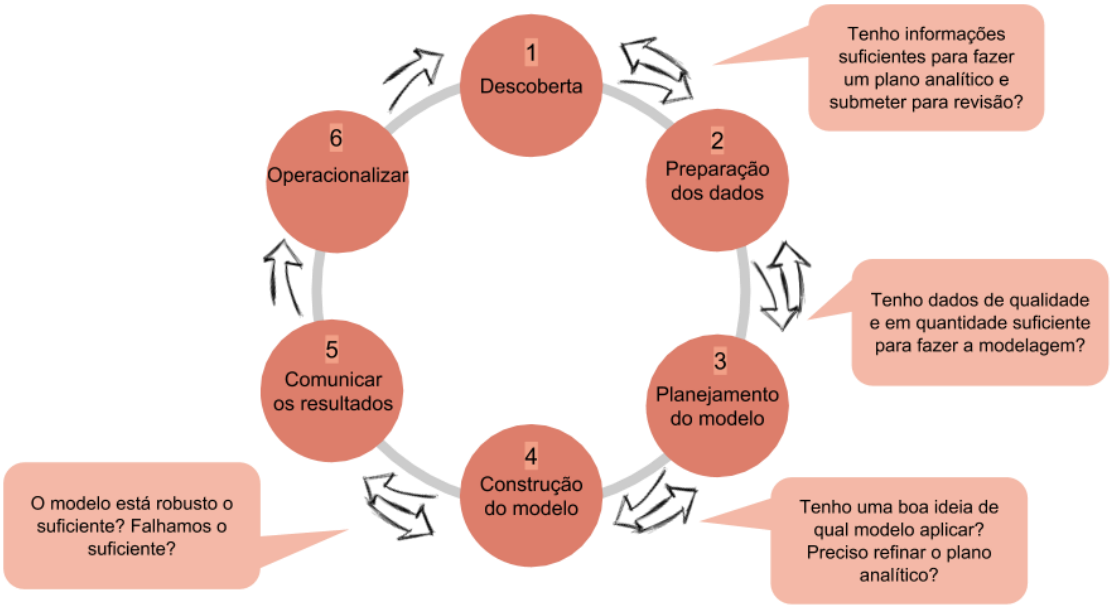
2.2.1.4 *Data Analytics Lifecycle*

O *Data Analytics Lifecycle* foi originalmente desenvolvido para o curso *Data Science & Big Data Analytics* da EMC por David Dietrich, e lançado no início de 2012 (DIETRICH, 2013).

O *Data Analytics Lifecycle* foi designado a resolver problemas relacionados a projetos que envolvam *Big Data* e *Data Science*. Seu ciclo de vida possui seis fases sendo que o projeto pode ser desenvolvido de maneira que muitas destas fases possam ocorrer ao mesmo tempo. Além disso, as etapas podem se movimentar para frente ou para trás, retratando de forma realista a iteratividade de um projeto, onde seus integrantes avançam de acordo com o aprendizado.

O ciclo de vida tem por objetivo definir práticas recomendadas para o processo de análise desde a descoberta até a conclusão do projeto. Vários dos processos que foram consultados incluem: “método científico; CRISP-DM; DELTA *framework* de Tom Davenport; Abordagem da *Applied Information Economics* (AIE) de Doug Hubbard; "*MAD Skills*" por Cohen et al.” (DIETRICH et al., 2015).

Figura 10 - *Data Analytics Lifecycle*



Fonte: Dietrich et al. (2015, p. 29) – adaptado pela autora

1. Descoberta (*Discovery*): esta etapa consiste em aprender sobre o domínio do negócio, analisando o histórico da organização em análise de dados. Deve-se também avaliar os recursos disponíveis pela organização: pessoas, tecnologia, tempo e dados. É nesta fase também que se formulam as hipóteses sobre os problemas de negócio que deverão ser desenvolvidas nas próximas etapas.
2. Preparação dos Dados (*Data Preparation*): esta etapa consiste em extrair os dados de um sistema-fonte, converter em um formato que possa ser analisado e armazenar em um armazém ou outro sistema. Nesta etapa aplicam-se métodos geralmente utilizadas quando os dados são provindos de fontes diferentes: o ETL, um tipo de *data integration* em três etapas (*extract, transform e load*) que combina dados de diversas fontes e o ELT (*extract, load, transform*), uma abordagem alternativa utilizada para aprimorar a performance. Quando as duas técnicas são utilizadas o processo também é conhecimento como ETLT (SAS). Esta é uma fase de familiarização com os dados onde é possível tomar medidas para condicionar os dados.
3. Planejamento do Modelo (*Model Planning*): esta etapa consiste em determinar os modelos e técnicas a serem aplicadas. É uma fase de exploração dos dados,

compreensão das relações entre as variáveis e seleção das que melhor atendem o modelo.

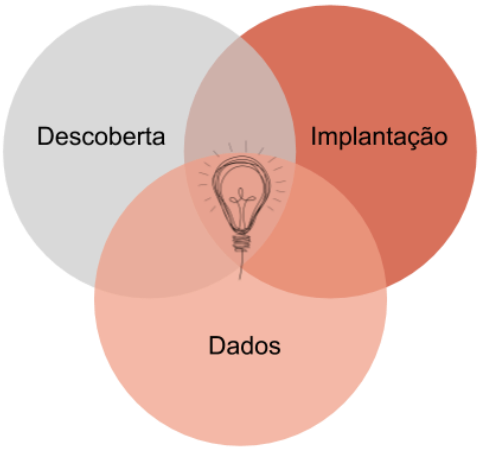
4. Construção do Modelo (*Model Building*): esta etapa consiste em executar o planejamento da fase anterior em cima de uma base de dados menor e selecionada para a realização de testes e treinamento dos modelos. Nesta etapa também são analisadas as ferramentas existentes para processamento dos modelos a fim de verificar se atendem os requisitos necessários, por exemplo: processamento paralelo.
5. Comunicar os Resultados (*Communicate Results*): esta etapa consiste em identificar os principais resultados, aferindo com o os objetivos de negócio levantados na etapa 1. Esta fase pode ser identificada como um sucesso ou um fracasso de acordo com os resultados obtidos. Os resultados devem ser resumidos e apresentados para as partes interessadas do projeto.
6. Operacionalizar (*Operationalize*): esta etapa consiste nas entregas finais do projeto que podem ser relatórios, algoritmos, instruções e documentos técnicos. Também pode ser executado um projeto piloto para implementar os modelos em um ambiente de produção.

2.2.1.5 *SAS Analytical Life Cycle*

A empresa SAS, além da já conhecida SEMMA, ainda apresenta o *SAS Analytical Life Cycle*, um processo iterativo de descoberta a partir dos dados que aplica novas informações no intuito de melhorar de forma contínua os modelos preditivos e seus resultados. (SAS, 2016).

Este processo foi criado pensando em um ambiente que possa ajudar a organização a lidar com todos os dados, modelos e decisões que precisam ser desenvolvidos em uma escala crescente. Onde:

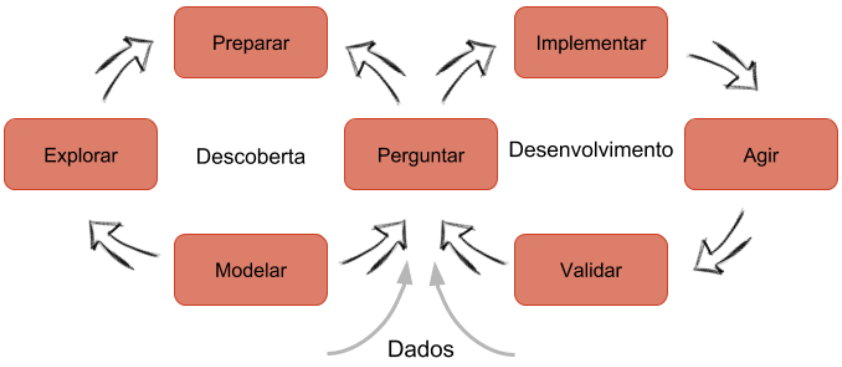
Figura 11 - Combinação integrada de dados, descoberta e implantação



Fonte: SAS (2016, p. 1) – adaptado pela autora

- Dados: são a base para a tomada de decisões.
- Descoberta: é o processo de identificação de novos *insights* nos dados.
- Implantação: é o processo de usar esses *insights* recém encontrados para impulsionar ações aprimoradas.

Figura 12 - *SAS Analytical Life Cycle*



Fonte: SAS (2016, p. 3) – adaptado pela autora

1. Perguntar (*Ask*): esta etapa consiste em fazer uma pergunta de negócio. A descoberta do processo é impulsionada a partir das perguntas de negócio. Este passo é focado em explorar o que você precisa saber, e como você pode aplicar a análise em seus dados para resolver um problema ou melhorar um processo.
2. Preparar (*Prepare*): esta etapa consiste em preparar os dados para serem analisados. Os dados podem vir de fontes diferentes e em formatos diferentes, para a realização da análise é preciso que os dados estejam em um mesmo local e prontos para serem analisados, ou seja, que as diferentes fontes de dados possam ter seus dados relacionados de alguma forma.
3. Explorar (*Explore*): esta etapa consiste em explorar os dados disponíveis e validar a hipótese criada no início do projeto desenvolvendo ideias de análises e testes. Ao realizar estes testes pode-se perceber a necessidade de buscar outros formatos de dados na base ou reformular a hipótese, por exemplo, retomando o ciclo de análise nas etapas anteriores.
4. Modelar (*Model*): esta etapa consiste em aplicar algoritmos de modelagem analítica que tenham capacidade de responder o problema de negócio proposto.
5. Implementar (*Implement*): esta etapa consiste utilizar os modelos propostos para gerar a informação necessária para resolução do problema proposto, seja através da automatização e integração do modelo ou através da aplicação direta e uso da análise em uma situação específica.
6. Agir (*Act*): esta etapa consiste em tomar decisões em cima da análise realizada. Elas podem ser estratégicas, ou seja, realizadas por um humano (SAS, 2016) ou transformadas em decisões operacionais que devem ser automatizadas e, por isso, não exigem intervenção humana, como em um sistema de recomendação das melhores ofertas.
7. Avaliar (*Evaluate*): esta etapa consiste no monitoramento contínuo dos resultados obtidos pela aplicação dos modelos analíticos, garantindo assim que eles continuem trazendo o resultado esperado.

Esses são os passos necessários definidos pela SAS para o ciclo de vida analítico, eles ainda trazem um oitavo passo, que tem por objetivo retomar o processo desde o princípio.

8. Pergunte de novo (*Ask again*): dada a característica mutante dos dados, o processo analítico nunca termina, com o tempo será necessária uma reavaliação dos dados obtidos e, por consequência, dos modelos gerados para que tenham a capacidade de atender ou a novas perguntas de negócio ou de maneira mais eficiente um problema já definido.

2.2.2 Comparativo dos métodos da literatura

Em busca de compreender as diferenças entre os métodos analisados, foi criada uma tabela comparativa, baseada nas etapas dos métodos descritos (Tabela 1). Para tal, a autora assume que as etapas dos métodos são equivalentes e por isso podem ser comparados.

Tabela 1 - Comparativo entre os métodos

<table>
<tr><th>Métodos</th><th>KDD Process (KDD-P)</th><th>CRISP-DM</th><th>SEMMA</th><th>Data Analitics Lifecycle (DAL)</th><th>SAS Analitical Life Cycle (SAS-ALC)</th></tr>
<tr><td>Nº etapas</td><td>5</td><td>6</td><td>5</td><td>6</td><td>7</td></tr>
<tr><td rowspan="11">Etapas</td><td></td><td>Entendimento de negócio</td><td></td><td>Descoberta</td><td>Perguntar</td></tr>
<tr><td>Seleção</td><td rowspan="2">Entendimento dos dados</td><td>Amostra</td><td rowspan="3">Preparação dos dados</td><td rowspan="3">Preparar</td></tr>
<tr><td>Processamento</td><td>Explorar</td></tr>
<tr><td>Transformação</td><td>Preparação dos dados</td><td>Modificar</td></tr>
<tr><td rowspan="2">Mineração de dados</td><td rowspan="2">Modelagem</td><td rowspan="2">Modelar</td><td>Planejamento do modelo</td><td>Explorar</td></tr>
<tr><td>Construindo o modelo</td><td>Modelar</td></tr>
<tr><td>Interpretação e avaliação</td><td>Avaliação</td><td>Avaliar</td><td>Comunicar resultados</td><td></td></tr>
<tr><td></td><td>Utilização e aplicação</td><td></td><td>Operacionalizar</td><td>Implementar</td></tr>
<tr><td></td><td></td><td></td><td></td><td>Agir</td></tr>
<tr><td></td><td></td><td></td><td></td><td>Avaliar</td></tr>
</table>

Fonte: elaborado pela autora

CRISP-DM, DAL e SAS-ALC possuem uma etapa inicial focada na compreensão do negócio, já o KDD-P, ainda que Fayyad, Piatetsky-Shapiro e Smyth (1996) falem sobre uma etapa inicial de compreensão de negócio, ela não está evidenciada no processo como um todo. KDD-P e SEMMA podem ser diretamente comparados com as três seguintes etapas (seleção, processamento e transformação),

já o CRISP-DM de acordo com as descrições, agrupa na etapa (entendimento dos dados) duas das etapas do KDD-P e SEMMA, assim como DAL e SAS-ALC agrupam as três etapas correspondentes.

Seguindo com o comparativo, as próximas etapas são relacionadas a mineração de dados, onde KDD-P, CRISP-DM e SEMMA são correspondentes. No que diz respeito a DAL e SAS-ALC, essa etapa foi dividida em duas partes (explorar e modelar), esse é um ponto importante na diferenciação dos métodos na literatura, a etapa de exploração/planejamento é uma fase anterior a criação dos modelos de análise e tem por objetivo compreender a base de dados, muitas vezes aplicando técnicas de aprendizado de máquina. Segundo Witten et al. (2016) o aprendizado de máquina fornece a base técnica da mineração de dados, é usado para extrair informações dos dados brutos em bancos de dados.

A parte de validação dos resultados está presente em quase todos os métodos, com exceção do SAS-ALC que possui a etapa de avaliar após a implementação do modelo. Outro ponto importante de diferenciação é a característica de ir e vir entre as etapas evidenciada nos métodos KDD-P, CRISP-DM, DAL e SAS-ALC, porém, a etapa de validação e retomada do modelo fica mais clara nos métodos KDD-P e DAL.

Por fim e não menos importante a implementação do projeto, que coloca a prova todo o processo desenvolvido, explicitamente presentes nos métodos CRISP-DM, DAL e SAS-ALC. Ainda temos as etapas (agir e avaliar) da SAS-ALC que não encontram equivalência com os outros métodos.

3 MÉTODO

Para o desenvolvimento deste trabalho realizou-se uma pesquisa qualitativa de caráter exploratória, buscando compreender a temática que envolve o problema proposto e os métodos descritos na literatura.

> “Este tipo de pesquisa tem como objetivo proporcionar maior familiaridade com o problema, com vistas a torná-lo mais explícito ou a construir hipóteses. A grande maioria dessas pesquisas envolve: (a) levantamento bibliográfico; (b) entrevistas com pessoas que tiveram experiências práticas com o problema pesquisado; e (c) análise de exemplos que estimulem a compreensão”. (GIL, 2007).

Em via de atender a problemática levantada por este trabalho, a descoberta dos principais métodos utilizados pelas equipes de *Data Science*, foi utilizado o método estudo de caso coletivo (multi-caso).

> “O estudo de caso único envolve a estratégia de pesquisa aplicada à compreensão de várias dimensões do fenômeno com foco em um caso singular como espaço amostral enquanto o estudo de caso coletivo (multi-casos) envolve o estudo das mesmas dimensões do fenômeno em mais de um caso simultaneamente com objetivo posterior de comparação de resultados.” (BARBOSA, 2008, p.3)

A primeira fase do projeto tratou-se de um levantamento bibliográfico, descrição e comparação dos métodos encontrados. As buscas pelos artigos foram feitas com as seguintes palavras-chave: “*knowledge discovery process*”; "*Data Science process*"; "*Data Science methodology*”; "*Data Science analytics*”; “*Big Data analytics*”. Não foi utilizado um processo exaustivo de busca.

