

UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE INFORMÁTICA
CURSO DE CIÊNCIA DA COMPUTAÇÃO

THIAGO RANGEL DAL MORO

**Data Science na prática:
Transformando problemas de negócio em soluções de dados**

Monografia apresentada como requisito parcial para a obtenção do grau de Bacharel em Ciência da Computação.

Orientadora: Prof. Dra. Renata Galante
Co-orientadora: Prof. Dra. Daniela F. Brauner

Porto Alegre
2022

UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL

Reitor: Prof. Carlos André Bulhões Mendes

Vice-Reitora: Prof^a Patricia Helena Lucas Pranke

Pró-Reitoria de Ensino (Graduação e Pós-Graduação): Prof^a Cíntia Inês Boll

Diretora do Instituto de Informática: Prof^a Carla Maria Dal Sasso Freitas

Coordenador do Curso de Ciência da Computação: Prof. Rodrigo Machado

Bibliotecária-Chefe do Instituto de Informática: Beatriz Regina Bastos Haro

AGRADECIMENTOS

Ao sair da IFRS aos 20 anos de idade, ao tomar a decisão de cursar Ciência da Computação da UFRGS, a todo empenho e dedicação para passar no vestibular, à todo empenho moral, financeiro e emocional da minha família, da confiança que seria possível, da expectativa em entrar na federal e a toda essa jornada, muito obrigado. Chegar ao fim de 2014 até hoje, 2022, é gratificante demais.

Sou imensamente grato a todos que cruzei durante esse período, aos colegas de cadeiras, aos excelentes professores, a todos os funcionários na qual tive contato, muito obrigado.

Sou imensamente grato às minhas orientadoras Renata Galante e Daniela F. Brauner, a leveza, sabedoria e parceria foram incríveis.

Sou imensamente grato a todos que apoiaram este trabalho acontecer, as empresas entrevistadas, as pessoas que me deram dicas, ideias e apoio, muito obrigado.

Sou imensamente grato à oportunidade de ter vivido quase 2 anos dentro da Escola de Administração da UFRGS no período que eu estive no movimento empresa júnior, na PS Júnior. Sem sombra de dúvidas transformou a visão que tinha sobre carreira, visão de mundo, sobre mim mesmo. Uma das experiências mais extraordinárias que a UFRGS me proporcionou.

Sou imensamente grato pelas experiências de estágio e de CLT que tive no meu período de estudante, podendo estar com empresas excepcionais. Ao desafio de lidar com a grade curricular da universidade e me provoca empreender no período da universidade, terminar, empreendendo já há 4 anos, me possibilitando ser um profissional versátil dentro do ramo da tecnologia também.

Ao meu pai, José, à minha irmã, Danielle, aos meus irmãos, Rafael e Frederico, à minha madrasta Lu, um super agradecimento por todo apoio, amo vocês. A todos da minha família expandida, muito obrigado, amo vocês.

À segunda família que são meus amigos, um muito obrigado por me aturarem dizendo que eu iria me formar.

Um super agradecimento à minha mãe e avós que me assistiram do outro lado da vida. É ter a ciência, baseado em fatos, que no fim, tudo valeu a pena. Muito Obrigado.

RESUMO

Em virtude da grande transformação da sociedade nos últimos anos, principalmente, a partir da sociedade da informação, cada vez mais o tema de dados tem sido presente no dia a dia, nas organizações e na vida das pessoas. Muitos estudos têm sido desenvolvidos nesta área de ciência de dados consequentemente a expansão do conceito tem crescido com grande velocidade. Em um contexto de negócio, esta evolução tem acontecido de forma ainda mais acelerada, dada a sua importância no contexto de diferencial competitivo de mercado. Com o objetivo de aprofundar na realidade de dados das organizações e entender como as empresas têm tratado seus desafios de negócio sob o prisma de dados. Com isso, o presente estudo se propõe a realizar um mapeamento das principais metodologias da literatura, uma pesquisa exploratória com empresas, a fim de compreender as principais informações para desenvolver soluções de negócio através de dados. Além disso, aplicar um protótipo de dados em uma empresa real com o objetivo de validar um Canvas de Descoberta de Dados.

Palavras-chave: Ciência de Dados. Problemas de negócio. Métodos. Ciclo de vida. Descoberta de dados. Produtos de dados. *Business Intelligence*.

Data Science in practice: Transforming business problem in data solution

ABSTRACT

This Monography had the purpose of visualizing how organizations have been transforming their business problem using Data. In the context of substantial society transformation in the last few years, a lot of studies have been done to improve the capacity to deal with Data Science, therefore, this study had the objective to explore the main methodologies recommended by literature about this kind of data projects. Furthermore, it did exploratory research with companies who are working in the Data area. Lastly, a prototype of data was applied in a company aiming to validate a method of Discovery Data developed throughout this study.

Keywords: Data Science. Business Problem. Method. Data. Lifecycle. Discovery of data. Product of data. Business Intelligence.

LISTA DE FIGURAS

Figura 2.1 – Diagrama de <i>Venn</i> da Ciência de Dados	25
Figura 2.2 – Comparação das últimas pesquisas realizadas	26
Figura 2.3 - Os Quatro níveis de abstração do CRISP-DM.....	26
Figura 2.4 - As 6 fases do CRISP-DM	27
Figura 2.5 - Guia Visual CRISP-DM	28
Figura 2.6 - Ciclo de vida do <i>Data Analytics Lifecycle</i>	30
Figura 2.7 - Proposta Canvas <i>Analytics</i> inspiração para o presente estudo.....	31
Figura 2.8 - Modelo Smart DLC	31
Figura 2.9 - Estudo Nível de Maturidade Gartner	39
Figura 4.1 - Canvas de Descoberta de Dados	57
Figura 4.2 - A representação do Canvas preenchido	64
Figura 4.3 - Painel Parte 1 das análises da Empresa	66
Figura 4.4 - Painel Parte 2 das análises da Empresa	68
Figura 4.5 - Representatividade financeira do Segmento	69
Figura 4.6 - Principais Produtos Vendidos no período	70
Figura 4.7 - Análise produtos Segmento Energia	71
Figura 4.8 - Relevância da Categoria de Produtos Outros no Setor Energia	71
Figura 4.9 - Análise produtos Segmento Serviços	71
Figura 4.10 - Categoria de Produto Outros no setor de Serviços	72
Figura 4.11 - Análise produtos Segmento Mineração	72
Figura 4.12 Produtos vendidos no setor de Mineração	73
Figura 4.13 - Análise de Recorrência de Consultoria Especializada	74
Figura 4.14 - Análise de Recorrência de Estudo Hidrológico	74
Figura 4.15 - Análise de Recorrência de Outros	75

LISTA DE TABELAS

Tabela 2.1 - Tabela comparativa dos métodos de dados pesquisados na literatura.....	33
Tabela 2.2 - Exemplos de problemas de negócio relacionados a projeto de dados.....	37
Tabela 3.1 - Quadro comparativo da análise das tabelas.....	55
Tabela 4.1 - Descrição das etapas do Canvas de Descoberta Dados.....	57
Tabela 4.2 - Resultados obtidos no encontro facilitado do Canvas de Descoberta de Dados..	62
Tabela 4.3 - Dicionário de dados do CRM utilizado nas análises do protótipo	65
Tabela 4.4 - Principais indicadores da área comercial calculados	67

LISTA DE ABREVIATURAS E SIGLAS

API	<i>Application Programming Interface</i>
BD	Banco de Dados
BI	<i>Business Intelligence</i>
CNPJ	Cadastro Nacional de Pessoas Jurídicas
CRISP-DM	<i>Cross-Industry Standard Process for Data Mining</i>
CRM	<i>Customer Relationship Management</i>
DW	<i>Data Warehouse</i>
DBA	Administrador de Banco de Dados
DAL	<i>Data Analytics Lifecycle</i>
DOD	Decisões Orientadas a Dados
DDD	<i>Data-Driven Decision Making</i>
DLC	<i>Data Lifecycle</i>
ETL	<i>Extract Load Transform</i>
FTP	<i>File Transfer Protocol</i>
HBR	<i>Harvard Business Review</i>
IA	Inteligência Artificial
LGPD	Lei Geral de Proteção de Dados
PO	<i>Product Owner</i>
TI	Tecnologia da Informação
UFRGS	Universidade Federal do Rio Grande do Sul
UX	<i>User Experience</i>
MVP	<i>Minimum Viable Product</i>

SUMÁRIO

1	INTRODUÇÃO	20
2	REVISÃO DA LITERATURA	23
2.1	Ciência de Dados	23
2.2	Metodologias	25
2.2.1	<i>CRISP-DM</i>	26
2.2.2	<i>Ciclo de Vida da Análise de Dados</i>	29
2.2.3	<i>Smart Data Lifecycle</i>	31
2.2.4	<i>Comparativo dos métodos da literatura</i>	33
2.3	Equipe de Ciência de Dados	35
2.4	Problemas em projetos de Ciência de Dados	37
2.5	Maturidade analítica	38
3	ESTUDO DE CASO COM EMPRESAS	39
3.1	Metodologia do Estudo de Caso	40
3.2	Entrevistas	41
3.2.1	<i>Empresa 1</i>	41
3.2.2	<i>Empresa 2</i>	44
3.2.3	<i>Empresa 3</i>	46
3.2.4	<i>Empresa 4</i>	49
3.2.5	<i>Empresa 5</i>	52
3.3	Comparação das empresas	54
4	APLICAÇÃO PRÁTICA EM UM CASO REAL	56
4.1	Canvas de Descoberta de Dados	56
4.1.1	<i>Ciclo de dados – Etapas 1 a 5</i>	59
4.1.2	<i>Pilares de sustentação – Etapas 6 a 11</i>	60
4.2	Protótipo	61
4.3	Aplicação do ETL (<i>Extract, Transform, Load</i>)	64
4.4	Organização do BI	66
4.5	Análises do BI	69
4.6	Feedback da empresa	75
4.7	Principais achados da pesquisa	76
5	CONCLUSÃO	78
	REFERÊNCIAS	80
	ANEXO A – ROTEIRO SEMI-ESTRUTURADO DAS ENTREVISTAS	83

1 INTRODUÇÃO

A Revolução Industrial transformou o trabalho manual e, com isso, muitas mudanças, novos conhecimentos e habilidades surgiram a partir disso. Nesse sentido, podemos dizer que a sociedade da informação está fazendo uma revolução intelectual do trabalho. Atualmente, cada vez mais as pessoas se tornam profissionais que detêm o conhecimento, profissionais especialistas, mestres, doutores, uma sociedade pautada na geração de valor a partir da geração do conhecimento. Dados, informações e conhecimento geralmente são três palavras que possuem significados similares, no entanto, com algumas diferenças. Para começar a aprofundar este trabalho de pesquisa, uma primeira e importante separação de conceitos e significados de cada uma delas. Segundo Schreiber et al (2000), dados são entendidos como sinais não interpretados que impactam nossos sentidos a todo momento. Computadores tem seu funcionamento inteiramente por meio de sinais, ao exemplo de *strings*, números e símbolos que são tratados o tempo todo em grandes quantidades. Informação é o momento em que conseguimos dar um significado para os sinais. Por exemplo, uma luz vermelha por si só é somente um sinal, no entanto, quando compreendida em um contexto de trânsito, este sinal passa a ter um significado, portanto, se torna informação. Conhecimento, por fim, é possível ser entendido a partir de dois diferentes aspectos: o primeiro, quando se tem um propósito por trás, quando se quer atingir um objetivo, gerar uma ação, por exemplo. Já, o segundo, quando é possível se desenvolver uma competência, ou seja, quando é o conhecimento produz uma nova informação.

Importantes estudos, por exemplo, de Rossetti e Morales (2007) estão intrinsecamente relacionados ao presente trabalho. A tecnologia da informação (TI) desempenha um papel fundamental dentro das organizações quando usada como instrumento para desenvolver, extrair, codificar, modelar, disponibilizar, compartilhar o conhecimento e a comunicação dentro da empresa. A TI como um fator estratégico de competitividade e sobrevivência nas empresas. Em face disso, é possível compreender como a profissão de cientista de dados passa ser considerada a mais desejada do século XXI e como a disciplina de Ciência de Dados surge nesse contexto (DAVENPORT; PATIL, 2012).

A taxa de crescimento dos dados gerados e armazenados tem crescido exponencialmente e essa explosão de dados tem gerado muitas oportunidades e combinações para se usar estes dados e se criar valor a partir deles. Nesse contexto, a Ciência de Dados surge como a área de

estudo a qual lida com esta grande quantidade de dados e busca extrair conhecimento e informações estratégicas para o negócio, a partir de um processo de descoberta ou de um processo de formulação e teste de hipóteses (NIST, 2015).

No entanto, como sinaliza um estudo realizado pelo Gartner (2019), previa que até 2022, 80% dos projetos de Ciência de Dados irão falhar, pois eles não solucionam um problema de negócio. Em busca de compreender, então, como as organizações estão lidando com este desafio, o presente trabalho faz a seguinte pergunta: **Quais são as informações fundamentais para se transformar um problema de negócio em uma solução de dados?**

Para entender melhor o problema a ser solucionado no presente trabalho, é importante entender a evolução tecnológica que o mundo vem passando nos últimos anos tem atingido a todos. Por meio desta, as organizações e as pessoas têm sido impactadas com um grande volume de informações por diversos meios, principalmente através da internet (ROSSETTI; MORALES, 2007). Com esta vasta quantidade de dados disponíveis, praticamente, todas as organizações estão explorando o tema de dados de forma a aproveitar uma vantagem competitiva de mercado. Esta variedade e volume de dados tem superado a capacidade de análises manuais e em alguns casos excedido a capacidade de alguns bancos de dados. Ao mesmo tempo, os computadores atuais têm tido cada vez capacidade de processamento e algoritmos tem sido desenvolvidos para conseguir realizar algumas análises não possíveis anteriormente. Este tem sido o fenômeno que tem permitido o crescimento do tema de ciência de dados aplicado para negócios. A temática envolve princípios, processos e técnicas para se realizar análises de dados. O mais recente objetivo da ciência de dados tem sido aperfeiçoar o processo de apoio à tomada de decisão para os negócios. Os estudos nessa área de Decisões orientada a dados (DOD) tem mostrado resultados conclusivos no prisma de como estas decisões têm impactado a performance da empresa, um aspecto muito importante para o contexto de negócio. Uma das conclusões do estudo foi que estatisticamente quanto mais a empresa for orientada à dados, mais produtiva ela é (PROVOST; FAWCETT, 2013).

Portanto, o estudo de dados aplicado aos negócios tem sido comprovado um campo fértil e abundante em diversos aspectos, principalmente, pela real vantagem competitiva que o tema de ciência de dados implica sobre a realidade da organização. Nesse sentido, o presente trabalho busca compreender como a literatura aborda este processo de tomada de decisões

orientada a dados, como as empresas pesquisadas estão desenvolvendo as suas soluções para seus desafios e por fim, quais são as informações são fundamentais para o processo de transformação de problemas de negócio em solução de dados.

Este trabalho tem por objetivo identificar os métodos mais utilizados por empresas para gerar valor a partir dos dados, como estes métodos se relacionam com os métodos da literatura e por fim, identificar quais são as informações fundamentais para se transformar um problema de negócio em soluções de dados. Os objetivos específicos são:

1. Compreender as metodologias que a literatura indica para o desenvolvimento de projetos de dados;
2. Mapear e entrevistar empresas que atuem com projetos de análise de dados;
3. Identificar o(s) métodos que a(s) empresas pesquisadas utilizam para realizar esse processo;
4. Entender e mapear o processo de transformação de problemas de negócio em solução de dados pelas empresas pesquisadas;
5. Recomendar informações fundamentais com base na análise deste estudo buscando diminuir possíveis gargalos e gaps deste processo;
6. Criar um protótipo utilizando a metodologia mais adequada buscando identificar como transformar um problema de negócio em solução de dados;

O restante deste trabalho está organizado da seguinte forma: no Capítulo 2 é realizada uma revisão da literatura buscando apresentar e aprofundar os conceitos fundamentais, principais metodologias e desafios em projetos abordando o tema central de ciência de dados. Além disso, são apresentadas as relações de ciência de dados com as demais disciplinas e estruturas de tecnologia. No Capítulo 3, é apresentado um estudo de caso realizado em cinco empresas, onde são abordados como as empresas estão tratando o tema de dados dentro dos seus times e como estão resolvendo os seus desafios de negócio. No Capítulo 4, são apresentados os principais achados e conclusões do estudo, assim como a recomendação de um método de descoberta de dados. Este método serve de base para criação do protótipo, também apresentado neste capítulo. Por fim, no Capítulo 5, são discutidas as conclusões da pesquisa e os encaminhamentos deste trabalho para estudos futuros.