Os dados foram coletados através de um questionário *online* e entrevistas em profundidade. O questionário foi utilizado como contextualização e complemento de análise e teve como objetivo compreender se as empresas, de forma geral, realizam algum tipo de análise de dados. Como orientação para os respondentes, foi utilizado um dos métodos da literatura, o *KDD Process*, por ser considerado como um dos mais antigos, cunhado em 1989 no primeiro congresso do KDD (FAYYAD et al., 1996). A estrutura das perguntas foi de caráter: abertas, fechadas e de múltipla escolha (Anexo A).

Para as entrevistas foi criado um roteiro com perguntas semiestruturadas (Anexo B) e escolhido o método mais difundido entre as empresas (MARBÁN, MARISCAL e SEGOVIA, 2009), o CRISP-DM, como forma de orientar as perguntas e as informações que não poderiam faltar durante a realização das entrevistas. Este

método também foi utilizado como base para o desenvolvimento do *Data Analytics Lifecycle* (DIETRICH et al., 2015), escolhido como base para a comparação dos resultados neste trabalho.

O questionário *online* foi compartilhado em mais de vinte grupos com foco em análise de dados, mineração de dados, aprendizado de máquina, inteligência artificial etc. As redes sociais onde o questionário *online* foi compartilhado foram as seguintes: *Facebook*, *LinkedIn*, *WhatsApp*, *Slack*, *Google+*. O formulário *online* ficou aberto pelo período de um mês.

Para as entrevistas em profundidade foram convidadas oito empresas da região de Porto Alegre e Grande Porto Alegre que trabalham com projetos de análise de dados. Das oito empresas convidadas, sete aceitaram participar da pesquisa. Os convites para participação foram enviados por e-mail direcionados aos responsáveis das empresas ou pelas áreas focadas em análise de dados. As entrevistas ocorreram no local indicado pelos gestores das empresas, durante o mês de setembro, com duração de 30 min a 60 min.

A análise foi feita através do método indutivo, que:

> “Considera que o conhecimento é fundamentado na experiência, não levando em conta princípios preestabelecidos. No raciocínio indutivo a generalização deriva de observações de casos da realidade concreta. As constatações particulares levam à elaboração de generalizações.” (SILVA; MENEZES, 2005, p. 26)

Desta forma, os processos das empresas foram descritos e subdivididos de acordo com a compreensão da autora. Também foi feita uma classificação das empresas entrevistadas levando em conta a forma de desenvolvimento dos projetos de análise de dados. O comparativo entre os métodos da literatura foi utilizado como forma de qualificar o método escolhido para as comparações e análises com os processos das empresas. Mais detalhes sobre como as inferências foram feitas se encontram nas próximas seções deste trabalho.

4 APRESENTAÇÃO DA PESQUISA E ANÁLISE DOS RESULTADOS

A apresentação da pesquisa se dá através da descrição dos métodos das empresas entrevistadas seguido da análise dos resultados obtidos através do questionário *online* e do comparativo entre os métodos das empresas.

Como um dos resultados deste trabalho, também é apresentada uma proposta de Canvas *Analytical* oriundo de uma necessidade evidenciada no desenvolver deste projeto.

4.1 Descrição dos métodos das empresas

Através da pesquisa qualitativa, buscou-se compreender não apenas o processo de análise dos dados, mas também todo o processo de relacionamento da empresa com o cliente, desde a fase de prospecção até a fase de entrega e acompanhamento. Para Godoy (1995, p. 62) “A palavra escrita ocupa lugar de destaque nessa abordagem, desempenhando um papel fundamental tanto no processo de obtenção dos dados quanto na disseminação dos resultados”.

Tabela 2 - Apresentação do perfil das empresas entrevistadas

Perfil das empresas entrevistadas						
Empresa	**Posicionamento da empresa**	**Perfil do profissional entrevistado**	**Formação**	**Cargo**	**Tempo de atuação com análise**	**Prospecção dos projetos de dados**
E1	Consultoria	Homem 40 anos	Licenciatura em Matemática Doutorado em Matemática foco Modelagem Matemática	Consultor	12 anos	Palestras; Mentorias; Congressos; Indicações
E2	Laboratório de *Data Science*	Homem 23 anos	Téc. em Tecnologia da Informação; Formando em Eng. de Produção	Data Scientist	5 anos	Palestras; Mentorias; Congressos; Indicações
E3	*Software* para análise de dados e tomada de decisão	Homem 36 anos	Administração Doutorado em Administração com foco em Sistemas	Diretor	17 anos	Eventos de negócios como patrocinador; Email marketing; Visitas técnicas; LinkedIn; Indicações

E4	*Software* para análise de dados e tomada de decisão	Homem 39 anos	Engenharia Civil Mestrado em Engenharia de Produção	Coordenador da Área de Inteligência	12 anos	Oferecem novas funcionalidades que agregam inteligência em dados dentro do produto para os clientes que já utilizam o *software*
E5	Consultoria	Mulher 30 anos	Formanda em Estatística e Administração de Empresas	CEO	13 anos	Palestras; Mentorias; Indicações
E6	Consultoria	Homem 27 anos	Ciência da Computação	Gestor de Projeto / Sócio	5 anos	Palestras; Indicações
E7	*Software* e Consultoria	Homem 47 anos	Administração MBA Gerência de Projetos e E-business	Diretor de Negócios	20 anos	Eventos; Visitações a possíveis clientes; Cursos

Fonte: elaborado pela autora

A Tabela 2 é um resumo do perfil dos profissionais e o posicionamento das empresas que participaram da entrevista. A descrição dos processos das empresas será dividida por etapas, conforme o entendimento da autora em relação aos relatos dos profissionais.

4.1.1 Caso E1

Modo de trabalho: consultoria e desenvolvimento de análises e algoritmos.

Entregas realizadas: algoritmo do modelo desenvolvido para a análise; relatórios estáticos.

Uso frequente das análises de dados pelos clientes: indústria: otimização de processos e custos, identificação de falhas; varejo: identificação de anomalias, controle de estoque, previsão de vendas, identificação do perfil consumidor; banco: processos de recursos humanos.

Etapa 1: entendo o problema e definindo o projeto

A empresa E1 realiza reuniões com a pessoa que entrou em contato e que tem uma necessidade relacionada a análise de dados dentro do cliente. Essas reuniões têm por objetivo compreender o problema e as dores do cliente, bem como compreender como é o funcionamento da empresa e quem são as pessoas responsáveis pela tomada de decisão. O consultor, então, é convidado para apresentar seus projetos

diretamente ao gestor do setor que possui a necessidade de extrair informação dos dados com o objetivo de demonstrar a capacidade de entrega dos projetos de análise. Muitas vezes durante esta reunião surgem outros problemas que são *linkados* pelos gestores como prioritários e o foco do projeto pode vir a mudar.

Vencida e etapa de compreensão do valor da aplicação do projeto para o cliente, o consultor inicia uma exploração dentro da empresa, conhecendo as áreas e pessoas envolvidas nos processos, muitas vezes olhando para setores e processos anteriores e posteriores ao setor que será analisado. Esta fase tem por objetivo buscar *insights* para compreender quais dados, além dos já disponíveis pelo setor, podem compor o projeto, podendo durar de três a quatro encontros, conforme a complexidade do problema e número de áreas envolvidas. Durante os encontros são realizadas entrevistas em profundidade com as pessoas que sofrem do problema para que os consultores tenham a compreensão do todo. Como forma de validar a compreensão do problema o consultor apresenta aos *stakeholders* do projeto uma sentença, ou frase, que resuma o problema a ser resolvido, é necessário que todos estejam de acordo e compreendam o que foi. Desta sentença inicia-se a fase de elaboração do escopo do projeto. No geral os projetos são de curta duração e tendo como principal objetivo entregar valor em informações úteis ao cliente. Quando necessário, o consultor convida parceiros com especialidades distintas para compor a compreensão do problema e quais métodos podem ser usados para resolvê-los.

Etapa 2: acessando os dados

A empresa E1 normalmente trabalha com um servidor exclusivo dentro do cliente com direitos de acesso remoto, este acesso permite o consultor trabalhar de forma mais independente, sem precisar interferir na TI da empresa com frequência durante o processo de análise. O cliente fica responsável por disponibilizar os dados solicitados pelo consultor dentro deste servidor, preferencialmente no formato (csv). Os dados são disponibilizados na sua forma bruta e de preferência com a base completa, ou seja, não amostral. Os dados podem ser provindos de fontes diferentes e setores diferentes, conforme os resultados esperados e o problema descrito.

Etapa 3: preparando os dados

Nesta fase o consultor explora os dados disponíveis, os formatos, classes etc. Se a TI do cliente tem possibilidade de entregar os dados nos formatos adequados o consultor deixa essa fase com o cliente, caso contrário, ele mesmo realiza a limpeza e organização dos dados. Aqui também é o momento para compreender os dados, o

que significam e para que são utilizados. É feito também uma pré-análise dos dados e modelos para ver se atendem as necessidades do problema.

Etapa 4: criando e validando um modelo

Nesta etapa o consultor cria os modelos matemáticos e estatísticos para a resolução do problema. Aqui o consultor determinar qual é o melhor modelo a ser aplicado para o problema (determinístico ou probabilístico). Dependendo do problema pode ser necessário uma ampliação da base de dados com a criação de métricas, neste caso o consultor volta a falar com o time de negócios compreender quais métricas tem capacidade de medir as premissas do problema. Além disso, algumas métricas são criadas para analisar possíveis relações causais. Também são aplicados algoritmos de *machine learning* para possíveis descobertas em cima dos dados, como a *clusterização* e identificação de padrões de comportamentos nos grupos criados. É feita uma análise qualitativa em cima dos resultados pelo consultor, criando uma história para explicar a relação e implicações entre as variáveis para ver se faz sentido. A empresa E1 leva as histórias das variáveis para os stakeholders para fazer a validação dos resultados. Muitas das histórias são problemas conhecidos e outras acabam trazendo relações causais que fazem sentido para as pessoas de negócio, porém que não faziam parte do conhecimento adquirido pelo time.

Etapa 5: implementando a solução

Após a validação o modelo vai sendo aplicado aos poucos e revalidando em todas as etapas, podendo alguns serem adaptados por alguma diferença identificada nas bases. Analisa-se o impacto da implantação da solução nos processos da empresa, pois podem interferir em processos de setores diferentes, por exemplo.

Maior dificuldade encontrada durante o desenvolvimento dos projetos: modelagem do problema; resolução do problema com um modelo que atenda uma quantidade muito grande de dados.

Tecnologias utilizadas para o desenvolvimento do projeto: R.

4.1.2 Caso E2

Modo de trabalho: consultoria, desenvolvimento de análises, transformação da cultura empresarial para uma cultura *data driven*, apoio no desenvolvimento de times de *Data Science*.

Entregas realizadas: desenvolvimento de times, processos, *dashboards*, modelos de tomada de decisão, sistemas preditivos para aplicação.

Uso frequente das análises de dados pelos clientes: melhoria em processos, predição, qualificação do time de *analytics*, estruturação de uma área de dados.

Etapa 1: entendo o problema e definindo o projeto

Projeto inicia com um evento estilo *hackathon* "uma maratona de programação onde pessoas se reúnem para resolver problemas de tecnologia em poucos dias" (Significado de Hackathon, 2012) onde os profissionais da empresa E2 e os tomadores de decisão do cliente são cocriadores do projeto de análise de dados. Durante o *hackathon* são explorados os tipos de decisão tomadas pelos gestores, quais as informações relevantes para a tomada de decisão e de onde elas vêm, onde estão armazenadas. Durante o evento o time determina alguns produtos possíveis de serem desenvolvidos pensando no cenário montado. O cliente fica então responsável por definir qual dos produtos será desenvolvido e entregue pela empresa E2. Os projetos podem variar de seis a dois anos de duração.

Etapa 2: acessando os dados

Nesta fase a empresa E2 realiza reuniões com o time de negócios e do TI do cliente e faz o levantamento de todos os dados disponíveis que envolvam o escopo do problema a ser resolvido. Normalmente trabalham com dados amostrais.

Etapa 3: preparando os dados

Esta etapa é realizada junto ao cliente, onde a empresa E2 auxilia a TI do cliente no processo de ajuste dos dados. Como a empresa E2 tem por objetivo que o cliente transforme sua organização e seus processos para que tenham a capacidade de realizar as análises de forma constante, é necessário que o cliente tenha dentro de sua estrutura os dados prontos para a análise, ou seja, o cliente é responsável por disponibilizar um ambiente onde os dados necessários para o projeto sejam integrados, limpos e transformados para aplicação dos modelos analíticos.

Etapa 4: criando e validando um modelo

Nesta etapa o time mapeia e desenvolve modelos que podem atender as necessidades do projeto. Durante a etapa de modelagem o time realiza a validação dos resultados em conjunto com o time de negócios do cliente. Cada aplicação de modelagem é avaliada de acordo com o entendimento do negócio podendo mudar as variáveis escolhidas para a modelagem ou não.

Etapa 5: desenvolvimento e implementação

É entregue para o cliente um plano de implementação contendo todas as etapas necessárias para a aplicação da modelagem e dos resultados que podem ser relatórios, *dashboards* entre outros. O planejamento atende aos três pilares de um projeto *data driven* da empresa E2: negócio, ciência, tecnologia. Além do plano também podem ser entregues: o modelo analítico desenvolvido; relatórios; *dashboards*; treinamentos.

Etapa 6: acompanhamento

O acompanhamento acontece com clientes que possuem projetos de longa duração e tendem a continuar o desenvolvimento em outras etapas.

Maior dificuldade encontrada durante o desenvolvimento dos projetos: processo de passar o conhecimento para o cliente e implementar a cultura de analisa e processar dados. Cada empresa tem uma maturidade analítica diferente, além dos setores terem capacidades analíticas distintas, o desafio é fazer com que todos consigam falar a mesma língua dentro do projeto de *Data Science*.

Tecnologias utilizadas para o desenvolvimento do projeto: R, *Python*, Excel.

4.1.3 Caso E3

Modo de trabalho: venda de *software* para análise de dados e tomada de decisão.

Entregas realizadas: relatórios eletrônicos dentro do *software* da empresa.

Uso frequente das análises de dados pelos clientes: *marketing*: pesquisa de perfil, satisfação, NPS; recursos humanos: avaliação 360, clima; qualidade: auditorias, análise de outliers; inteligência: captura de sinais, ouvidoria, SAQ.

Etapa 1: entendo o problema e definindo o projeto

Nesta fase a empresa E3 faz uma primeira aproximação criando um protótipo que mostre o valor da solução. Após é desenvolvida a proposta para o projeto e realizado

o fechamento. Essas reuniões podem acontecer presencialmente ou por videoconferência. Os projetos podem ser de ciclos curtos (1 a 2 meses) e ciclos longos (fluxo contínuo).

Alguns projetos de cunho mais simples já vêm com a necessidade formatada e o cliente busca mais uma forma de automatizar o processo através da ferramenta que a empresa E3 oferece. Já nos projetos maiores existe um detalhamento maior da parte técnica. Para a compreensão do problema do cliente a empresa E3 agenda uma reunião para entender os processos e as ferramentas utilizadas pelo cliente, verifica possibilidades de integração das ferramentas do cliente com o *software*, para então ser feito um detalhamento preliminar do que precisa ser feito no projeto.

Utilizam diagrama de caso de uso e o desenvolvimento dos protótipos junto aos usuários para a compreensão da entrega.