2 REVISÃO DA LITERATURA

Segundo estudo de Rossetti e Morales (2007) "Sob a ótica do valor do conhecimento e na busca de orientar indivíduos e empresas a conquistar espaços nas instituições e nos mercados, muitas pesquisas têm sido conduzidas". Essa visão corrobora com o objetivo do presente trabalho. Neste Capítulo, são abordadas as metodologias referências da literatura para o tema de pesquisa, juntamente com a compreensão dos principais conceitos dentro de um contexto de busca por desenvolver soluções de dados através de desafios de negócio.

2.1 Ciência de Dados

Na presente introdução deste trabalho é feita uma separação importante sobre três palavras que possuem significados distintos, porém, correlacionadas entre si: dados, informação e conhecimento. Para (SCHREIBER *et al.*, 2000), a palavra "dados" é entendida como um sinal não interpretado que impacta a nossa realidade a todo momento. Esta visão vai ao encontro de como (DHAR, 2013) entende que dados é a parte do material bruto da ciência, o insumo básico para construção dos projetos.

Como sinaliza (DHAR, 2013), a palavra "*Science*" está implicada sobre o conhecimento ganho através de um estudo sistemático. Em uma definição rápida, um estudo sistemático está no processo de construir e organizar conhecimento a partir de testes e predições.

De uma forma geral, existem diferentes visões de autores sobre o conceito Ciência de Dados. Uma outra visão é a de Provost e Fawcett (2013), eles entendem que Ciência de Dados é um conjunto de princípios fundamentais que auxiliam e direcionam o processo de extração de informações e conhecimento a partir dos dados. No entanto, já para NIST (2015), Ciência de Dados pode ser entendido como atividades acontecendo na camada de arquitetura do sistema, ao mesmo tempo, essas atividades consumindo dados armazenados com objetivo de extrair conhecimento para a organização (NIST, 2015).

Um importante ponto a considerar nesse momento é que a definição exata das fronteiras de Ciência de Dados não é o ponto principal. Data Science tem sido debatido academicamente de forma profunda e têm sido desenvolvidos inúmeros estudos sobre as definições e fronteiras formais do tema. No entanto, para o presente trabalho, a visão de Provost e Fawcett (2013) corrobora com o objetivo deste estudo. Para se entender de ciência de dados aplicada a

negócios, primeiramente, é importante entender (i) a relação do tema com outros importantes conceitos que estão intrinsecamente conectados e (ii) os princípios fundamentais da ciência de dados.

Para entender o item (i), são abordados os três principais conceitos interconectados com Ciência de Dados na visão dos autores Provost e Fawcett (2013). O primeiro deles é a relação de *Big Data* e Ciência de Dados, que na visão de DHAR (2013) existe uma grande correlação, pois possuem características importantes quando relacionadas à grande quantidade de dados e suas características. As principais características que fazem alusão a essa correlação são os 'V's do *Big Data* (Volume, Velocidade, Variabilidade e Variedade). De forma objetiva, O primeiro 'V' significa Volume, isto é, a referência do tamanho do conjunto de dados. O segundo 'V' significa velocidade, isto é, uma taxa de processamento. O terceiro 'V' significa variabilidade, isto é, as mudanças que vão tendo nos dados com o passar do tempo e por fim, Variedade, isto é, dados vindos a partir de múltiplos repositórios, domínios ou tipos (NIST, 2015). O segundo importante conceito e, muitas vezes, confundido com Ciência de Dados, é o de mineração de dados. Para os autores Provost e Fawcett (2013), mineração de dados é o processo de extração do conhecimento que se utiliza dos princípios de ciência de dados. O terceiro, não menos importante, é a correlação de ciência de dados com *Data-Driven Decision Making* (DDD). Em outras palavras, o objetivo final de ciência de dados é implementar e melhorar o processo de tomada de decisão. O (DDD) refere-se à prática de tomar decisões baseadas em dados que são analisados com um viés puramente intuitivo, ou seja, a partir de um processo que considera o contexto que aqueles dados estão relacionados (PROVOST; FAWCETT, 2013).

Já para entender o item (ii), os princípios fundamentais da ciência de dados, os autores Provost e Fawcett (2013) caracterizam em três perspectivas os princípios fundamentais de ciência de dados: a primeira delas é que o relacionamento entre Ciência de Dados e negócio busca resolver problemas estratégicos e ter uma vantagem competitiva, pois, a ação de extrair conhecimento útil de dados para resolver problemas de negócio pode ser resolvido usando um processo sistemático com estágios bem definidos.

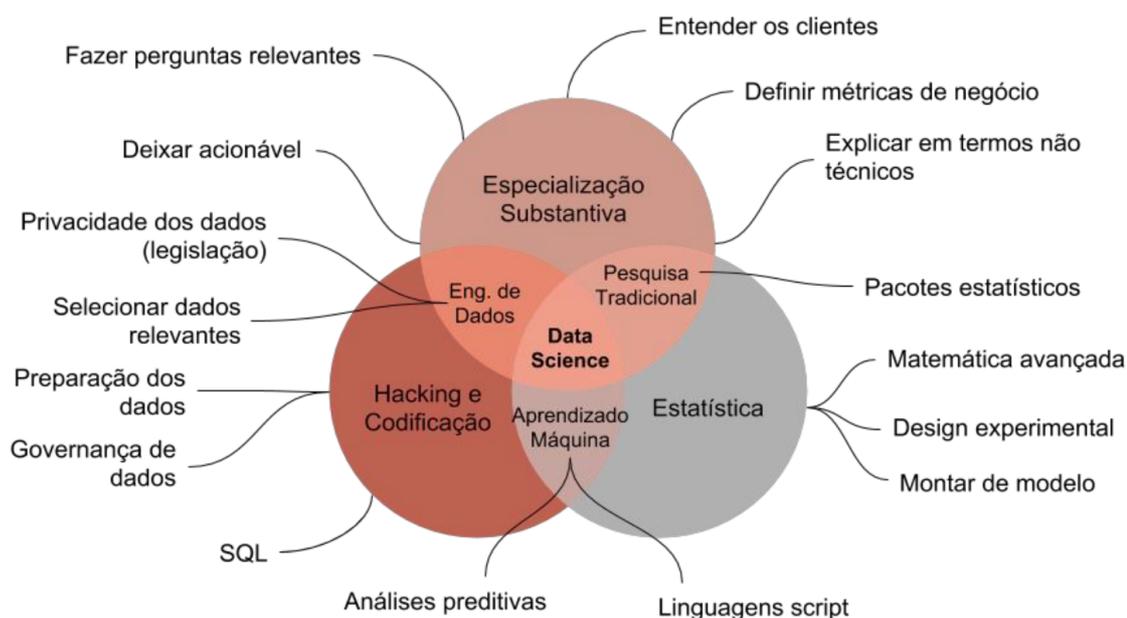
A segunda é que as descobertas de conhecimento a partir do pensamento analítico são as bases para uma solução técnica, em outras palavras, os autores querem dizer: "Se você

realmente olhar profundamente para um conjunto de dados, você irá encontrar algo relevante." (PROVOST; FAWCETT, 2013).

Por fim, a terceira perspectiva é que alguns destes princípios fundamentais estão atrelados a correlações e perspectivas, em outras palavras significa dizer que, para se considerar sucesso nos resultados de Ciência de Dados é fundamental o uso de contexto aplicado a seu uso prático (PROVOST; FAWCETT, 2013).

Para ter uma visão mais sistêmica do que um projeto de dados exige e com o que a ciência de dados se relacione com outros temas, o Gartner (TAYLOR, 2016) propôs um diagrama de *Venn* para se ter essa visualização, conforme Figura 2.1.

Figura 2.1 – Diagrama de *Venn* da Ciência de Dados



Fonte: Karina Moura (2018) – Adaptado Taylor (2016)

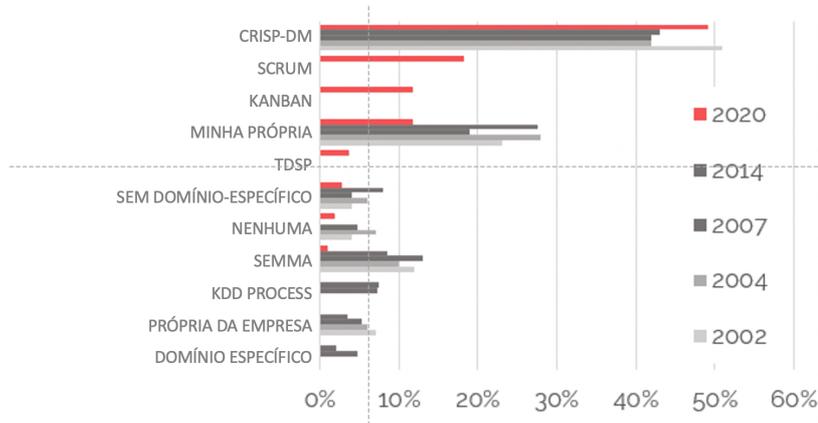
2.2 Metodologias

Nesta seção são apresentadas as metodologias pesquisadas na literatura que buscam uma visão mais ampla do ciclo de vida de um projeto de dados.