Etapa 2: acessando e preparando os dados

O primeiro acesso aos dados é feito em cima de uma amostra da base de dados do cliente, em um arquivo no formato (csv). Nesta fase a empresa E3 já faz uma análise da estrutura do dado entregue, verificando se os dados estão adequados para realizar o processo. Muitas vezes o cliente entrega dados consolidados quando necessitam estar em seu estado bruto o que pode alterar o resultado esperado. Verificam-se nessa fase as: colunas, estruturas, unidades de análises etc. Geralmente a empresa E3 pede para o cliente fazer a adequação dos dados, porém em alguns projetos a empresa fica responsável por realizar esta etapa utilizando *softwares* como apoio a transformação dos dados.

Etapa 3: criando e validando um modelo

A empresa E3 normalmente utiliza seu *software* para realizar as análises. Outras vezes ela desenvolve uma estrutura personalizada para atender a necessidade do cliente. As análises buscam outliers, diferença entre médias significativas entre outras informações que podem ser relevantes para resolver o problema identificado junto ao cliente. A modelagem dos dados acontece dentro do *software* que já possui modelos pré-prontos. A validação dos modelos é aplicando em bases de dados distintas da que passou pela prototipagem e verificando em cima de dados históricos que os resultados estão consistentes.

Etapa 4: desenvolvimento e implementação

Na maior parte das vezes a empresa E3 entrega para o cliente relatórios eletrônicos. Para outras soluções é entregue um módulo do *software* que permite ter acesso aos

relatórios *online* com filtros. Esses relatórios ficam disponíveis aos clientes através de um link e permite que sejam realizadas outras análises através de menus disponíveis na plataforma. Quando existe a necessidade de as informações obtidas integrarem outra plataforma do cliente, é disponibilizado então a exportação desses dados ou a conexão através de uma API.

Maior dificuldade encontrada durante o desenvolvimento dos projetos: desenvolvimento da parte de visualização dos dados, gráficos e tabelas que sejam capazes de entregar a informação de forma rápida e precisa.
Tecnologias utilizadas para o desenvolvimento do projeto: PHP, *Python*, Java, R, *software* de análise desenvolvido pela empresa.

4.1.4 Caso E4

Modo de trabalho: venda de *software* para análise de dados e tomada de decisão.
Entregas realizadas: produto com módulos de análises, listas de execução, índices calculados baseados nos dados que estão na base do *software* da empresa.
Uso frequente das análises de dados pelos clientes: predição de compra de produtos, gestão de estoques, abastecimento de produtos, gestão dos distribuidores.

Etapa 1: entendo o problema e definindo o projeto
O *software* já atende vários problemas conhecidos pelo setor que a empresa E4 atende. O produto pode ser adquirido de acordo com o módulo que o cliente precisa para resolver o problema.
Para desenvolver novos módulos dentro do sistema a empresa E4 busca atender algumas necessidades específicas de alguns clientes que possam ser replicadas para outras empresas do mesmo setor. Para compreender o problema a ser resolvido a empresa E4 utiliza o processo *Google Design Sprint*[1]. A empresa E4 também realiza estudos de campo para melhorias dentro do *software* e entregas das análises já realizadas.
Etapa 2: acessando e preparando os dados
A empresa E4 possui uma grande base de dados no formato bruto (60 Terabyte), dada esta realidade, dificilmente a empresa precisa solicitar novos dados. Os dados são

[1] http://www.gv.com/sprint/

tratados pela própria empresa deixando prontos para a aplicação de modelos de aprendizagem ou análises estatísticas.

Etapa 3: criando e validando um modelo

Com os resultados do Google Design Sprint a empresa E4 cria novos modelos e realiza treinamento e testes em cima da base de dados já existente. Os resultados são validados a partir de uma prova de conceito junto ao cliente podendo ser alterados quantas vezes for necessário. Estão avaliando outras etapas com dados que ainda não possuem na base.

Etapa 4: desenvolvimento e implementação

As modificações entram no processo de desenvolvimento do produto dentro da empresa E4. Os módulos ficam disponíveis para contratação pelos clientes da empresa E4.

Maior dificuldade encontrada durante o desenvolvimento dos projetos: entregar uma solução que atenda as reais necessidades do cliente, compreensão dos problemas enfrentados pelo cliente no dia a dia.

Tecnologias utilizadas para o desenvolvimento do projeto: AWS, *Python*, R, *software* de análise desenvolvido pela empresa.

4.1.5 Caso E5

Modo de trabalho: consultoria, análises, diagnósticos.

Entregas realizadas: relatórios, *dashboards*, consultoria cultural.

Uso frequente das análises de dados pelos clientes: gerenciamento de crise, marketing, novos negócios.

Etapa 1: entendo o problema e definindo o projeto

Inicia com uma conversa para compreender a necessidade do cliente. Faz perguntas investigativas para identificar o nível de maturidade da empresa em captar e analisar dados. Dependendo da necessidade de análise do cliente e da maturidade a empresa E5 define o escopo do projeto. Muitas vezes o projeto se trata mais de adaptação cultural do que um processo de análise de dados. Após o entendimento da maturidade a empresa E5 realiza um orçamento e define as entregas. Após são feitas novas reuniões de compreensão do problema, participando desta reunião a consultora e as pessoas indicadas pelo cliente.

Etapa 2: acessando os dados

Para esta etapa a empresa E5 compreende o que o cliente tem de dados e quais são suas políticas de acesso. Dependendo do projeto a consultora trabalha dentro da empresa com sua equipe, ou acessa os dados remotamente. A empresa E5 também entende como necessária a participação de alguns funcionários do cliente durante este processo, para que o acesso aos dados seja facilitado. Além disso, a empresa E5 preocupa-se em fazer com que o cliente absorva a cultura e o entendimento do processo de análise.

Etapa 3: preparando os dados

A parte de preparação dos dados é realizada junto com o cliente. A compreensão dos tipos de dados e da importância deles é feita em parceria com um funcionário que tenha conhecimento da base. A preparação dos dados que envolve limpeza e transformação é realizada pela empresa E5.

Etapa 4: criando e validando um modelo

Nesta etapa a consultora mapeia o tipo de análises estatísticas que precisam ser realizadas e aplicada aos dados para chegar até a solução do problema proposto. Os modelos gerados para as análises não são entregues aos clientes, apenas servem de guia para a consultora.

Etapa 5: desenvolvimento e implementação

As entregas são feitas de acordo com a necessidade e capacidade de cada cliente. As vezes são entregues arquivos em Excel com *dashboards* dentro, outras ela utiliza a versão gratuita do *Qlick Sense* (*software* de *Business Intelligence*). *Dashboards* personalizados são entregues em parceria com o cliente, onde ou algum outro parceiro cria ou o próprio cliente, a empresa E5 auxilia na construção. Algumas das entregas são mais relacionadas a como implementar a cultura dedos dentro da empresa, uma conscientização da importância dos dados e explicação de alguns mitos relacionados ao controle que envolve a análise de dados.

Etapa 6: acompanhamento

A empresa E5 faz acompanhamento do cliente para ver se existiu uma mudança no comportamento na análise de dados.

Maior dificuldade encontrada durante o desenvolvimento dos projetos: cultura do cliente, a compreensão do contexto da análise de dados dentro da empresa e como tomar decisões dentro baseadas em dados.

Tecnologias utilizadas para o desenvolvimento do projeto: *Python*, R, Qlick Sense, Excel, SAS.

4.1.6 Caso E6

Modo de trabalho: consultoria, desenvolvimento de análises, automatização de processos e integração (preparação para análise de dados).

Entregas realizadas: relatórios, *dashboards*, produtos de *softwares* personalizados, algoritmos, auxiliar no processo de *data driven*.

Uso frequente das análises de dados pelos clientes: otimização de distribuição, estoque, previsão e sugestão de venda de produtos etc., análise de sentimento.

Etapa 1: entendo o problema e definindo o projeto

Esta etapa consiste em realizar um diagnóstico do cliente, normalmente os encontros são presenciais com duração de duas a três semanas. Para o desenvolvimento da análise a empresa E6 disponibiliza uma equipe mista e segue um processo pré-determinado, que tem por objetivo responder duas questões: 1) o que o cliente tem hoje: base de dados, tipos dos dados etc.; 2) qual o objetivo da análise. Durante esta etapa a empresa E6 faz algumas sugestões para o projeto baseada nas duas primeiras respostas ou em cima de algumas análises de dados amostrais. Já nesta etapa se tem uma ideia de qual será a entrega realizada para o cliente. É também nesta etapa que se desenvolve a proposta de projeto baseada no diagnóstico e faz a validação do desenvolvimento do projeto junto ao cliente. Os projetos seguem o conceito de MVP e buscam entregar o maior valor possível em um curto prazo, a empresa E6 disse se basear no método *Hypothesis-Driven Development,* que segundo (O'REILLY, 2014) “significa pensar sobre o desenvolvimento de novas ideias, produtos e serviços como uma série de experimentos para determinar se um resultado esperado foi alcançado.”

Etapa 2: acessando os dados

Os dados podem ser acessados pela empresa E6 tanto alocando pessoas dentro do cliente quanto por acesso remoto a uma base de dados ou amostra.

Etapa 3: preparando os dados

Nesta fase a empresa E6 explora os dados históricos para ter uma pré-análise histórica em cima dos dados do cliente, após realiza uma padronização dos dados para que na próxima etapa seja possível aplicar os modelos de análise em cima de

dados que precisam ser processados de forma constante. A fase de preparação dos dados inclui limpeza, transformação e padronização da base.

Etapa 4: criando e validando um modelo

Nesta etapa o time mapeia e desenvolve modelos que podem atender as necessidades do projeto. Durante a etapa de modelagem o time realiza a validação dos modelos verificando a acurácia dos resultados em conjunto com o cliente.

Etapa 5: desenvolvimento e implementação

Durante esta fase é desenvolvido de fato o modelo e o produto final para o cliente, podendo ser em formato de *dashboards*, integrações de sistemas, análises, dependendo do objetivo e maturidade de cada cliente.

Etapa 6: acompanhamento

Existe uma fase de acompanhamento pós projeto para solucionar dúvidas ou dificuldades encontradas durante a fase de adaptação com novos processos de análises e tomadas de decisão.

Maior dificuldade encontrada durante o desenvolvimento dos projetos: diversidade das áreas de negócio, a fase de diagnóstico e a clareza de todo o processo de desenvolvimento do projeto com as entregas que atendam a necessidade do cliente.

Tecnologias utilizadas para o desenvolvimento do projeto: R, *Python*, Excel, *Hadoop*, *Spark*.

4.1.7 Caso E7

Modo de trabalho: consultoria em análise de dados, venda de *software Business Intelligence*, cursos e mentorias em análise de dados e usabilidade da ferramenta de BI.

Entregas realizadas: pequenas análises como demonstração; implantação do *software* de BI.

Uso frequente das análises de dados pelos clientes: índices de marketing, vendas, custos e finanças.

Etapa 1: entendendo o problema e definindo o projeto

A empresa E7 trabalha com uma venda consultiva. Durante esta fase, ela traz os futuros clientes mais próximos de si com iniciativas como: liberação gratuita do *software* para testes e validações; oferecimento de cursos e suporte para a utilização

do *software* etc. Essas ações visam a compreensão da principal dor do cliente e a "conquista do campeão" – termo utilizado para identificar a pessoa que compra a solução dentro da empresa e quem irá defende-la perante o tomador de decisão.

Após a empresa E7 busca compreender o *core business* do futuro cliente e oferece um "*pocket*" – termo utilizado para definir uma entrega de inteligência com uma base de dados relativamente pequena, utilizando o *software* proposto – como demonstração do potencial de entrega da ferramenta. A prova de conceito é entregue em poucas semanas pela capacidade de processamento da ferramenta. A partir dessa demonstração faz-se o escopo do projeto. Esta etapa pode levar de 6 meses a 1 ano, dependendo do tamanho do projeto.

Etapa 2: acessando os dados

O *software* é implantado e conectado com as ferramentas de gestão do cliente para a captura dos dados necessários para a etapa do projeto. Para esta fase a empresa E7 auxilia a criar um *roadmap* de quais dados devem ser priorizados de acordo com a necessidade da área de negócios, para isso, a empresa utiliza o Data Driven Canvas[2].

Etapa 3: preparando os dados

É de responsabilidade da TI do cliente deixar os dados em um formato que os responsáveis pelas análises do negócio consigam ler e criar seus gráficos. Para esta fase o *software* possui um módulo capaz de facilitar a limpeza necessária dos dados de forma visual.

Etapa 4: criando e validando um modelo

Para análises preditivas, a empresa cria modelos na linguagem de programação R e conecta os resultados ao *software* de BI para a exibição dos resultados. Para análises descritivas e diagnósticas, o *software* atende com tranquilidade sem precisar maiores esforços em desenvolvimento. Os responsáveis por explorar os dados dentro do *software* e criar as análises são os profissionais do cliente que estão habituados a entregar constante informação a partir da manipulação de dados brutos e transformando em dados visuais, chamados de "planilheiros" – termo utilizado para identificar o profissional com capacidade analítica que trabalha criando gráficos. Os índices e gráficos criados por estes profissionais são validados em conjunto com a pessoa de negócio, que irá consumir estes dados a fim de tomar decisões.

Etapa 5: implementando a solução

[2] https://www.datadrivencanvas.org/

A implementação do projeto pode durar de seis meses a três anos, dependendo do tamanho do cliente e do objetivo do projeto, podendo ter uma relação contínua. A fase de implementação está focada na instalação do *software* dentro do cliente e passa pelas fases 3, 4, 5 e 6 descritas acima.

Etapa 6: acompanhamento

A empresa E7 faz o acompanhamento dos processos de análise dos clientes, muitas vezes como forma de suporte e mentoria.

Maior dificuldade encontrada durante o desenvolvimento dos projetos: Foi identificado como maior dificuldade a cultura empresarial relacionada a: capacidade das pessoas de ler, trabalhar, analisar e argumentar com dados. Falou-se sobre a necessidade de alfabetização de dados.

> "*The emergence of data and analytics capabilities, including artificial intelligence, requires creators and consumers to "speak data" as a common language. Data and analytics leaders must champion workforce data literacy as an enabler of digital business, and treat information as a second language*." (GARTNER, 2018)

Tecnologias utilizadas para o desenvolvimento do projeto: KNIME, R, AWS, *softwares* de BI vendidos pela empresa.

Para todas as empresas, foi questionado o conhecimento dos métodos de análise de dados descritos neste trabalho. Duas das empresas entrevistadas disseram conhecer o método de KDD, os demais processos são desconhecidos.

4.2 Análise dos resultados

As análises serão divididas em duas seções, a primeira será sobre os dados levantados no questionário *online* e a segunda uma análise comparativa dos métodos das empresas entrevistadas com o *Data Analytics Lifecycle*, um dos métodos descritos da literatura, alinhado com os objetivos do método multi-casos.

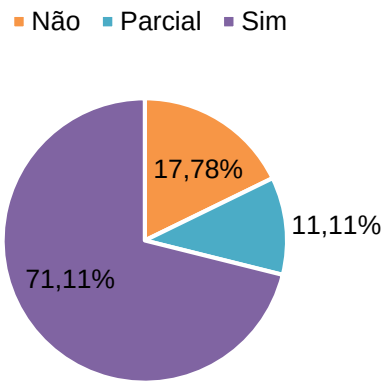
Os resultados obtidos pelo questionário e pelas entrevistas foram cruzados de forma a embasar as conclusões obtidas nas análises.

4.2.1 Questionário *online*

Com o questionário *online* buscou compreender, de uma forma mais generalista, como as empresas vêm trabalhando com análise de dados. Por isso, o perfil dos respondentes do questionário *online* não pode ser diretamente comparado com o perfil das empresas entrevistadas.

Ao todo, 45 pessoas responderam o questionário *online*, destas 17,78% não trabalham com análise de dados e, por isso, não responderam a parte do questionário que detalha como se dá a análise dentro da empresa. Ainda que não possamos fazer comparações diretas entre os entrevistados e os respondentes do questionário *online*, é interessante trazer para a análise uma das dificuldades encontradas pelas empresas de consultoria, a cultura em analisar dados. O que corrobora com a porcentagem dos respondentes que informaram não realizar nenhum tipo de análise dentro da empresa.

Gráfico 1 - Realizam análise de dados nas empresas

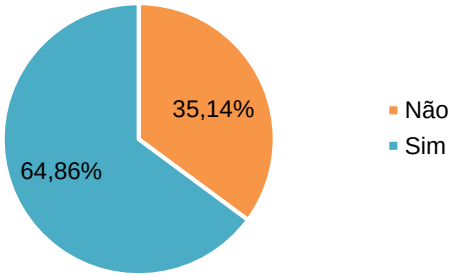


Fonte: elaborado pela autora

Perfil dos respondentes

O perfil dos respondentes foi bastante diverso, sem despontar para nenhum cargo ou área dentro das empresas. Podemos inferir que diversas áreas e pessoas com perfis diferentes realizam análise de dados, ainda que, pelo Gráfico 2, 64,86% dos respondentes tenham dito que as empresas possuem um setor específico para a realização das análises, onde, do levantamento entre as áreas, o setor de inteligência ganha certo destaque com 13% das indicações (Tabela 4).

Gráfico 2 - Empresa possui setor específico para análise



Fonte: elaborado pela autora

Tabela 3 - Perfil dos respondentes por profissão

Cargo	Frequência	%
Analista BI	4	9%
Cientista de Dados	4	9%
Desenvolvedor	4	9%
Gerente	4	9%
Analista de Dados	3	7%
Consultor	3	7%
Analista de Sistemas	2	4%
Assistente de Diretoria	2	4%
Engenheiro de Dados	2	4%
Técnico em Informática	2	4%
Outras ocorrências	15	33%
Total	**45**	**100%**

Fonte: elaborado pela autora

Tabela 4 - Perfil dos respondentes por área da empresa

Área da Empresa	Frequência	%
Business Intelligence	4	9%
TI	4	9%
Comunicação Pública	2	4%
Inteligência	2	4%
Operacional	2	4%
Suporte ao Usuário	2	4%
Outras ocorrências	29	64%
Total	**45**	**100%**

Fonte: elaborado pela autora

Uma questão que parece relevante é o grau de instrução dos respondentes, onde 51% informaram possuir algum tipo de qualificação além da graduação (Tabela 5). O que pode estar relacionado ao perfil necessário para ser um cientista de dados, conforme nos mostra a Figura 5 da sessão 3.2 deste trabalho. No entanto, 42,22%

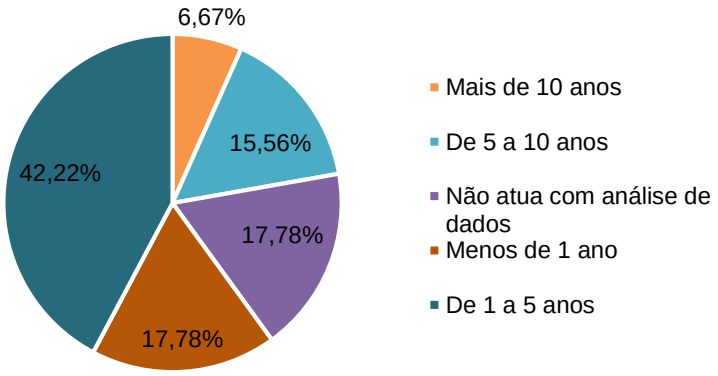
dos respondentes afirmam trabalhar com análise de dados de 1 a 5 anos (Gráfico 3), isso pode demonstrar, além de uma imaturidade do mercado em desenvolver análises, uma alta na busca por perfis analíticos nos últimos anos.

Tabela 5 - Perfil dos respondentes por formação

Formação	Frequência	%
Pós-graduação	18	40%
Graduação	16	36%
Técnico	4	9%
Especialização	3	7%
Doutorado	2	4%
Ensino Médio	1	2%
Graduando	1	2%
Total	**45**	**100%**

Fonte: elaborado pela autora

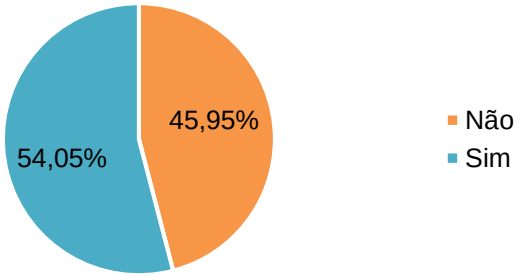
Gráfico 3 - Perfil dos respondentes por tempo de trabalho



Fonte: elaborado pela autora

Outra informação que pode corroborar com esse aumento da busca por pessoas capazes de analisar dados é que 45,95% dos respondentes afirmaram que as empresas não contratam pessoas qualificadas para o cargo (Gráfico 4). Já nas empresas entrevistadas, todos os profissionais atuam com análise a mais de 5 anos.

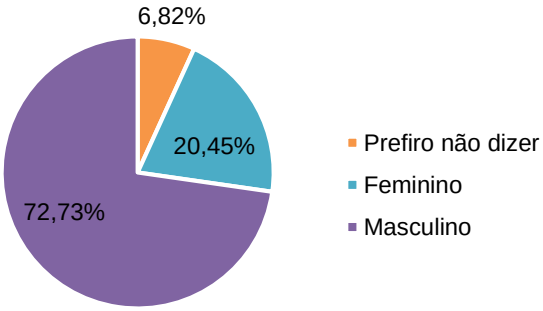
Gráfico 4 - Empresas contratam pessoas qualificadas?



Fonte: elaborado pela autora

A participação feminina é baixa, apenas 20,45% dos respondentes (Gráfico 5). Segundo uma pesquisa realizada pela UNESCO, no Brasil, elas representam 33,1% dos graduados em carreiras STEM (ciências, tecnologia e matemática) (Época Negócios, 2018).

Gráfico 5 - Perfil dos respondentes por gênero



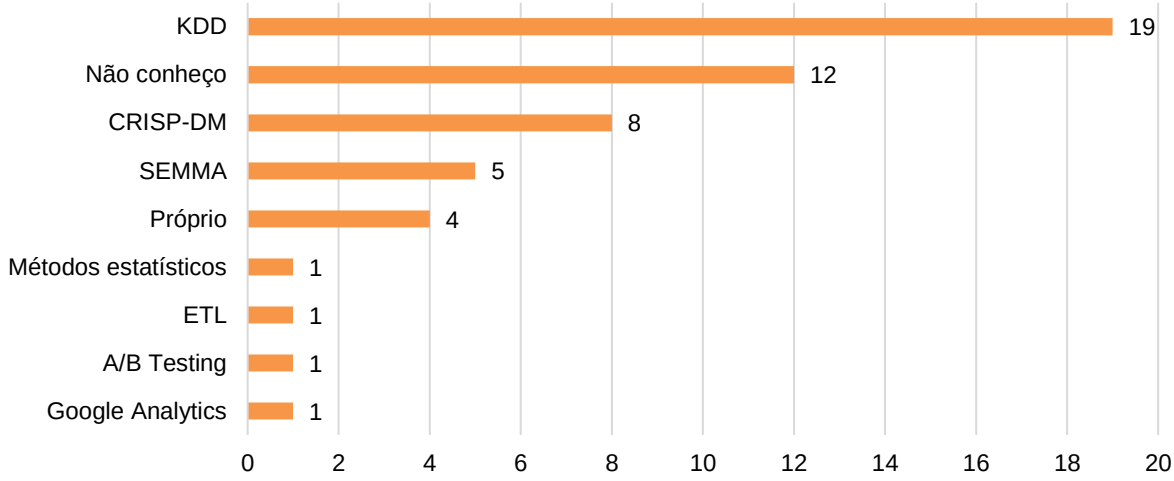
Fonte: elaborado pela autora

As análises dos dados nas empresas

No questionário *online* foram feitas algumas perguntas relacionadas as análises realizadas pelas empresas, o objetivo era, além de identificar se as empresas conheciam os três métodos mais utilizados em projetos de *Data Science* - *KDD Process*, CRISP-DM e SEMMA - conforme pesquisa da (KDNUGGETS, 2014), também era compreender a motivação dessas análises.

Os dados coletados com perguntas abertas foram padronizados pela autora para que pudessem ser expostos em formato de gráficos e tabelas. Como podemos observar no Gráfico 6, a maioria dos respondentes tinham conhecimento de pelo menos um dos métodos citados, com destaque para o KDD com 19 votos.

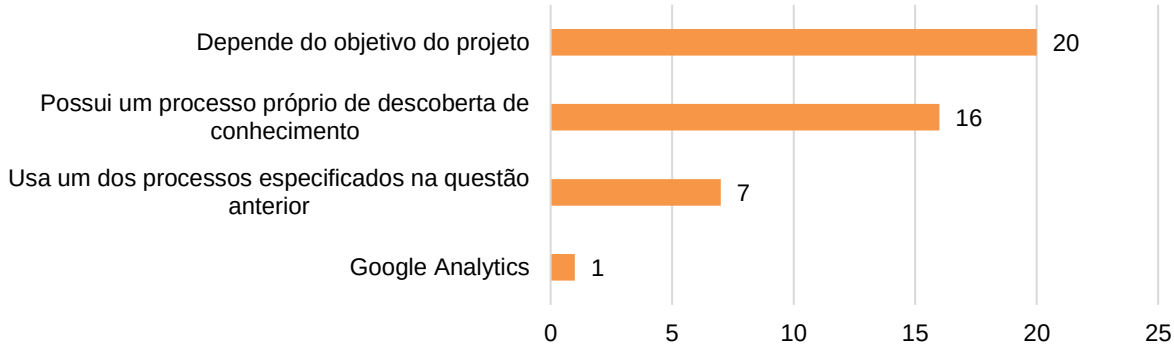
Gráfico 6 - Conhecem os dados citados na literatura



Fonte: elaborado pela autora

Dos respondentes, 44% afirmaram que a empresa aplica processos diversificados para análise de acordo com o objetivo do projeto, contra 35% dos respondentes que afirmam possuir um processo próprio de descoberta de conhecimento (Gráfico 7). Este valor corrobora com a pesquisa da comunidade (KDNUGGETS, 2014), onde 27,5% disseram possuir métodos próprios.

Gráfico 7 - Possui algum processo de análise de dados

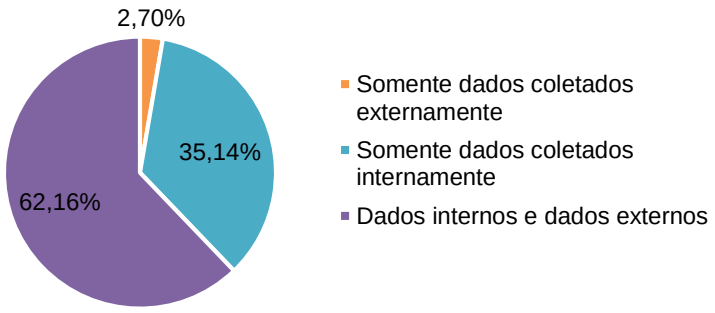


Fonte: elaborado pela autora

Sobre os objetivos para a realização das análises e os tipos de dados coletados pelas empresas, 62,16% afirmaram coletar dados internos e externos, onde o objetivo principal para as análises é para apoio a tomada de decisão. O que corrobora com as áreas mais indicadas dos respondentes, inteligência. Em segundo lugar ficou a entrega de valor para os clientes seguida de melhorar a qualidade do produto. É

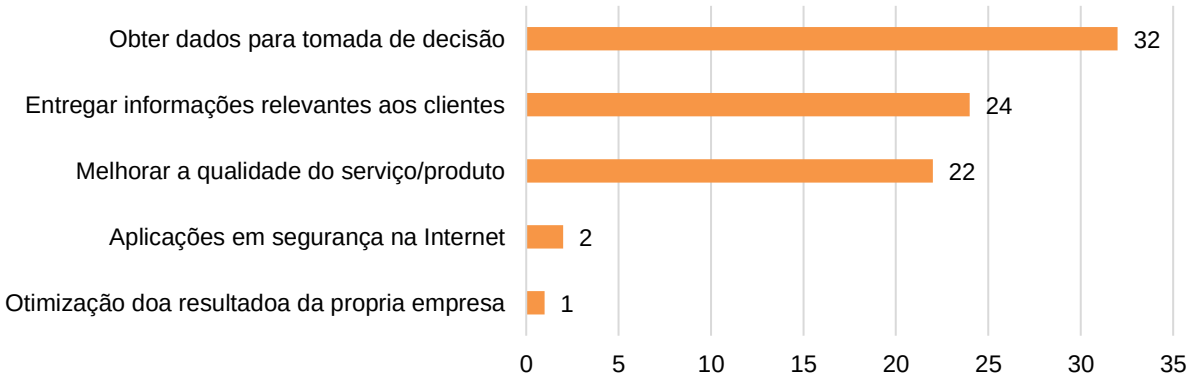
interessante ressaltar que com a chegada da lei de proteção dos dados[3] em junho de 2018 a proporção de aplicações das análises de dados pode ter um aumento significativo no quesito “Aplicações em segurança na *Internet*” (Gráfico 9) item pouquíssimo votado pelos respondentes.

Gráfico 8 - Tipo de dados analisados



Fonte: elaborado pela autora

Gráfico 9 - Objetivos das análises de dados nas empresas



Fonte: elaborado pela autora

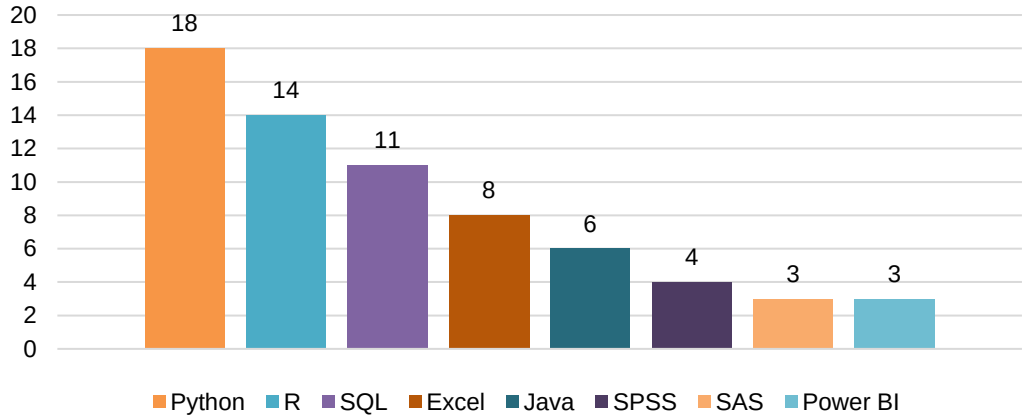
Uma das ferramentas mais votadas para análise de dados segundo os respondentes é a linguagem de programação *Python*, seguida da linguagem R e SQL Lembrando que as empresas podiam indicar mais de uma ferramenta (Gráfico 10). O que segue a tendência dos projetos de análise se compararmos com uma pesquisa realizada pela comunidade *KDnuggets* (PIATETSKY, 2018) com 2052 participantes,

[3] https://www12.senado.leg.br/noticias/materias/2018/07/10/projeto-de-lei-geral-de-protecao-de-dados-pessoais-e-aprovado-no-senado

onde a linguagem *Python* liderou com 65,6% dos votos seguida por *RapidMiner* 52,7% e R 48,5%.

Tanto as empresas entrevistadas quanto as que participaram do questionário *online* indicaram utilizar mais de uma ferramenta para suas análises, o que faz sentido de acordo com Gráfico 7, onde os respondentes indicam utilizar métodos que variam de acordo com o objetivo do projeto.