2.2.1 CRISP-DM

Uma pesquisa, com 109 respondentes, realizada pela *Data Science Project Management* (2020), indica que CRISP DM (*Cross-Industry Standard Process for Data Mining*) continua sendo a metodologia mais utilizada nos projetos de ciência de dados com 49% das respostas da pesquisa. Em segunda opção de escolha Scrum com 18% e seguido de metodologias próprias e Kanban com 12%. A Figura 2.2 apresenta o resultado em comparação com as últimas pesquisas.

Figura 2.2 – Comparação das últimas pesquisas realizadas

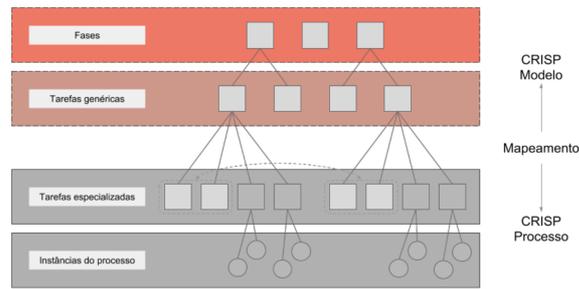


Fonte: Data Science Project Management (2020) – Adaptado pelo autor

CRISP-DM foi concebido em 1996 por três empresas que eram veteranas frente ao novo mercado de mineração de dados que surgia, DaimlerChrysler, SPPS e NCR. (CHAPMAN *et al.*, 2000). CRISP-DM é uma metodologia que é descrita a partir de um modelo hierárquico de processos, consistindo em uma série de atividades descritas em 4 níveis de abstração: Fase, Tarefa genérica, Tarefa especializada e instância de processos, como mostra na

Figura 2.3.

Figura 2.3 - Os Quatro níveis de abstração do CRISP-DM

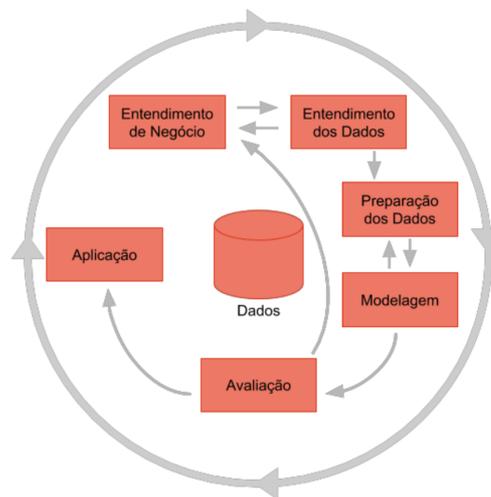


Fonte: Karina Moura (2018) – Adaptado Chapman et al. (2000)

O processo do CRISP-DM apresenta uma visão mais ampla do ciclo de um projeto de mineração de dados, pois consiste em uma visualização das fases do projeto, com suas respectivas atividades e a conexão existente entre as atividades (CHAPMAN *et al.*, 2000).

Um ciclo de vida de um projeto de dados consiste em 6 fases, nas quais sua sequência não é rígida, o que significa que avanços ou recuos - em termos de progresso – não acarretam um problema de projeto, conforme Figura 2.4.

Figura 2.4 - As 6 fases do CRISP-DM



Fonte: Karina Moura (2018) – Adaptado Chapman et al. (2000)

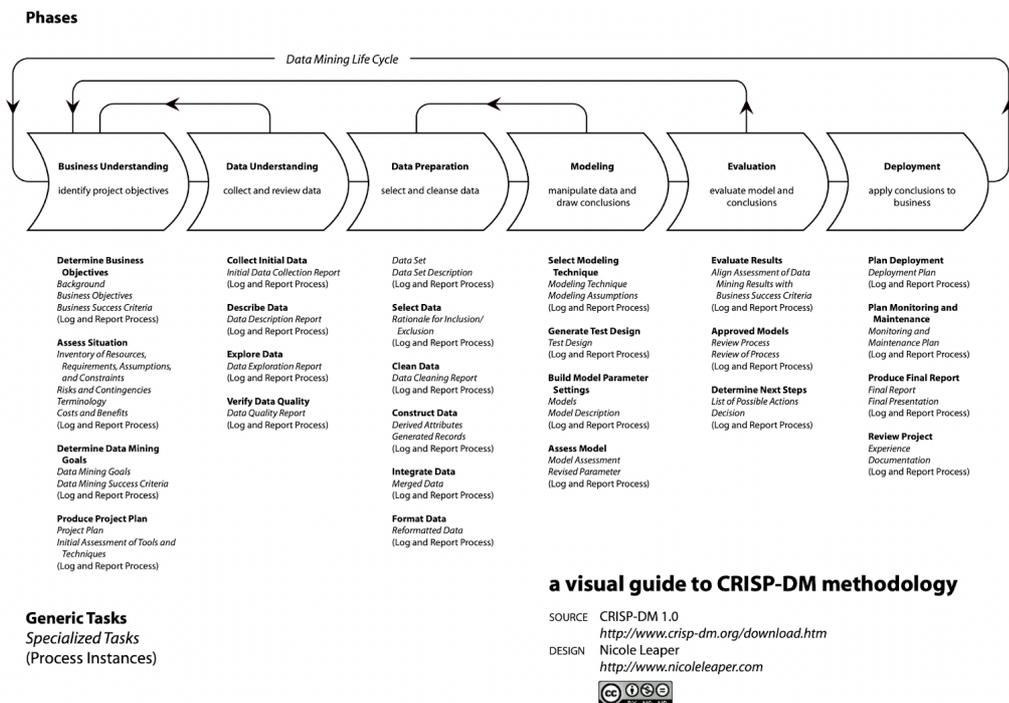
De forma sintetizada, as 6 fases do CRISP-DM consistem em:

- 1) **Entendimento de negócio:** Explorar o contexto e desafios do negócio, bem como os objetivos do projeto, com o intuito de transformar esses fatores em um projeto de dados.

- 2) **Entendimento dos dados:** Explorar o contexto dos dados, quais são as familiaridades existentes com essas informações, quais os problemas de qualidade de dados. Isso significa ter as primeiras impressões com os dados de um projeto.
- 3) **Preparação dos dados:** É o momento de se construir um conjunto final destes dados, conhecido como *datasets*. Neste momento, para chegar ao *dataset*, é importante que os dados passem por uma série de limpezas e transformações.
- 4) **Modelagem:** Técnicas de modelagem são implementadas neste momento, como ajustes de parâmetros e otimização de valores.
- 5) **Avaliação:** Nesta etapa criam-se modelos com uma alta qualidade de análise de dados, com o entendimento que o principal objetivo é resolver um desafio de negócio.
- 6) **Execução (Deployment):** Ter apenas um modelo não significa que o projeto está concluído. É preciso colocar em execução, o que significa apresentar de uma forma que as informações disponíveis agreguem valor ao desafio de negócio.

A Figura 2.5 apresenta uma visão sistêmica da metodologia CRISP-DM (A VISUAL GUIDE TO CRISP-DM METHODOLOGY, 2009).

Figura 2.5 - Guia Visual CRISP-DM



Fonte: *Visual Guide Methodology* (2009)

2.2.2 Ciclo de Vida da Análise de Dados

O método de Ciclo de Vida da Análise de Dados, do Inglês *Data Analytics Lifecycle* (DAL), foi desenvolvido por David Dietrich especialmente para problemas de *Big Data* e projetos de ciência de dados. O objetivo do método é consolidar processos a partir de métodos estabelecidos, buscando obter análises para tomar melhores decisões. O ciclo de vida é dividido em 6 fases e o movimento nestas fases podem ser tanto no sentido de avançar, quanto para recuar (DIETRICH *et al.*, 2015).

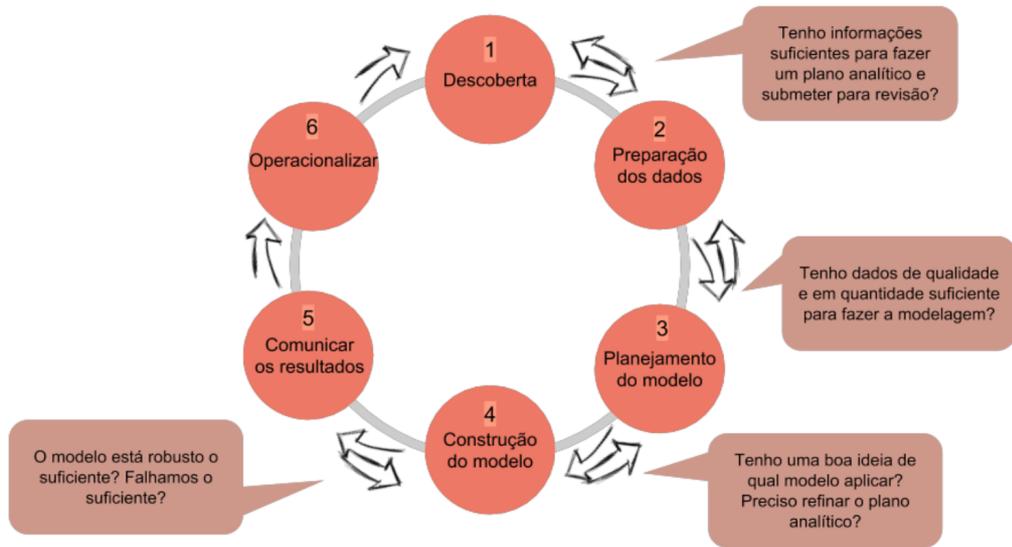
Segundo Dietrich (2015), as fases do DAL, descritas de forma sucinta são as seguintes:

- 1) **Descoberta** - nesta primeira fase, busca-se o entendimento sobre o negócio, a fim de entender o histórico e outros projetos similares que já aconteceram para levantar os aprendizados. Os principais temas a serem investigados nesta fase são sobre pessoas, tecnologias, tempos e informações relevantes. Esta etapa é fundamental para aprofundar no real problema de negócio que se pretende resolver, iniciando uma fase de levantamento de hipóteses sobre este projeto.
- 2) **Preparação de dados** - nesta segunda fase, é necessário um apoio técnico do time de dados. O time envolvido precisa desenvolver a criação de uma ETL (*Extract, Transform and Load*) para conseguir capturar os dados que são relevantes para resolver o desafio levantado na primeira fase.
- 3) **Definindo a modelagem** - esta fase em que se determinam os métodos, as técnicas e o processo que o projeto de dados deve seguir. Nesta etapa, o time busca explorar o relacionamento entre variáveis, identificando quais são as variáveis chave para o desafio.
- 4) **Construindo a modelagem** - este é o momento de desenvolver os conjuntos de dados que queremos testar, treinar e extrair significados relevantes. Nesta etapa também o time desenvolve o que foi construído na fase anterior.
- 5) **Comunicando os resultados** - nesta quinta fase, de forma colaborativa com os principais *stakeholders*, definem-se os resultados do projeto são satisfatórios ou não baseados no critério definido na primeira fase. Aqui é a etapa em que o time busca encontrar os principais resultados, conectando-os com o negócio e, a partir disso, desenvolver uma narrativa para compartilhar com os *stakeholders*.

- 6) **Operação** - esta é a fase em que se entrega um relatório final consolidado com os resultados, incluindo os códigos que foram desenvolvidos junto à documentação técnica. Nesta fase, o time implementa o projeto piloto que deve ir para um ambiente de produção.

A Figura 2.6 apresenta uma versão geral do método DAL.

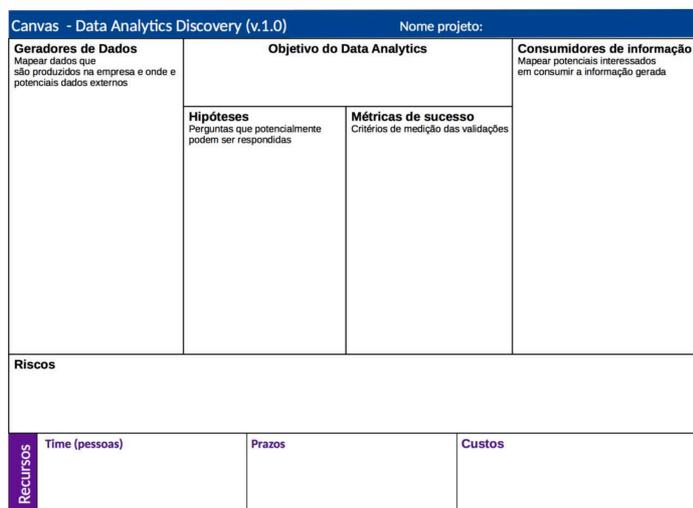
Figura 2.6 - Ciclo de vida do *Data Analytics Lifecycle*



Fonte: Karina Moura (2018) – Adaptado de Dietrich et al. (2015)

Com base no *Data Analytics Lifecycle*, na Figura 2.7, é apresentado o *Canvas Analytics*, publicado no trabalho de conclusão da Karina Moura (2018) com a orientação da Dra. Daniela F. Brauner.

Figura 2.7 - Proposta Canvas *Analytics* inspiração para o presente estudo



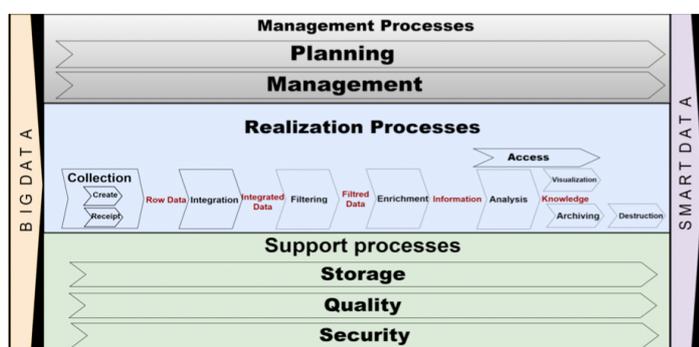
Fonte: Karina Moura (2018)

2.2.3 Smart Data Lifecycle

O modelo foi proposto no estudo de Arass e Souissi (2018) com o objetivo de responder todos os aspectos de um projeto de *Big Data* e transformar para *Smart Data*.

Smart Data Lifecycle (Smart DLC) é entendido em 3 níveis: processos de negócio, processos operacionais e, por fim, processos de suporte. A Figura 2.8 mostra a organização dos processos dentro de cada um dos níveis.

Figura 2.8 - Modelo Smart DLC



Fonte: Arras e Souissi (2018)

Segundo os autores Arass e Souissi (2018), de forma sucinta, cada um dos níveis pode ser entendido da seguinte forma:

- 1) **Processos de negócio** - nível relacionado às atividades de planejar e administrar os processos, que pode ser vista como uma fase transversal durante todo o ciclo de vida dos dados. Esta é uma parte fundamental de entendimento dos dados, pois são descritos mais detalhadamente os dados e quais aspectos de negócio serão abordados, bem como quais decisões serão tomadas ao final do projeto.
- 2) **Processos Operacionais** - neste nível é onde acontece o ciclo de vida do projeto de dados, que consiste nas seguintes etapas:
 - **Captação dos dados** - esse processo geralmente é o primeiro no projeto de ciência de dados, pois consiste em receber os dados brutos e começar a sua organização.
 - **Integração** - etapa que busca disponibilizar um coerente padrão a partir de múltiplas, distribuídas e heterogêneas fontes de dados.
 - **Filtrar** - esta etapa consiste em restringir o conjunto de dados, ou seja, limpar os dados sujos e dados errados.
 - **Aperfeiçoamento** - etapa que consiste em realizar mudanças estruturais e hierárquicas nos dados. Nesta fase, é fundamental ter repositórios prontos para serem aperfeiçoados os dados coletados.
 - **Análises** - etapa considera a mais importante por vários métodos, como é descrito pelos autores. As conclusões e interpretações das decisões definidas no nível de processos de negócio são realizadas nesta etapa.
 - **Acesso** - o foco das aplicações de Big Data é comunicar a interação e os significados para quem é o consumidor destes dados.
 - **Disponibilização** - esta etapa consiste em apresentar os dados de forma didática e inteligente para que seja fácil de entender e interpretar.
 - **Arquivamento** - esta etapa se propõe a guardar os dados por um período mais longo, sendo possível utilizá-los futuramente, se necessário.
 - **Destruição** - etapa que se propõe a excluir os dados que já foram utilizados com sucesso e que possivelmente não tem mais valor.
- 3) **Processos de suporte** - neste nível acontecem processos de apoio e transversais, com 3 processos que são contínuos ao longo do projeto de dados.