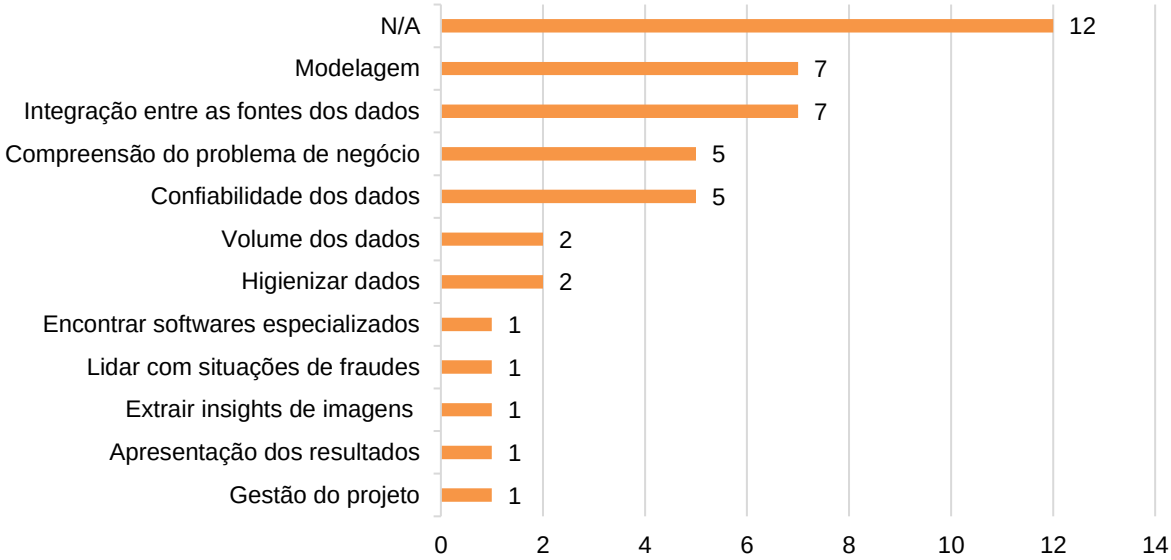
Gráfico 10 - *Software* utilizados para as análises



Fonte: elaborado pela autora

Sobre as dificuldades encontradas nos projetos, diferentemente das empresas entrevistadas, a modelagem foi uma das etapas indicada como a mais difícil, podendo ser uma evidência da diferença dos perfis das empresas entrevistadas e dos respondentes do questionário *online*. Empatada aparece a integração dos dados seguido da compreensão do problema de negócio e confiabilidade dos dados.

Gráfico 11 - Dificuldades encontradas no processo de análise de dados nas empresas



Fonte: elaborado pela autora

Na descrição dos processos, os respondentes não foram fidedignos e alguns se recusaram a responder, mas podemos perceber com a padronização dos dados que a maior parte ainda trabalha com o modelo tradicional, baseados em sistemas de apoio a tomada de decisão (BI). O que faz sentido se olharmos para a Tabela 4, onde temos a ocorrência da área de *Business Intelligence* como setor de atuação, com 9%. Nas demais descrições os métodos parecem seguir o ciclo do CRISP-DM com etapas como: entendimento do negócio, entendimento dos dados, preparação dos dados, modelagem, avaliação e aplicação.

Tabela 6 - Descrição dos processos de análise de dados

Processo realizado	Frequência	%
Extração de amostras de dados, BI	6	13%
Aplicação de modelos estatísticos	4	9%
CRISP-DM	3	7%
Consulta em banco de dados	2	4%
Coletar, ETL, BI	1	2%
ETL Kimball	1	2%
Data Lake conectado com analytics, OLAP, DS	1	2%
Extrair, processar, minerar, conhecimento	1	2%
Negócio, extrair, limpar, analisar	1	2%
Negócio, separar, extrair, BI	1	2%
Aplicação de métodos científicos para análise	1	2%
Sistema próprio para análise	1	2%
Coletar, explorar (ML), avaliar	1	2%

ETL, minerar, explorar (ML), analytics	1	2%
Limpar, minerar, validar	1	2%
Outras ocorrências	19	42%
Total	**45**	**100%**

Fonte: elaborado pela autora

Foi criada uma nuvem de palavras com os dados obtidos pelo formulário. Como foco principal temos a palavra “dados” seguida das palavras: decisão, clientes, relevantes, tomada, entregar, obter, informações, conhecimento, melhorar, qualidade etc. Esta nuvem de palavras nos ajuda a compreender as principais motivações das empresas com a análise dos dados.

Figura 13 - Nuvem de palavras

Fonte: elaborado pela autora com o *software* NVivo

4.2.2 Comparativo dos métodos das empresas

Para fim de compreensão das diferenças entre os processos das empresas entrevistadas, foi feita uma subdivisão em duas categorias: empresas de consultora e representantes de *software*. Essa classificação foi realizada baseada nas respostas dos entrevistados para a pergunta cinco do roteiro de entrevista (Anexo B) e pela observação da similaridade nos processos de análise de dados entre essas categorias.

Deste modo, foram entrevistadas quatro empresas de consultoria e três representantes de *softwares*. O ciclo do projeto foi separado em três fases, a saber:

Fase 1 – Negócio: consiste em toda a etapa de compreensão do problema a ser resolvido, prototipação, definição do escopo do projeto até o fechamento da venda do projeto.

Fase 2 – Desenvolvimento: compreende as etapas de acesso, exploração, manipulação e validação aplicados aos dados.

Fase 3 – Entregas: compreende os produtos finais dos projetos, o que a empresa entrega para seu cliente.

Essas fases são comuns para as duas categorias criadas (consultoria e *software*), o que as diferencia são as etapas realizadas dentro de cada fase. As empresas de consultoria, no geral, possuem uma equipe que trabalha em cima dos dados e na busca por *insights* de negócio. Já as empresas de *software* atuam mais fortemente na implantação do *software* e no treinamento do cliente para o uso do mesmo, conforme podemos observar nas tabelas a seguir.

Tabela 7 - Etapas projeto empresas de consultoria

Empresas de consultoria (E1; E2; E5; E6)		
Fase 1 - Negócio	**Fase 2 - Desenvolvimento**	**Fase 3 - Entregas**
Venda	Mapear dados	Relatórios
Entendimento do Negócio	Separar dados	*Dashboards*
Compreensão do Problema	Limpar dados	Algoritmos
Capacidade da empresa em realizar análises	Explorar dados	Planos de implementação
Prototipação	Modelar dados	Treinamento de equipes
Escopo do projeto	Validar o modelo	

Fonte: elaborado pela autora

Tabela 8 - Etapas projeto representantes de *software*

Representantes de *software* (E3; E4; E7)		
Fase 1 - Negócio	**Fase 2 - Desenvolvimento**	**Fase 3 - Entregas**
Venda	Integrar *software* com os bancos de dados	Relatórios
Entendimento do Negócio	Preparar os dados: ou para o *software* ou para os analistas	*Dashboards*
Compreensão do Problema	Construir modelos em outras linguagens que integrem com a ferramenta	*Software* para análise
Capacidade da empresa em implementar o *software*	Validação dos resultados obtidos pelo *software* com os stakeholders	Treinamento de equipes
Disponibilização do *software* gratuitamente		
Escopo do projeto		

Fonte: elaborado pela autora

Outra característica importante das empresas estudadas é que seus processos de análise ocorrem com um olhar de “fora” da organização que aplica a análise efetuada. Caso os processos fossem internos, ou seja, análises realizadas pelas próprias empresas que iriam agir sobre os resultados, existe a possibilidade de o ciclo do projeto ser diferente.

Neste cenário, o processo escolhido para comparação com as etapas das empresas entrevistadas foi o *Data Analitics Lifecycle* (DAL), pois possui algumas características relevantes que atendem as especificidades dos métodos analisados, a saber:

- Descoberta (fase 1): fica claro a questão sobre aprender o domínio do negócio, tantas vezes relatado pelas empresas entrevistadas;
- Preparação dos Dados (fase 2): aborda as técnicas utilizadas para a criação de um local onde a integração de várias fontes e tipos de dados seja possível, outro fator relevante para as empresas entrevistadas, que muitas vezes focam seus esforços principalmente na preparação dos dados;
- Planejamento do Modelo (fase 3): relatada pelas empresas como fase de compreensão dos dados, onde os analistas podem tirar *insights* e fazer investigações de *outliers*;
- Comunicar os Resultados (fase 5): nesta fase fica clara a necessidade de os resultados serem apresentados para os especialistas de negócio e de terem que entregar dados que possam ser úteis para a tomada de decisão;
- Operacionalizar (fase 6): aborda formatos de entrega diretamente correlacionados aos formatos entregues pelas empresas entrevistadas.

A (fase 4) Construindo um Modelo não foi referenciada pois tem relacionamento direto com a mineração de dados e é muito parecida nas descrições de todos os outros métodos.

O processo SAS-ALC não foi escolhido pois não possui uma fase de verificação dos resultados logo após a modelagem. Já o processo SEMMA não foi escolhido pois, como a própria empresa SAS pronunciou, não é uma metodologia e sim uma organização lógica dentro de um *software*. Já o CRISP-DM foi utilizado pelo autor do

DAL como inspiração, por isso a autora considerou o processo DAL como uma atualização do CRISP-DM mais focada em tecnologias de *Big Data* e *Data Science*, que são o foco da fundamentação teórica deste trabalho e também especificado pelo próprio autor. Por fim o KDD-P, como ele não possui de forma explicita as etapas de entendimento do negócio nem a operacionalização do processo de análise, entende-se que ele não está dentro do escopo das empresas entrevistas.

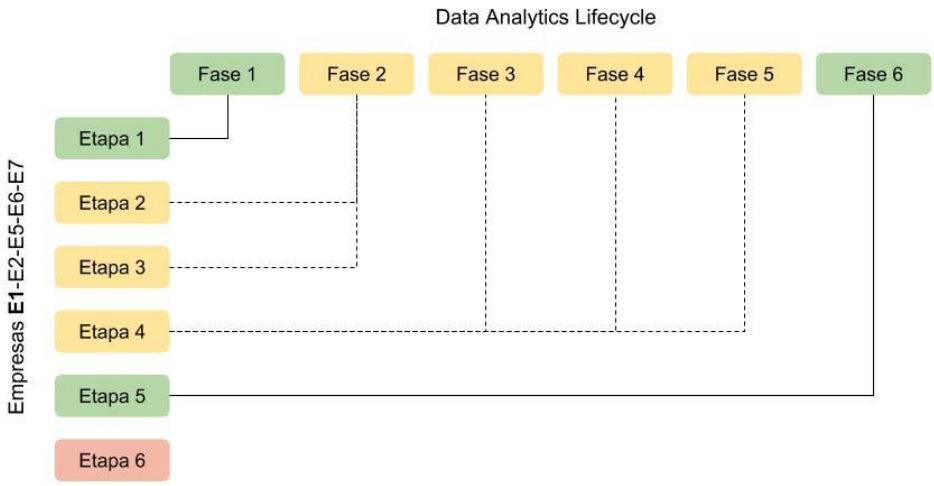
Como forma visual de comparação foi desenvolvida uma matriz onde as etapas das empresas são comparadas com as fases do processo DAL. As cores em verde significam correspondência direta, enquanto as cores em amarelo significam correspondência parcial. A etapa em vermelho significa que não existe correspondência.

Foram feitas duas matrizes, onde, a matriz da Figura 14 representa a correspondência de processos com as empresas de consultoria, com exceção da empresa E7, que foi classificada como uma representante de *software*, porém as etapas realizadas se aproximam das etapas das empresas de consultoria. Já a segunda matriz da Figura 15 apresenta a correspondência dos processos das representantes de *software* com o DAL.

Para as empresas E1; E2; E5; E6 e E7 as etapas 1 e 5 tem correspondência direta com as fases 1 e 6 respectivamente do processo DAL. No que diz respeito as etapas 2 e 3, elas se encontram condensadas dentro da fase 2 do DAL, onde a etapa 2, acessando os dados, é uma etapa importante no processo das consultorias, pois envolve a extração dos dados de dentro das empresas clientes, o que justifica um passo exclusivo para esse procedimento.

A etapa 4 está dividida entre as fases 3, 4 e 5 do DAS, a etapa 4 está relacionada a modelagem dados e as tarefas realizadas dentro desta etapa parecerem estar agrupas em uma mesma força tarefa no processo de análise das empresas. De todas as empresas deste agrupamento, apenas a empresa E1 não possui a etapa 6, relativa ao acompanhamento dos clientes após o projeto. Esta etapa de acompanhamento também não está descrita no método DAS e por isso não possui fase correspondente.

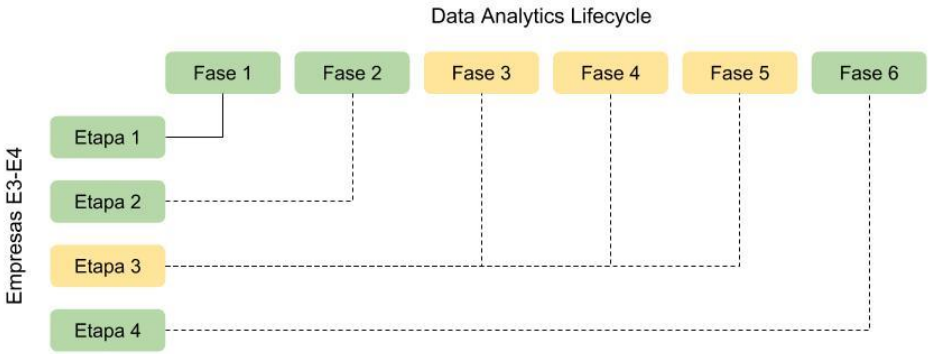
Figura 14 - Matriz de comparação dos métodos das empresas (E1; E2; E5; E6; E7) com o *Data Analytics Lifecycle*



Fonte: elaborado pela autora

Para as empresas E3; E4 as etapas 1, 2 e 5 tem correspondência direta com as fases 1, 2 e 6 respectivamente do método DAL. A etapa 3 está dividida entre as fases 3, 4 e 5 do DAS, a etapa 3 está relacionada a modelagem dados e as tarefas realizadas dentro desta etapa parecerem estar agrupas em uma mesma força tarefa no processo de análise das empresas.

Figura 15 - Matriz de comparação dos métodos das empresas (E3; E4) com o *Data Analytics Lifecycle*



Fonte: elaborado pela autora

Ainda que poucos dos entrevistados tenham indicado conhecer os métodos abordados pela literatura, pode-se observar a semelhança entre às etapas realizadas pelas empresas entrevistadas e as fases do DAL. O processo como um todo é importante, pois como são fases subsequentes, se iniciado o projeto com informações errôneas ou incoerentes pode afetar o resultado obtido, por isso a característica cíclica e iterativa são pontos cruciais. Na tentativa de entregar valor acionável para os clientes, as empresas entrevistadas trabalham com protótipos e entregas curtas, isso permite que o ciclo de vida da *Data Science* aconteça diversas vezes em um mesmo projeto, ainda que com profundidade de ações distintas, como no desenvolvimento de protótipos comparado ao desenvolvimento do projeto como um todo.

Além do comparativo entre os processos um dos problemas relatado com maior recorrência entre os entrevistados foi relacionado a capacidade do cliente em implementar projetos analíticos dada a cultura de tomada de decisão estabelecida e a maturidade dos processos de análise de dados, desde *software*s utilizados até a variedade e quantidade dos dados obtidos. Essa é uma questão muito relevante para as empresas, pois, conforme relatado nas entrevistas, muitas vezes os clientes desistem dos projetos por não se sentirem preparados para lidar com o desafio ou nem chegam a manter o processo de análise feito em sua operação, o que inviabiliza o projeto e desqualifica o esforço realizado pelas organizações.