- **Armazenamento** - consiste em possibilitar que os dados possam ser armazenados, ou seja, qual a infraestrutura dos dados e onde eles estão armazenados.
- **Qualidade** - é implementado alguns controles para mensurar a qualidade dos dados disponibilizados.
- **Segurança** - este processo consiste em implementar os três principais parâmetros de segurança: integridade dos dados, controle de acesso e privacidade.

2.2.4 Comparativo dos métodos da literatura

Com o objetivo de compreender qual método mais se adequa para servir como base para as demais etapas, foi elaborada a Tabela 2.1, para obter uma visualização sistêmica de como as etapas se relacionam entre si.

Tabela 2.1 - Tabela comparativa dos métodos de dados pesquisados na literatura

MÉTODOS	CRISP-DM	DATA ANALYTICS LIFECYCLE (DAL)	SMART DATA LIFECYCLE (SMART DLC)
Nº de Etapas	6	6	14
Etapas	Entendimento de negócio	Descoberta	Planejar
	Entendimento de dados	Preparação dos dados	Administrar
	Preparação de dados	Planejando modelo	Captação dos dados
	Modelagem	Construindo o modelo	Integração dos dados
	Avaliação	Comunicar os resultados	Filtrar os dados
	Utilização e Aplicação	Operacionalizar	Aperfeiçoar os dados
			Arquivar
			Destruir
			Analizar
			Acessar
			Disponibilização

			Armazenamento
			Qualidade
			Segurança

Fonte: Elaborado pelo autor

As três metodologias, CRISP-DM, DAL e *Smart DLC* possuem uma primeira etapa que busca endereçar as necessidades de negócio. No caso do método *Smart DLC*, os autores buscam ser mais específicos, deixando um processo de administração e um processo de planejamento. No entanto, nessa metodologia, os autores descrevem que essas duas primeiras etapas são transversais do projeto e serão observadas durante todo o ciclo de vida de dados.

Analisando as etapas que tratam dos dados, as três metodologias também abordam o processo necessário de manipulação dos dados, novamente, *Smart DLC* tendo uma abordagem bem mais detalhada, tendo explicitado em três processos que vai desde a captação, integração e filtragem que para o método DAL significa preparação dos dados, já para CRISP-DM esse processo acontece em dois momentos. Nesse caso, o CRISP-DM contempla uma etapa para entender os dados e outra para preparar, quando o DAL realiza as duas na mesma etapa.

Na análise da etapa modelagem, o método *Smart DLC* entende que é o processo de aperfeiçoamento dos dados e para DAL é entendido em duas etapas distintas, planejar o modelo e construir o modelo. Já para CRISP-DM, essa etapa é entendida como somente modelagem. Uma particularidade nessa etapa é que o método *Smart DLC* ainda determina que os processos de destruição e arquivamento são etapas a serem consideradas no método, o que difere dos outros dois modelos.

No contexto de avaliação, todos os métodos possuem essa etapa e são entendidos da mesma forma, momento de se olhar para os dados e se desenhar ideias, insights e ações práticas buscando resolver o desafio de negócio.

Por fim, na etapa de utilização ou também o *deploy*, como é conhecido dentro do contexto de TI, os três métodos também possuem a sua etapa, no entanto, como citado, o método *Smart DLC* busca detalhar mais os processos e, nesse caso, após a etapa de *deploy*, ainda ele sugere que tenham mais três processos transversais, armazenamento, qualidade e segurança, o que não é explicitamente definido nos métodos de CRISP-DM e DAL.

2.3 Equipe de Ciência de Dados

Tanto no mercado de trabalho como na academia, tem-se escutado frequentemente a pergunta: "Onde é possível encontrar um bom cientista de dados?" (MASON, 2010). Não à toa, como mesmo descreve Davenport e Patil (2012) no artigo da Harvard Business Review (HBR), o cientista de dados se tornou a profissão mais desejada do século XXI. Em uma pesquisa realizada pela mesma revista foi também questionado: O que um cientista de dados de fato faz? (BOWNE-ANDERSON, 2018).

A ciência de dados é uma temática que abrange diversos conhecimentos, como demonstrado na Figura 1. Em uma pesquisa realizada com trinta e cinco cientistas, buscou-se responder o que cientistas de dados fazem. A conclusão é que o papel dos cientistas de dados vai desde estabelecer uma base de dados sólida e robusta até o desenvolver análises de performance para a empresa, Como resultado, o cientista de dados é hábil a construir algoritmos de aprendizado de máquina personalizados e produtos de dados para melhor entender do negócio, dos consumidores, e como é possível tomar melhores decisões (BOWNE-ANDERSON, 2018).

Em outras palavras, Dietrich (2015) resume em três conjuntos de atividades:

- 1) **Desafios de negócio e desafios de análise** - o profissional da ciência de dados precisa ter a habilidade de diagnosticar um problema de negócio conseguindo identificar qual é o principal desafio a ser resolvido, a partir disso, determinando quais métodos de análise são candidatos para resolver esse problema.
- 2) **Projetar, implementar e implantar modelos estatísticos e técnicas de mineração de dados** - este conjunto de atividades são as mais evidentes para as pessoas que pensam sobre as atividades do cientista de dados. De fato, aplicar complexos e avançados métodos de análise para resolver problemas de negócio é um dos principais papéis deste profissional.
- 3) **Desenvolver ideias e ações recomendadas** - é importante entender que aplicar métodos de análise a partir de dados não necessariamente gerará valor para o negócio. Por isso, a habilidade de criar ideias e ações a partir dos dados que se apresentam e comunicar de forma clara para os parceiros chave.

Em termos de habilidades deste profissional, Dietrich (2015) resume as cinco principais:

- **Habilidades quantitativas** - sendo a matemática e estatística as duas principais.

- **Habilidades técnicas** - tendo como principal uma visão de engenharia de software, aprendizado de máquina e habilidades de programação.
- **Visão sistêmica e pensamento crítico** - é importante o cientista de dados conseguir ver além de uma única perspectiva do problema, tendo um senso crítico para avaliar os dados que se apresentam.
- **Curiosidade de criatividade** - é fundamental criar ideias e novas visualizações de como utilizar os dados para resolver o problema de negócio que se tem pela frente.
- **Comunicação e colaboração** - o cientista de dados deve ser articulado para trabalhar em conjunto com o time de projeto e se comunicar de forma clara e assertiva os resultados que foram desenvolvidos durante o projeto. Além disso, ter uma narrativa consistente para convencer os stakeholders do valor de negócio que o projeto consegue gerar é fundamental.

Por fim, Dietrich (2015) apresenta uma visualização dos sete papéis chave dentro de um time de dados que gere resultados satisfatórios: Os papéis são os seguintes:

- **Usuário de negócio** - pessoa entende das necessidades do negócio e consegue identificar os benefícios a partir dos resultados-chave do projeto. Ela pode consultar o time durante o projeto, pode orientar, fazer ajustes; ou seja, essa pessoa é um stakeholder fundamental dentro do projeto de dados.
- **Patrocinador do projeto** - responsável pela criação do projeto, disponibilizando os requisitos do projeto e definindo o principal problema de negócio a ser resolvido.
- **Gerente de projeto** - quem garante que as etapas vão acontecer dentro do tempo esperado e da qualidade desejada.
- **Analista de Inteligência de negócio** - pessoa como domínio e experiência profunda do entendimento dos dados, compreendendo os principais indicadores de sucesso do negócio e construindo atualizações e relatórios com a inteligência de negócio.
- **Administrador do Banco de dados (DBA)** - responsável pela configuração da base de dados, que é onde todos os dados são armazenados e fornecem o ambiente necessário para que seja possível fazer as análises dos dados.

- **Engenheiro de dados** - quem possui a capacidade de otimizar as consultas no banco de dados, com uma responsabilidade de administração dos dados e suporte aos diferentes níveis de acesso ao longo do projeto.
- **Cientista de dados** - aquele com habilidade e experiência de técnicas de análise, modelagem de dados, que busca conectar os dados com os problemas de negócio que se quer resolver.

2.4 Problemas em projetos de Ciência de Dados

Os projetos de Ciência de Dados geralmente têm uma característica em comum: são projetos com carácter exploratório mais do que projetos de *Business Intelligence* (BI) ou mesmo de projetos de análise de dados. Um erro muito comum em projetos de ciência de dados é ir direto para coleta e análise de dados antes de planejar e definir um escopo do problema de negócio que se pretende resolver (DIETRICH *et al.*, 2015).