Em segundo lugar está a preocupação em entregar valor ao cliente de forma clara, abordando as técnicas analíticas e a capacidade de representação visual dos dados, principalmente quando se tratam de grandes volumes de informações. Como os clientes atendidos são de áreas distintas com objetivos distintos, a curva de aprendizado é alta, pois os analistas precisam compreender o negócio do cliente suficientemente bem a ponto de poder analisar e sugerir melhorias. Como a análise tem valor se alcança o principal objetivo estratégico da empresa cliente, é preciso estar a par dos negócios da empresa como um todo para não correr o risco de entregar uma solução grande para um problema pequeno.

4.3 Proposta Canvas *Analytics*

Com base nas experiências deste projeto e motivado por um dos problemas relatados pelas empresas: a compreensão dos objetivos de negócio atrelados aos objetivos de análise de dados, foi feita uma proposta de Canvas *Analytics* pelo grupo de pesquisa da prof. Dra. Daniela Francisco.

O Canvas foi nomeado como *Data Analytics Discovery* (DADCanvas) e tem como objetivo auxiliar de forma visual e resumida a estruturação dos projetos de análise de dados. Foi utilizado como base de desenvolvimento a fase 1 (descoberta) do método *Data Analytics Lifecycle* (DAL) e dividido em sete partes, a saber:

Geradores de Dados: área destinada para registrar as fontes dos dados, ou seja, listar os sistemas e setores responsáveis por produzir os dados e que tipo de dados esses sistemas produzem, além de elencar possíveis dados externos à empresa que possam auxiliar na composição da análise.

Hipóteses: área destinada para registrar os potenciais problemas de negócio que o projeto busca responder.

Consumidores de informação: área destinada para registrar os *stakeholders* do projeto, as pessoas que tem interesse nos resultados obtidos com a aplicação da análise e que possivelmente irão tomar decisões baseados nelas.

Métricas de sucesso: área destinada para registrar as métricas que tem capacidade de medir e validar os resultados encontrados para as perguntas de negócio.

Objetivo do *Data Analytics*: área destinada para registrar o objetivo geral da análise, o problema central que se busca resolver.

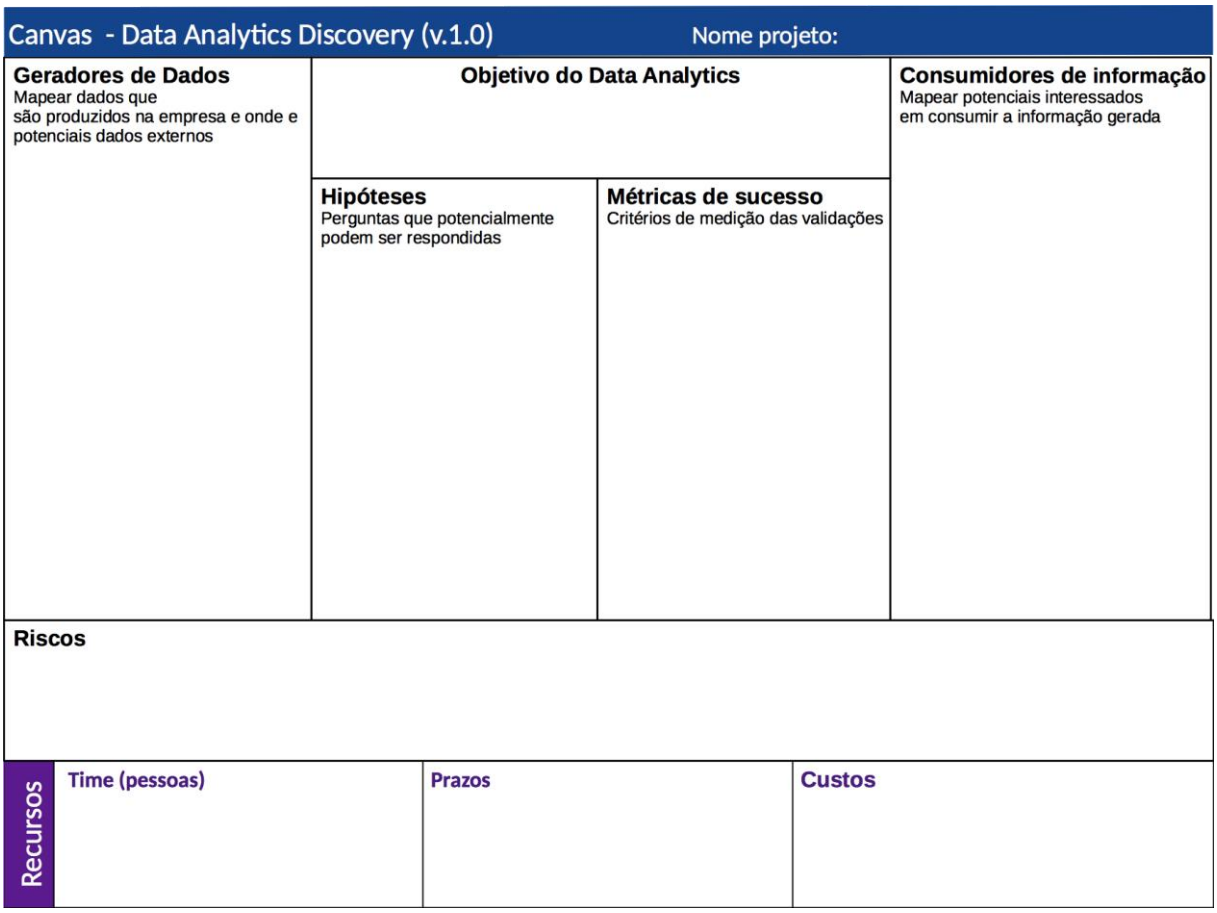
Recursos: área destinada para registrar os recursos disponíveis para a realização do projeto como: pessoas, prazos e recursos.

Riscos: área destinada para registrar riscos potencias ao projeto, como problemas dados insuficientes para a análise, qualidade dos dados coletados, conhecimento da equipe, tempo e custos vinculados ao desenvolvimento do projeto.

O preenchimento do DADCanvas segue uma ordem, de acordo com o encadeamento dos pré-requisitos: 1) geradores de dados; 2) hipóteses; 3)

consumidores de informação; 4) métricas de sucesso; 5) objetivo do *data analytics*; 6) recursos; 7) riscos.

Figura 16 - Canvas *Data Analytics Discovery*



Fonte: elaborado pelo grupo de pesquisa

Aplicação do DADCanvas

O DADCanvas foi aplicado com os alunos da disciplina de Sistemas de Informação Gerenciais das turmas de Administração e Administração Pública noturnas no semestre de 2018/2.

Durante a disciplina os alunos têm como projeto principal desenvolver uma análise em cima de dados abertos coletados em sistemas públicos na Internet, este projeto é desenvolvido em grupo. As ferramentas utilizadas pelos alunos para a realização das análises são: *Weka* e *Qlik Sense Cloud*. Antes de iniciarem com o desenvolvimento do trabalho foi explicado os objetivos do DADCanvas e como deveria ser preenchido e, após, solicitado aos alunos que preenchessem o Canvas com as informações necessárias para o projeto da disciplina.

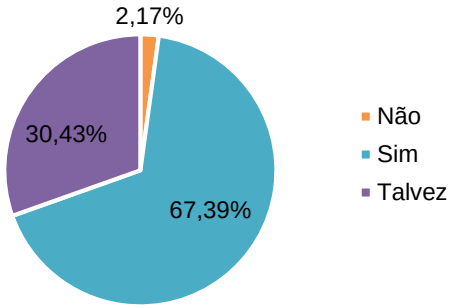
Os alunos também preencheram uma pesquisa *online* (Anexo C) sobre o uso do DADCanvas. O objetivo da pesquisa foi entender se o Canvas ajudou os alunos na compreensão e desenvolvimento do projeto e os objetivos da análise. Ainda que o trabalho tenha sido desenvolvido em grupo, todos os integrantes tiverem que responder a pesquisa realizada.

Resultados da aplicação do DADCanvas

Além da aplicação da pesquisa *online*, foi feita uma avalição qualitativa em cima do preenchimento do DADCanvas, buscando avaliar através das respostas dos alunos a compreensão de cada etapa. Foram avaliados 26 canvas ao total, pertencentes aos grupos das duas disciplinas.

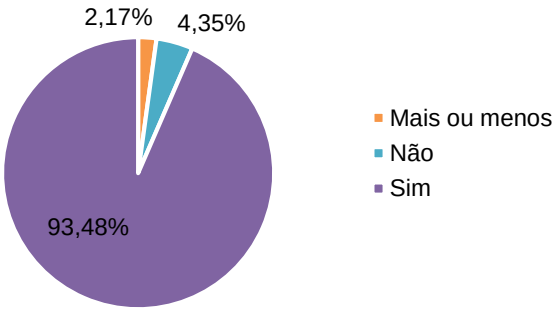
A pesquisa *online* contou com 46 respostas no total onde podemos perceber de forma geral resultados positivos em relação ao *feedback* dos alunos. Dos 46 respondentes, 67,39% disseram que indicariam o DADCanvas (Gráfico 12), e 93,48% disseram que o preenchimento do Canvas contribuiu para o desenvolvimento do projeto (Gráfico 13).

Gráfico 12 - Você recomendaria o DADCanvas?



Fonte: elaborado pela autora

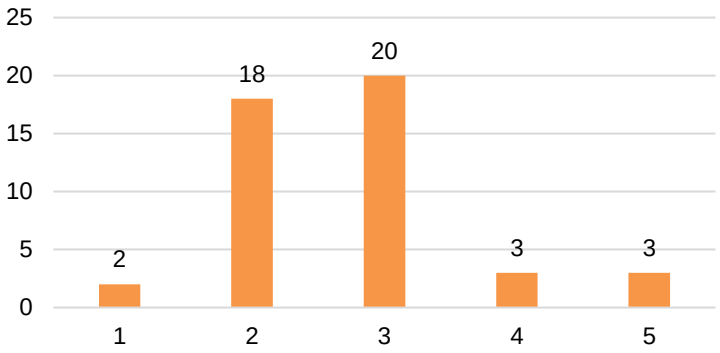
Gráfico 13 - Você acha que o DADCanvas contribuiu para a concepção do projeto?



Fonte: elaborado pela autora

Também foi solicitado aos alunos que classificassem o preenchimento do Canvas dando uma nota de 1 a 5, onde 1 (super fácil) e 5 (muito difícil). A grande maioria dos respondentes classificaram entre 2 e 3, o que nos permite inferir que a dificuldade no preenchimento do Canvas é média tendendo a fácil.

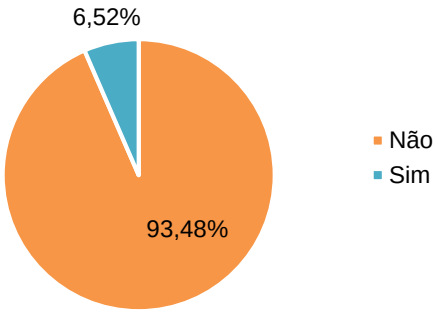
Gráfico 14 - Nível de dificuldade de preenchimento



Fonte: elaborado pela autora

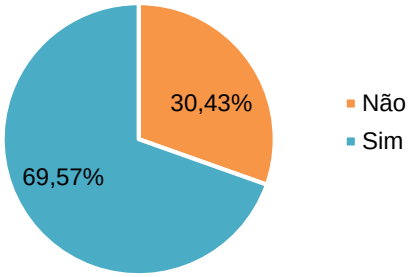
A pergunta sobre o histórico de preenchimento de outros Canvas sobre análise de dados corrobora a percepção da facilidade do DADCanvas quando 93,48% dos alunos afirma não ter preenchido outro Canvas para o mesmo fim (Gráfico 15). Além disso, 30,43% dos respondentes afirmaram não ter recebido explicação sobre o uso da ferramenta, o que não baixou o percentual de facilidade e contribuição (Gráfico 16).

Gráfico 15 - Você já tinha preenchido algum canvas para análise de dados?



Fonte: elaborado pela autora

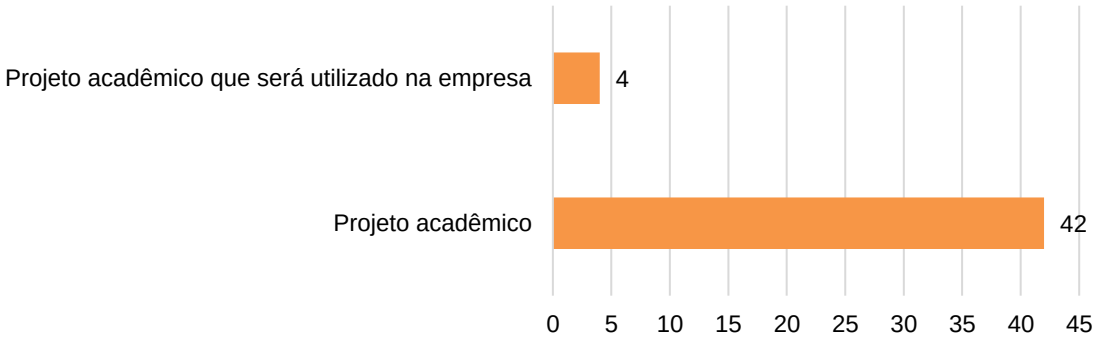
Gráfico 16 - Você recebeu explicação de preenchimento do DADCanvas?



Fonte: elaborado pela autora

Por fim, podemos perceber que alguns dos trabalhos foram de cunho prático, ou seja, o trabalho teve aplicação na empresa dos respondentes.

Gráfico 17 - Contexto de preenchimento do DADCanvas



Fonte: elaborado pela autora

Sobre as avaliações de preenchimento dos trabalhos. Dos 26 Canvas avaliados, 3 grupos utilizaram o modelo errado e 2 não compreenderam o objetivo do

Canvas. Os demais preencheram de forma correta apresentando maior dificuldade nos itens: geradores de dados; métricas de sucesso, riscos e custos. As duas primeiras mais relacionadas a questão de projetos de análises de dados as duas últimas mais relacionadas a gestão de projetos como um todo.

No item geradores de dados, notou-se a dificuldade de preenchimento por se tratarem de bases de dados abertos, dessa forma, é maior a dificuldade para os alunos de saberem quais os setores e *softwares* que geram os dados. Já no item métricas de sucesso, notou-se a dificuldade em transformar a validação das hipóteses em variáveis mensuráveis.

5 CONSIDERAÇÕES FINAIS

Este trabalho teve por objetivos a compreensão dos processos de análise de dados, a descoberta dos métodos utilizados pelas empresas e descritas na literatura, bem como a identificação de gargalos. Para isso, foi realizado um estudo qualitativo de natureza exploratória, com entrevistas (multi-caso) e um questionário *online*.

Por se tratar de um assunto que aborda diversas áreas do conhecimento, a busca por bibliografias foi bastante complexa. A compreensão dos termos específicos da área da *Data Science*, bem como a forma com que cada artigo encontrado desenvolvia suas teorias de acordo com a disciplina a qual estava vinculado, foi um dos pontos mais difíceis no desenvolvimento e escolha dos métodos. Por isso, optou-se por trabalhar com os métodos mais conhecidos como o *KDD-Process*, CRISP-DM e SEMMA e, por consequência, com métodos que podem ser considerados como seus sucessores, com pequenos incrementos e atualizações, SAS *Analytical Life Cycle* e *Data Analytics Lifecycle*.

Outro ponto foi o nível de aprofundamento na descrição dos métodos da literatura. Para cada método existia uma série de detalhes que poderiam ter sido expostos neste trabalho, porém, entendeu-se que seria muito extenso e sem conexão com os objetivos. É preciso salientar que para o quadro comparativo entre os métodos da literatura, a autora utilizou-se do conhecimento não descrito por completo nas fases dos métodos.

Através das descrições dos métodos da literatura, cumpriram-se os objetivos específicos de compreender e identificar os métodos de análise de dados. As entrevistas com as empresas foram baseadas nesse conhecimento desenvolvido durante a etapa de descrição dos métodos.

As entrevistas foram gravadas para consulta posterior e tiveram duração distintas, o que impactou no nível de detalhamento fornecido por cada empresa. Por conta disso, algumas etapas dos processos das empresas ficaram mais condensadas, tendo efeito na descrição e divisão dos seus processos. Outra consequência importante se dá no comparativo dos métodos das empresas com os da literatura, pois, devido à falta de detalhes, as etapas podem ter sido separadas de forma a ficarem distintas dos métodos da literatura, quando poderiam ter correspondência direta ao invés de parcial.

No geral, observou-se uma alta compatibilidade entre as etapas dos métodos da literatura e os das empresas. Levando em consideração que as empresas entrevistadas têm um olhar de fora da instituição que aplica os resultados, seria interessante para compreender se, em um contexto interno, essa compatibilidade seria integral. Dado que, no processo de consultoria, existem as etapas de venda e acompanhamento dos clientes, que não fazem parte do ciclo de vida da *Data Science*, além dos treinamentos e dos projetos que entregam apenas uma etapa do processo, como nos casos em que os clientes não possuem nem a estruturação de uma base de dados.

O questionário *online* serviu como um bom recorte da situação vivida pelos profissionais dentro das empresas. Ainda que a quantidade de respondentes não tenha sido significativa, foi possível identificar algumas diferenças entre as empresas entrevistadas e os perfis dos respondentes da pesquisa *online*, qualificando alguns *insights* obtidos nas análises das entrevistas e comparações entre os processos. As análises dos dados das entrevistas e do questionário *online*, junto com a comparação entre empresas e literatura, cumpriram com o objetivo de compreender o processo de geração de conhecimento a partir dados nas empresas.

Com o cenário desenhado neste trabalho, foi possível identificar que o senso comum sobre as dificuldades das empresas em armazenar, processar e analisar dados estar diagnosticando-as em um estágio inicial. Tanto na sua cultura, ou seja, na operacionalização dos processos e na tomada de decisão baseada em dados, quanto na transformação das análises passando de um sistema tradicional baseado em sistemas de apoio gerencial (*Business Intelligence*) para um processo de análise de um grande volume de dados, podendo assim ser classificado como *Big Data* e *Data Science*. Kozyrkov (2018) afirma que “*Many people only use data to feel better about decisions they’ve already made*”.

A qualificação dos profissionais e a compreensão das empresas do que significa a *Data Science* parece ser outro ponto importante no desenvolvimento deste trabalho. As empresas entrevistadas afirmam não precisar prospectar clientes, o que demonstra não somente a dificuldade dos clientes em realizar análise dos seus dados, como a falta de profissionais desenvolvidos tempo o suficiente para assumir as demandas da área. Outro ponto interessante é a baixa representação feminina na área, tanto nas empresas entrevistas quanto nos respondentes do questionário *online*, a representatividade foi abaixo dos 21%.

Durante o desenvolvimento deste trabalho, a autora participou ativamente de grupos focados em *Data Science* e pode testemunhar muitos relatos sobre as expectativas das empresas e dos profissionais da área. No geral, foi observado que existe uma confusão entre as várias etapas da *Data Science* e, na maior parte das vezes, as vagas são destinadas à área de engenharia de dados, focada na estruturação das fontes de dados para a análise, corroborando com o que foi levantado neste trabalho sobre a dificuldade das empresas na compreensão dos termos da *Data Science*.

Como gargalos e dificuldades identificadas com este trabalho, destacam-se a cultura das empresas em analisar um grande volume de dados, a falta de profissionais experientes para assumirem as demandas levantadas e a dificuldade na compreensão dos problemas de negócio, item que apareceu tanto nas entrevistas quanto no questionário *online*. Para este último item, a proposta do DADCanvas, feita pelo grupo de pesquisa, vem a ser uma alternativa na hora de esclarecer as informações necessárias para a compreensão do problema no início do ciclo do projeto.

Como forma de amenizar o problema de cultura e compreensão, a autora sugere atividades como: cursos, *meetups*, palestras, formações etc., que disseminem no ecossistema mais informações sobre o *Big Data* e a *Data Science*, bem como a necessidade de os projetos de *software* serem baseados em análise de dados.

Para trabalhos futuros seria interessante também entender:

1. Por que as empresas não buscam informações sobre os ciclos de análise de dados, já que muitas conheciam ao menos o *KDD-Process*?
2. Não deveriam todos os *softwares* desenvolvidos terem como base um projeto de análise de dados?
3. Como a aplicação do DADCanvas pode auxiliar em projetos práticos nas empresas?

Por fim, a autora compreende que o trabalho teve seus objetivos atingidos, auxiliando na compreensão da *Data Science*, dos métodos, dos cenários das empresas e de suas dificuldades.

REFERÊNCIAS

AGÊNCIA SENADO. Projeto de lei geral de proteção de dados pessoais é aprovado no Senado. **Senado Federal**, 2018. Disponivel em: <https://www12.senado.leg.br/noticias/materias/2018/07/10/projeto-de-lei-geral-de-protecao-de-dados-pessoais-e-aprovado-no-senado>. Acesso em: 30 Novembro 2018.

ALMEIDA, M. B. Uma introdução ao XML, sua utilização na Internet e alguns conceitos complementares. **CI.Inf.**, Brasília, Maio/Agosto 2002. 5-13.

ANGELONI, M. T. Elementos intervenientes na tomada de decisão. **Ci. Inf.**, Brasília, Janeiro/Abril 2003. 17-22. Disponivel em: <https://www.researchgate.net/publication/26349980_Elementos_intervenientes_na_tomada_de_decisao>. Acesso em: 16 nov. 2017.

AZEVEDO, A.; SANTOS, M. F. KDD, SEMMA AND CRISP-DM: A PARALLEL OVERVIEW. **IADIS**, 2008.

BARBOSA, S. D. L. O Estudo de Caso e a Evolução da Pesquisa em Administração: Limitações do Método. **EnANPAD**, Rio de Janeiro, p. 6-10, Setembro 2008.

BARKER, A.; WARD, J. S. Undefined By Data: A Survey of Big Data Definitions. **Cornell University Library**, UK, 20 Setembro 2013.

BLOEM, J. et al. **Creating clarity with Big Data**. SOGETI. Netherlands. 2012.

BLOEM, J. et al. **Your Big Data Potential**. SOGETI. Groningen. 2013.

BRACHMAN, R. B.; ANAND, T. The Process of Knowledge Discovery in Databases: A First Sketch. **AAAI Technical Report**, 1994.

BRYNJOLFSSON, E.; MCAFEE, A. Big Data: The Management Revolution. **Harvard Business Review**, Outubro 2012. Disponivel em: <https://hbr.org/2012/10/big-data-the-management-revolution>. Acesso em: 11 out. 2017.

CHAPMAN, P.; KHABAZA, T.; SHEARER, C. **CRISP-DM 1.0: Step-by-step data mining guide**. SPSS Inc. [S.l.]. 2000.

DAVENPORT, T. H. How strategists use "big data" to support internal business decisions, discovery and production. **Strategy & Leadership**, 42, Julho 2014. 45-50.

DHAR, V. Data science and prediction. **Communications of the ACM**, New York, v. 56, p. 64-73, Dezembro 2013.

DIETRICH, D. The Genesis of EMC's Data Analytics Lifecycle. **InFocus DELL EMC**, 2013. Disponivel em: <https://infocus.dellemc.com/david_dietrich/the-genesis-of-emcs-data-analytics-lifecycle/>. Acesso em: 19 Outubro 2018.

DIETRICH, D.; HELLER, B.; YANG, B. **Data Science & Big Data Analytics:** Discovering, Analyzing, Visualizing and Presenting Data. Indiana: WILEY, 2015.

EMC DIGITAL. Discover the digital universe of opportunities: rich and the increasing value of the Internet of Things. **DELL EMC**, 2014. Disponivel em: <https://www.emc.com/leadership/digital-universe/>. Acesso em: 01 Janeiro 2018.

FARIAS, F. Google Trends: o que é a ferramenta e como usá-la na sua estratégia. **Resultados Digitais**, 2017. Disponivel em: <https://resultadosdigitais.com.br/blog/google-trends/>. Acesso em: 06 Janeiro 2018.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From Data Mining to Knowledge Discovery in Databases. **AI Magazine**, v. 17, n. 3, 1996.

FRABASILE, D. Apenas 17% dos programadores brasileiros são mulheres. **Época Negócios**, 2018. Disponivel em: <https://epocanegocios.globo.com/Economia/noticia/2018/02/apenas-17-dos-programadores-brasileiros-sao-mulheres.html>. Acesso em: 05 Dezembro 2018.

GANTZ, J.; REINSEL, D. **Extracting Value from Chaos**. IDC iView. [S.l.]. 2011.

GANTZ, J.; REINSEL, D. **THE DIGITAL UNIVERSE IN 2020: Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East**. IDC's Digital Universe Study. [S.l.]. 2012.

GARTNER. Fostering Data Literacy and Information as a Second Language. **Gartner**, 2018. Disponivel em: <https://www.gartner.com/technology/research/data-literacy/>. Acesso em: 23 Novembro 2018.

GERHARDT, T. E.; SILVEIRA, D. T. **Métodos de Pesquisa**. 1. ed. [S.l.]: UFRGS, 2009.

GIL, A. C. **Como elaborar projetos de pesquisa**. 4. ed. São Paulo: Atlas, 2007.

GODOY, A. S. INTRODUÇÃO À PESQUISA QUALITATIVA E SUAS POSSIBILIDADES. **RAE**, São Paulo, v. 35, n. 2, p. 57-63, Março/Abril 1995.

GONZÁLEZ, M. V. CRISP-DM: The methodology to put some order into Data Science projects. **s|ngular**, 2018. Disponivel em: <https://sngular.com/en/crisp-dm-the-methodology-to-put-some-order-into-data-science-projects/>. Acesso em: 23 Novembro 2018.

HAMPTON, J. SEMMA AND CRISP-DM: DATA MINING METHODOLOGIES. **Jesshampton**, 2011. Disponivel em: <https://jesshampton.com/2011/02/16/semma-and-crisp-dm-data-mining-methodologies/>. Acesso em: 10 Outubro 2018.

ISOTANI, S.; BITTENCOURT, I. I. **Dados Abertos Conectados**. [S.l.]: NOVATEC, 2015. Disponivel em: <http://ceweb.br/livros/dados-abertos-conectados/>.

KDNUGGETS. What main methodology are you using for your analytics, data mining, or data science projects? Poll. **KDnuggets**, 2014. Disponivel em: <https://www.kdnuggets.com/polls/2014/analytics-data-mining-data-science-methodology.html>. Acesso em: 11 Novembro 2017.

KITCHIN, R. **The Data Revolution:** Big Data, Open Data, Data Infrastructures & Their Consequences. Londres: SAGE, 2014. Disponivel em: <https://books.google.com.br/books?hl=pt-BR&lr=&id=GfOICwAAQBAJ&oi=fnd&pg=PP1&dq=big+data+definitions&ots=pdrcQTVeTW&sig=mC4NTodY48KtceCziFY2WvuA-eY#v=onepage&q&f=false>. Acesso em: 21 Novembro 2017.

KOZYRKOV, C. Data-Driven? Think again. **Hackernoon**, 2018. Disponivel em: <https://hackernoon.com/data-inspired-5c78db3999b2>. Acesso em: 05 Dezembro 2018.

LANEY, D. **3D Data Management: Controlling Data Volume, Velocity, and Variety**. META Group. Stamford. 2001.

LOUKIDES, M. What is data science? **O'reilly**, 2010. Disponivel em: <https://www.oreilly.com/ideas/what-is-data-science>. Acesso em: 05 Janeiro 2018.

MANYIKA, J. et al. Big data: The next frontier for innovation, competition, and productivity. **McKinsey**, 2011. Disponivel em: <https://www.mckinsey.com/business-functions/digital-mckinsey/our-insights/big-data-the-next-frontier-for-innovation>. Acesso em: 31 Dezembro 2017.

MARBÁN, Ó.; MARISCAL, G.; SEGOVIA, J. A Data Mining & Knowledge. **InTech**, Madrid, Janeiro 2009. 438. Disponivel em: <https://www.intechopen.com/books/data_mining_and_knowledge_discovery_in_real_life_applications/a_data_mining__amp__knowledge_discovery_process_model>. Acesso em: 09 Outubro 2017.

MATTMANN, C. A. A vision for data science. **NATURE**, Califórnia , 24 Janeiro 2013. 473-475.

MAURO, A. D.; GRECO, M.; GRIMALDI, M. A formal definition of Big Data based on its essential features. **Library Review**, 65, Março 2016. 122-135.

MAYO, M. The Data Science Process, Rediscovered. **KDnuggets**, 2016. Disponivel em: <https://www.kdnuggets.com/2016/03/data-science-process-rediscovered.html/2>. Acesso em: 08 Janeiro 2018.

MICROSOFT. What is the Team Data Science Process? **Microsofte Azure**, 2017. Disponivel em: <https://docs.microsoft.com/en-us/azure/machine-learning/team-data-science-process/overview>. Acesso em: 07 Janeiro 2018.

NIST. NIST Big Data Interoperability Framework: Volume , Definitions. **NIST Special Publication 1500-1**, Gaithersburg, Setembro 2015. Disponivel em: <https://bigdatawg.nist.gov/_uploadfiles/NIST.SP.1500-1.pdf>. Acesso em: 20 Dezembro 2017.

O'NEIL, C.; SCHUTT, R. **Doing Data Science:** Straight Talk from the Frontline. 1. ed. [S.l.]: O'Reylly, 2014. Disponivel em: <https://books.google.com.br/books?hl=pt-BR&lr=&id=ycNKAQAAQBAJ&oi=fnd&pg=PR2&dq=%22data+science%22+definitions&ots=-5D5yRixqu&sig=oAR-

A3fhSGZEjQ9RYVsYFz8ckHU#v=onepage&q=%22data%20science%22%20definitions&f=false>. Acesso em: 23 Novembro 2017.

O'REILLY, B. Experimentação é a chave da inovação. **ThoughtWorks**, 2014. Disponivel em: <https://www.thoughtworks.com/pt/insights/blog/how-implement-hypothesis-driven-development>. Acesso em: 26 Outubro 2018.

PIATETSKY, G. Python eats away at R: Top Software for Analytics, Data Science, Machine Learning in 2018: Trends and Analysis. **KDnuggets**, 2018. Disponivel em: <https://www.kdnuggets.com/2018/05/poll-tools-analytics-data-science-machine-learning-results.html>. Acesso em: 02 Dezembro 2018.

PORRAS, J.; YLIJOKI,. Perspectives to Definition of Big Data: A Mapping Study and Discussion. **Journal of Innovation Management**, 2016. 69-91.

PROVOST, F.; FAWCETT, T. **Data Science para Negócios**. Rio de Janeiro: Alta Books, 2016.

ROSSETTI, A. G.; MORALES, A. B. T. O papel da tecnologia da informação na gestão. **Ci. Inf.**, Brasília, Janeiro/Abril 2007. 124-135.

RUSSOM, P. **TDWI Best Practices Report. Big Data Analytics**. TDWI. Renton. 2011.

SAS. Eterprise Miner: SEMMA. **SMU**, 2006. Disponivel em: <http://faculty.smu.edu/tfomby/eco5385_eco6380/data/SPSS/SAS%20_%20SEMMA.pdf>. Acesso em: 10 Outubro 2018.

SAS. **Data Mining From A to Z: How to Discover Insights and Drive Better Opportunities**. SAS Institute Inc. [S.l.]. 2016.

SAS. Enterprise Miner. **SAS**, 2017. Disponivel em: <https://www.sas.com/content/dam/SAS/en_us/doc/factsheet/sas-enterprise-miner-101369.pdf>. Acesso em: 02 Outubro 2018.

SAS. ETL. O que é e qual a sua importância? **SAS**. Disponivel em: <https://www.sas.com/pt_br/insights/data-management/o-que-e-etl.html>. Acesso em: 26 Outubro 2018.