O presente trabalho busca justamente endereçar esse *gap* dentro dos projetos de dados, fornecendo um método para descoberta de dados buscando elencar as informações relevantes e necessárias para se resolver o desafio de negócio.

Na Tabela 2.2, Dietrich (2015) exemplifica problemas de negócio em que hoje os projetos de ciência de dados auxiliam as empresas a se tornarem mais orientadas a dados e mais competitivas nos seus mercados.

Tabela 2.2 - Exemplos de problemas de negócio relacionados a projeto de dados

Necessidade do negócio	Exemplos
Otimizar a operação de negócio	Vendas, preços, eficiência, novas receitas
Identificar risco	Churn, fraudes, padrões
Prever novas oportunidades de negócio	Up-sell, cross-sell, novos prospects
<i>Compliance</i> e leis que regulamentam o negócio	Identificação de lavagem de dinheiro, regras de empréstimos, entre outros

Fonte: Dietrich (2015) - Adaptado pelo autor

2.5 Maturidade analítica

Um estudo realizado pela consultoria Gartner indica como é possível visualizar a maturidade analítica que as empresas possuem. Os dados têm sido utilizados cada vez mais nas empresas, no entanto, nem todos os dados são utilizados para análise e tomada de decisão (GARTNER, 2019).

Diante disso, a consultoria traz uma definição para o termo "*Data Analytics*" que para eles é entendido como dar um sentido, dar uma análise, a partir dos dados que se disponibilizam no momento. Com intuito de visualizar de forma mais clara, foi desenvolvido um modelo de maturidade analítica, conforme mostrado a seguir:

Análise Descritiva - é o primeiro estágio do nível de maturidade sugerido pela consultoria Gartner, é o nível mais básico, que é o descritivo. O objetivo principal da análise descritiva é olhar para os dados com uma perspectiva do que já aconteceu. Em geral, esse tipo de análise é muito comum nas empresas, são estudos realizados, a partir de relatórios, painéis e KPI (Indicadores-chave de performance). As perguntas centrais que se buscam responder neste primeiro nível são: O que aconteceu? O que está acontecendo?

Análise Diagnóstica - o segundo estágio proposto é o nível de diagnóstico. O objetivo é buscar entender a perspectiva do porque algo aconteceu. Este estágio requer mais habilidades de ciência de dados para responder, requer aprofundar as informações e buscar encontrar alguns padrões. Por exemplo, as pessoas mais seniores da empresa podem usar essas informações para entender o comportamento dos usuários, dos clientes, do time comercial. A pergunta principal neste segundo nível é: Por que aconteceu?

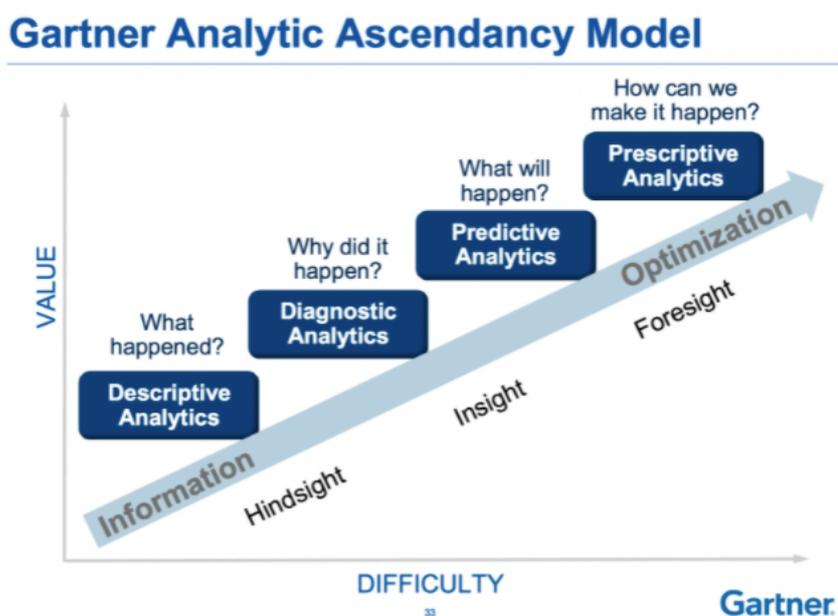
Análise Preditiva - o terceiro estágio é o nível preditivo. Este nível de maturidade já é um estágio desejado por muitas empresas, pois neste momento a empresa para de olhar para o passado e busca de forma mais intencional usar os dados para suas decisões. Esse tipo de análise, geralmente, lida com probabilidades e podem ser usados para prever alguns resultados, o que é chamado de *forecasting* ou também pode ser interpretado como uma simulação. Neste nível já se é muito utilizado padrões estatísticos, processos de aprendizado de máquina e diversas técnicas de modelagem. As perguntas principais neste terceiro nível são: O que irá acontecer? O que gostaria que acontecesse?

Análise Prescritiva - o quarto e último nível é onde grandes empresas orientada a dados se encontram, como Spotify, Amazon, Netflix entre outras. É o estágio mais avançado de

análise de dados, conforme apresentado pelo Gartner. A análise preditiva tem por objetivo influenciar ou calcular o melhor caminho para se conseguir uma vantagem competitiva, isso significa, de fato, orientar decisões com base em dados, conforme o famoso termo muito dito pelas empresas e pelo mercado "*Data-Driven*". Neste estágio, é buscado se fazer uma combinação de técnicas de análises preditivas com análises prescritivas, obviamente, com o auxílio avançado de técnicas de modelagem de dados, inteligência artificial e motores complexos de processamento e recomendação. A principal pergunta nesse estágio de maturidade é: Como podemos fazer isso acontecer?

A Figura 2.9 ilustra o que a empresa Gartner desenvolveu (GARTNER, 2019).

Figura 2.9 - Estudo Nível de Maturidade Gartner



Fonte: Gartner (2019)

3 ESTUDO DE CASO COM EMPRESAS

O objetivo deste estudo será entender como as empresas hoje estão atuando com seus projetos de dados, investigando desde o seu contexto, interação e realidade com dados, aprofundando suas metodologias, processos e times de dados, e por fim, compreender sobre cultura de dados, maturidade analítica e gargalos e desafios dos projetos. O seguinte Capítulo está organizado da seguinte maneira: a primeira Seção descreve a metodologia utilizada para realizar o estudo junto às empresas. A segunda Seção está a descrição detalhada das

entrevistas com as empresas. E a terceira Seção está um comparativo entre elas e uma análise consolidada do estudo.

3.1 Metodologia do Estudo de Caso

Para realização deste estudo, além da pesquisa abrangendo os principais pontos das metodologias existentes na literatura, realizou-se também uma pesquisa de caráter exploratório, bem como Gil (2007) recomenda, com o intuito de ter um maior aprofundamento nos desafios reais do mercado. Essa fase de estudos de caso buscou compreender a perspectiva de quem já teve experiências práticas no contexto de dados. Ao final das entrevistas, foram realizadas análises comparativas a fim de compreender como as empresas utilizam soluções de dados.

Esta fase exploratória de estudos de caso teve por objetivo um aprofundamento na realidade de cinco empresas que atuam com dados no seu dia a dia, investigando três principais temáticas. A primeira, no contexto de dados da empresa do seu entrevistado, qual sua relação com dados e o dia a dia da empresa com o tema. O segundo, no sentido do seu método, processo, passo-a-passo para lidar com projetos de dados no dia a dia. Além disso, entender também quais referências da literatura as empresas se inspiram hoje em dia e quem são as pessoas que estão envolvidas nesse processo. Por fim, a terceira temática, a cultura de dados da organização, os níveis de maturidade analítica do time e da empresa em geral, quais análises que hoje são realizadas e entregues e quais etapas do processo são os maiores desafios e gargalos.

Para as entrevistas foram convidadas 5 empresas buscando atender diferentes segmentos e realidades para explorar de forma diversa as suas realidades. Todas as entrevistas foram individuais, sempre conduzidas pelo autor e o critério de escolha foi a combinação de disponibilidade dos profissionais juntamente com a realidade de dados das empresas. Os setores das empresas que foram entrevistados foram: Varejo, Financeiro, Consultoria e Tecnologia. Além disso, também entre as empresas escolhidas para as entrevistas foi buscado trazer diferentes tamanhos de organização para explorar também o nível de maturidade de dados relacionado aos times e aos seus processos. Em geral, os convites foram feitos por e-mail e pelo LinkedIn e todos se mostraram muito abertos e receptivos para apoiar a pesquisa. O compromisso que o autor teve com todas as entrevistas realizadas era de que seriam

mantidos em sigilo os nomes das empresas e dos seus entrevistados. O tempo de entrevista variou de 40 minutos a 1 hora e 30 minutos.