SAS SUPPORT. Data Mining and SEMMA. **SAS**. Disponivel em: <http://support.sas.com/documentation/cdl/en/emcs/66392/HTML/default/viewer.htm#n0pejm83csbja4n1xueveo2uoujy.htm>. Acesso em: 02 Outubro 2018.

SCHMARZO , B. 5 Ways the Internet of Things Drives New $$$ Opportunities. **InFocus DELL EMC**, 2014. Disponivel em: <https://infocus.emc.com/william_schmarzo/5-ways-the-internet-of-things-drives-new-opportunities/>. Acesso em: 01 Novembro 2017.

SFERRA, H. H.; CORRÊA, Â. M. C. J. Conceitos e Aplicações de Data Mining. **REVISTA DE CIÊNCIA & TECNOLOGIA**, v. 11, n. 22, p. 19-34, Julho/Dezembro 2003.

SIGNIFICADO de Hackathon. **Significados**, 2012. Disponivel em: <https://www.significados.com.br/hackathon/>. Acesso em: 25 Outubro 2018.

SILVA, E. L. D.; MENEZES, E. M. **Metodologia da Pesquisa e Elaboração de Dissertação**. 4. ed. Florianópolis: UFSC, 2005.

SIMOUDIS, E. Insightful applications: The next inflection in big data. **O'Reilly**, 2016. Disponivel em: <https://www.oreilly.com/ideas/insightful-applications-the-next-inflection-in-big-data>. Acesso em: 01 Janeiro 2018.

SMITH, J. **Data Analytics:** What Every Business Must Know About Big Data And Data Science. [S.l.]: CreateSpace Independent Publishing Platform, 2016.

TAYLOR, D. Battle of the Data Science Venn Diagrams. **Prooffreader**, 2016. Disponivel em: <http://www.prooffreader.com/2016/09/battle-of-data-science-venn-diagrams.html>. Acesso em: 16 Novembro 2018.

WIRTH, ; HIPP, J. CRISP-DM: Towards a standard process model for data mining, Janeiro 2000.

WITTEN, I. H. et al. **Data Mining:** Practical Machine Learning Tools and Techniques. 4. ed. [S.l.]: Morgan Kaufmann, 2016. Disponivel em: <https://books.google.com.br/books?hl=pt-BR&lr=&id=1SylCgAAQBAJ&oi=fnd&pg=PP1&dq=machine+learning+para+insights&ots=8IDLyfowBf&sig=dW4IEy4-N_uhbSZV10F9sCvwQBw#v=onepage&q=machine%20learning%20para%20insights&f=false>.

ANEXO A – FORMULÁRIO ANÁLISE DE DADOS

Processos de análise de dados

Este questionário tem por objetivo obter uma noção sobre quais são os processos de análise de dados ou de descoberta de conhecimento em banco de dados realizados dentro das empresas. Também buscamos identificar o nível de conhecimento dos profissionais sobre este tema tão incipiente.

Gostaríamos de identificar se processos como KDD, CRISP-DM ou SEMMA ainda são os mais utilizados, ou se já foram atualizados por novas metodologias que ainda não foram exploradas pelo meio acadêmico.

Os resultados deste questionário serão utilizados em um trabalho de conclusão de curso da Escola de Administração da Universidade Federal do Rio Grande do Sul.

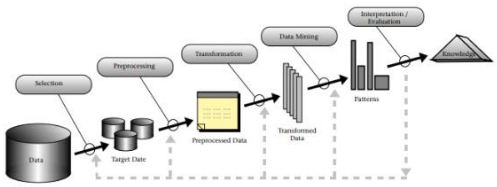
Desde já agradecemos a colaboração de todos.
Aluna de graduação: Karina Moura
Professora orientadora: Daniela Brauner

Breve contextualização

Quando abordamos acima o termo "processo de análise de dados" estamos nos referenciando a todo o ciclo realizado pelos profissionais de Data Science para extrair informações que possam servir como insumo na tomada de decisão pelas empresas.

Quando fizemos uma breve pesquisa sobre este assunto na base de dados de artigos acadêmicos, pudemos achar várias referências com palavras chaves distintas, que podem tanto significar o processo como um todo, ou apenas uma parte dele, como exemplo: "data mining", "knowledge discovery process model", "analytical methods", "project management methodologies" etc.

Um dos processos mais antigos bem documentados pela literatura acadêmica é o KDD (Fayyad, 1996). Vou utilizá-lo como exemplo para demonstrar os tipos de informações que buscamos.



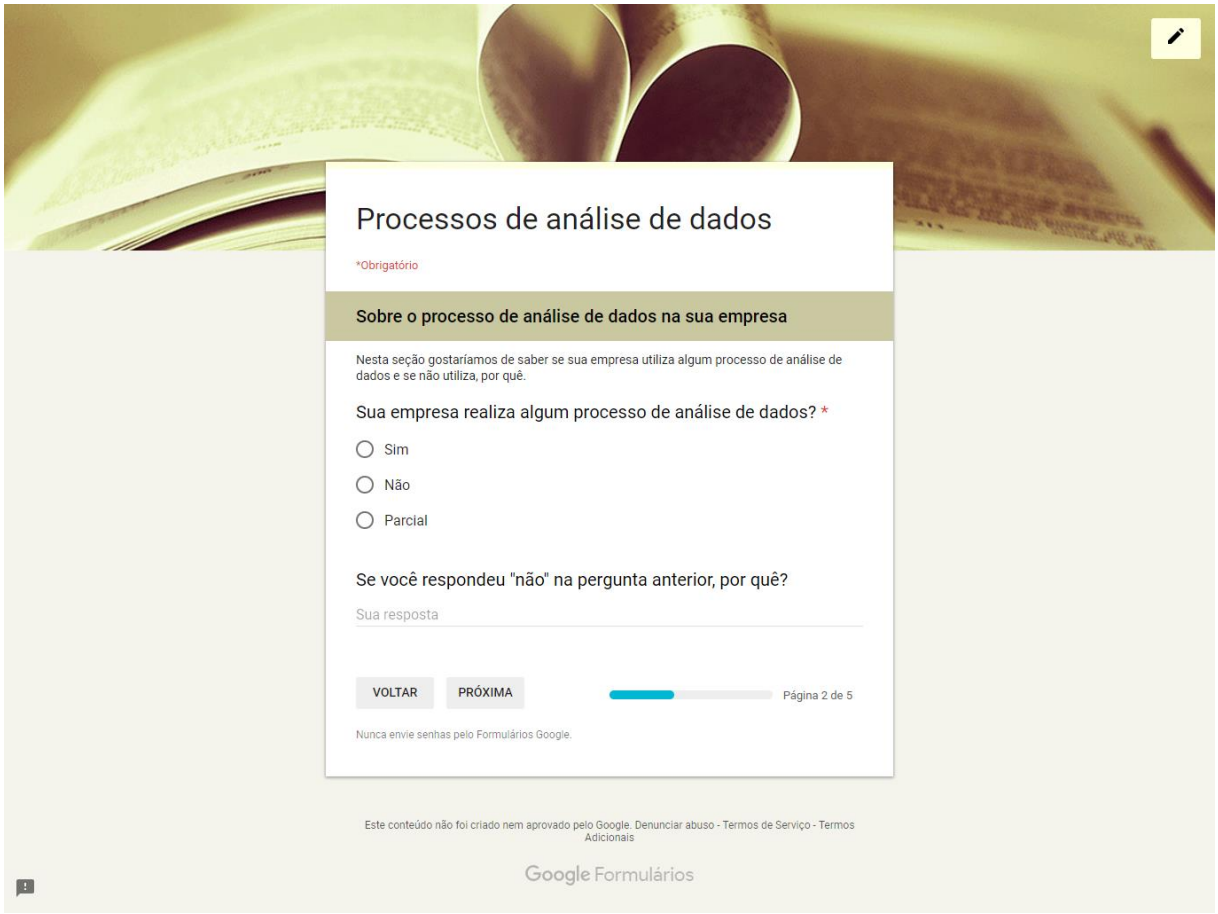
Estes são os passos que compõem o processo de KDD, a saber:

1. Seleção: Este estágio consiste em selecionar um conjunto de dados de destino ou concentrar-se em um subconjunto de variáveis ou amostras de dados, no qual a descoberta deve ser executada.
2. Processamento: Esta etapa consiste na limpeza e pré-processamento dos dados de destino para obter dados consistentes.
3. Transformação: Esta etapa consiste na transformação dos dados usando redução de dimensionalidade ou métodos de transformação.
4. Mineração de dados: Esta etapa consiste na busca de padrões de interesse em uma forma representacional específica, dependendo do objetivo da mineração de dados (geralmente, previsão).
5. Interpretação: Esta etapa consiste na interpretação e avaliação dos padrões minerados.

PRÓXIMA

Página 1 de 5

Nunca envie senhas pelo Formulários Google.



Processos de análise de dados

*Obrigatório

Sobre o processo de análise de dados na sua empresa

Nesta seção gostaríamos de saber se sua empresa utiliza algum processo de análise de dados e se não utiliza, por quê.

Sua empresa realiza algum processo de análise de dados? *

- Sim
- Não
- Parcial

Se você respondeu "não" na pergunta anterior, por quê?

Sua resposta

VOLTAR PRÓXIMA Página 2 de 5

Nunca envie senhas pelo Formulários Google.

Processos de análise de dados

*Obrigatório

Sobre os processos de análise de dados

Nesta seção buscamos entender o nível de compreensão que você e sua empresa possuem sobre os processos existentes na literatura acadêmica.

Você conhece algum dos processos citados sobre análise de dados? *

Marque somente os que você conhece, ou já ouviu falar.

- [ ] KDD
- [ ] CRISP-DM
- [ ] SEMMA
- [ ] Outro: _____

Sua empresa possui algum processo de análise de dados? *

- [ ] Usa um dos processos especificados na questão anterior
- [ ] Possui um processo próprio de descoberta de conhecimento
- [ ] Depende do objetivo do projeto
- [ ] Outro: _____

Descreva como é realizada a análise de dados em sua empresa: *

Sua resposta

Que tipo de dados sua empresa analisa? *

- ( ) Somente dados coletados internamente
- ( ) Dados internos e dados externos
- ( ) Somente dados coletados externamente
- ( ) Outro: _____

Quais softwares e linguagens de programação são utilizados no processo de análise de dados? *

Sua resposta

Com que objetivos sua empresa faz a análise dos dados? *

- [ ] Melhorar a qualidade do serviço/produto
- [ ] Entregar informações relevantes aos clientes
- [ ] Obter dados para tomada de decisão
- [ ] Aplicações em segurança na Internet
- [ ] Outro: _____

A empresa possui uma área específica de análise de dados? *

- ( ) Sim
- ( ) Não

A empresa contrata profissionais qualificados (ou terceiriza) para a realização dos projetos de análise de dados? *

- ( ) Sim
- ( ) Não

VOLTAR PRÓXIMA Página 3 de 5

Nunca envie senhas pelo Formulários Google.

Processos de análise de dados

*Obrigatório

Sobre a empresa

Nesta seção buscamos saber informações sobre a área de atuação da empresa para identificar setores da indústria que estão investindo em análise de dados.

Setor de atuação *

- ○ Indústria
- ○ Comércio
- ○ Serviços
- ○ Outro: _____

Tamanho da empresa *

- ○ Até 9 empregados
- ○ De 10 a 49 empregados
- ○ De 50 a 99 empregados
- ○ De 100 a 499 empregados
- ○ Mais de 500 empregados

Faturamento médio anual

Sua resposta

VOLTAR PRÓXIMA Página 4 de 5

Nunca envie senhas pelo Formulários Google.

Processos de análise de dados

*Obrigatório

Sobre você

Nesta seção buscamos mais informações sobre quem está respondendo ao nosso questionário.

Você já trabalhou em algum projeto que usou análise de dados? *

- ◯ Sim
- ◯ Não

Cargo *

Sua resposta

Área de atuação dentro da empresa *

Sua resposta

Formação acadêmica *

Selecione todos os graus de formação que você possui.

- ☐ Técnico
- ☐ Graduação
- ☐ Pós-graduação
- ☐ Especialização
- ☐ Outro:

Área de formação que está atuando *

Se você possui mais de uma formação em áreas diferentes, qual delas você aproveita melhor os conhecimentos para o trabalho que você realiza no atual momento?

Sua resposta

Há quanto tempo trabalha com projetos que implicam a análise de dados? *

- ◯ Menos de 1 ano
- ◯ De 1 a 5 anos
- ◯ De 5 a 10 anos
- ◯ Mais de 10 anos
- ◯ Não atua com análise de dados

Qual foi seu maior desafio trabalhando com projetos de análise de dados? *

Sua resposta

Gênero

Escolher

Idade *

Sua resposta

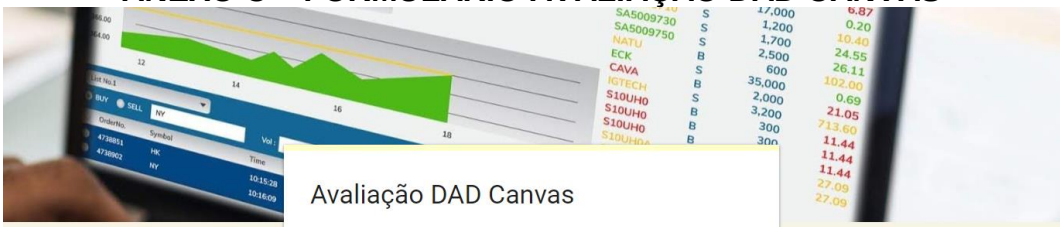
VOLTAR ENVIAR Página 5 de 5

Nunca envie senhas pelo Formulários Google.

ANEXO B – ROTEIRO DAS ENTREVISTAS

1. Permite que seu nome e o da empresa seja publicado no TCC?
2. Nome da empresa
3. Nome do funcionário
4. Formação
5. Cargo
6. Quanto tempo trabalha com análise de dados
7. Projeto atual dentro da empresa
8. Como a empresa trabalha (faz a análise ou faz projetos para as empresas realizarem a análise, tempo médio de duração dos projetos)
9. Como é o processo de análise de dados, etapas desde o contato com o cliente até a solução.
10. Quais as principais aplicações da análise de dados?
11. Quais tecnologias utiliza com frequência nos projetos?
12. O que é mais desafiador?
13. Conhece alguns dos processos registrados na literatura (KDD, SEMMA, CRISP-DM)?

ANEXO C – FORMULÁRIO AVALIAÇÃO DAD CANVAS



Avaliação DAD Canvas

Esta avaliação serve para validar o Canvas proposto.
Solicitamos o e-mail apenas para garantir respostas únicas.

*Obrigatório

Endereço de e-mail *

Seu e-mail

Você preencheu o DAD Canvas em qual contexto? *

- ◯ Projeto acadêmico
- ◯ Projeto da empresa
- ◯ Projeto acadêmico que será utilizado na empresa
- ◯ Outro:

Qual curso você participa? *

- ◯ Graduação
- ◯ Pós-graduação
- ◯ Outro:

Você recebeu explicação de como preencher o Canvas? *

- ◯ Sim
- ◯ Não

Qual nível de dificuldade no preenchimento DAD Canvas? *

	1	2	3	4	5	
Super fácil	◯	◯	◯	◯	◯	Muito Difícil

O que você achou do DAD Canvas? (e sugestões de melhoria) *

Sua resposta

Você recomendaria o DAD Canvas? *

- ◯ Sim
- ◯ Não
- ◯ Talvez

Você acha que o DAD Canvas contribuiu para a concepção do objetivo do projeto de Data Analytics? *

- ◯ Sim
- ◯ Não
- ◯ Outro:

Você já tinha preenchido algum Canvas para análise de dados? *

- ◯ Sim
- ◯ Não

Se sim, qual?

Sua resposta

ENVIAR

Nunca envie senhas pelo Formulários Google.