As análises foram feitas na mesma estrutura em que foram formuladas as perguntas do roteiro das entrevistas. As entrevistas foram analisadas em 3 grandes blocos, buscando resumir e trazer os principais achados. Além disso, com o resultado dessa análise, em conjunto com o estudo e levantamento das metodologias da literatura, é proposto um método de descoberta de dados, com as informações relevantes levantadas neste estudo. Para guiar essas entrevistas, foi utilizado um roteiro semiestruturado (Anexo A) e tendo como base para comparação das empresas o método do *Data Analytics Lifecycle* (DIETRICH *et al.*, 2015). A escolha do método DAL foi somente pela preferência do autor.

Os principais tópicos que foram levados em consideração nas entrevistas foram os três grandes blocos:

- 1) Do profissional, da empresa, da sua relação com dados, de uma visão mais ampla do seu cenário.
- 2) Da literatura, das metodologias, do seu processo de dados atual, de quem são as pessoas que participam desse processo.
- 3) Das análises que são entregues pelos times, do nível de maturidade das mesmas e dos principais desafios e gargalos de projeto de dados.

3.2 Entrevistas

3.2.1 Empresa 1

1) Do profissional, da empresa, da sua relação com dados, de uma visão mais ampla do seu cenário.

O entrevistado 1 descreve o início da sua carreira e do seu contato com dados, ainda em um contexto de que não se tinha muita coisa sendo falado no Brasil, por volta de 2006. O *Business Intelligence* (BI) era algo que ainda era pouco falado. Até o momento em que foi

fazer um intercâmbio e teve contato com o mercado internacional que mudou muito a sua compreensão sobre dados.

Em 2019 entrou na atual empresa 1, uma das maiores varejistas do Brasil e rapidamente foi crescendo dentro da sua carreira na área de dados até atualmente ser responsável por um produto de remarcações que está dentro da área de analítica avançada na qual seu produto hoje entrega de resultado semanal em torno de R\$1,2 milhões de receita incremental.

A empresa está estruturada em uma diretoria e em três grandes áreas: A primeira, engenharia de dados, na qual tem a responsabilidade de controlar os acessos aos dados, liberar as permissões para quem pode e para quem não acessar as informações. Em resumo, é a área que faz o *compliance* da empresa cuidando da Lei Geral de Proteção de Dados (LGPD) e controle de acessos. A segunda, área de *Insights*, na qual é responsável por apoiar as demais áreas a terem ideias, possibilidades, análises sobre seus desafios. E por fim, a terceira, a área que o entrevistado 1 está alocado que é o *Data Lab*, na qual tem por objetivo criar produtos de analítica avançada, ou seja, produtos com inteligência avançada para apoiar e recomendar tomadas de decisão.

2) Da literatura, das metodologias, do seu processo de dados atual, de quem são as pessoas que participam desse processo.

O entrevistado 1 conta como conduz os seus rituais de trabalho com o seu time de remarcações. O time está composto por três diferentes grupos. O primeiro grupo composto por pessoas que atuam como *Product Owner* (PO), referência direta dos métodos ágeis e olhar de negócio dentro do produto que está sendo desenvolvido e por pessoas que atuam em *User Experience* (UX), ou seja, designers e profissionais que cuidam da experiência do usuário, eles buscam entender como esse produto que é desenvolvido (remarcações) é entregue, além de como deixar a experiência agradável, criar engajamento e adesão no mesmo. O segundo grupo é composto por pessoas que atuam como desenvolvedores, estes são responsáveis por desenvolver o produto de forma técnica, parte fundamental em projetos de dados. E o terceiro grupo por pessoas que atuam dentro do contexto de inteligência artificial, aprendizado de máquina, estes são responsáveis por desenvolver os algoritmos de análise avançada com processos de recomendações, *clustering*, *forecast* entre outros.

De forma concisa, o ritual do time se dá com base nas metodologias ágeis, o seu ciclo de projeto de dados é de 3 meses, ou seja, é o tempo que a empresa 1 entende ser suficiente para testar o que foi levantado dentro de um processo de descoberta de dados. Eles realizam um processo de descoberta, inspirado no CRISP-DM, no entanto, adaptado para a sua realidade. Nesta etapa de descoberta, o objetivo é entender os principais desafios que surgem dentro da área. Um exemplo seria se o produto está performando bem diante dos concorrentes.

A parte seguinte a se explorar o problema é começar a criar premissas que estarão associadas a testes que entrarão em um processo de validação. A ideia é ter um experimento para se realizar, pois é somente testando que será possível de se entender o seu impacto e sua relevância. Um ponto relevante que o entrevistado 1 aponta é que esta etapa inicial de descoberta de dados é algo que é constantemente revisado, pois o contexto de dados e a realidade da empresa 1 é muito dinâmica e não é possível somente em uma semana de imersão com o time todo que se consegue esgotar o problema.

Se comprovando o teste, o próximo passo é criar um MVP daquele teste e ele sendo validado, o objetivo é tornar ele um produto e levar para os usuários. Na área de remarcação é disponibilizado dentro da plataforma interna e após este momento é rodado um ciclo de feedback para entender como o usuário recebe este produto e como é possível melhorá-lo.

3) Das análises que são entregues pelos times, do nível de maturidade das mesmas e dos principais desafios e gargalos de projeto de dados.

Um grande desafio que o entrevistado 1 possui hoje na sua realidade é no que tange a cultura de dados da empresa como um todo. Apesar de ser uma grande corporação existem diferentes níveis de entendimento das pessoas sobre dados e isso impacta profundamente o seu trabalho com dados. O que ele tem visto atualmente é que existem pessoas que não entendem o estudo que é realizado para se entregar uma lista de recomendações, que foram desenvolvidos algoritmos de aprendizado de máquina, de inteligência artificial para que chegassem a estes resultados. O que fica claro é que apesar de se ter um time de analítica avançada, existe uma falta de conhecimento sobre dados. Além disso, essa falta de conhecimento, na visão do entrevistado 1, está associado ao quanto a empresa investe na cultura de dados. Um desafio atual é que por dados estar na moda, outras áreas estão

começando a utilizar, no entanto, começam sem saber exatamente o que querem estudar, o que querem resolver.

Disclaimer importante:

Um destaque importante é a área de remarcações que produz análises avançadas, busca entregar para os seus usuários (demais áreas do negócio) recomendações, algoritmos de aprendizado de máquina de forma gerar produtos para ser escalável e rentável para a organização. Existem outras áreas dentro da diretoria de dados que fornecem outros tipos de análises, sejam elas descritivas ou diagnósticas.

3.2.2 Empresa 2

1) Do profissional, da empresa, da sua relação com dados, de uma visão mais ampla do seu cenário.

O entrevistado 2 atualmente atua como coordenador do time de dados da empresa 2. Começou há dois anos e meio e é responsável por um dos produtos da organização. A empresa 2 atua com três principais produtos, um deles no ramo hoteleiro, um segundo em plataforma de relacionamento de clientes (CRM - *Customer Relationship Management*) e o terceiro de serviços de computação em nuvem. A organização atualmente chega a ter cerca de 350 a 400 funcionários.

O principal objetivo da área de dados hoje é criar e suportar processos e rotinas de integração dos dados dos clientes na plataforma. Recentemente foi feita uma mudança que fez com que a área de dados pudesse apoiar cada um dos produtos. A área como um todo eram 22 pessoas e 5 pessoas foram para o produto de hotelaria, 6 pessoas no time de CRM e os demais no time de computação em nuvem.

2) Da literatura, das metodologias, do seu processo de dados atual, de quem são as pessoas que participam desse processo.

O trabalho principal hoje do time de dados é mais voltado para consulta do que descoberta. O entrevistado 2 inicia sua fala dizendo que não existe hoje um processo de mineração de dados e de descoberta dentro da rotina de trabalho do seu time.

A demanda mais comum é chegar um direcionamento ou uma necessidade de alguma das áreas do negócio e a partir disso o time de dados faz o seu trabalho de desenvolver métodos e

acessar esses dados de maneira mais performática possível. Os papéis que compõem o time de dados do entrevistado são cinco analistas de banco de dados, dois deles já em um nível sênior, os demais em um nível pleno e não possui nenhum júnior na equipe.

Sobre os rituais do time, atualmente, estão em um processo de mudança para o modelo ágil, possui profissionais internos que estão apoiando a implementação do SCRUM, até então era um modelo sob demanda. Como coordenador, o entrevistado conta que seu papel também é apoiar a priorização das demandas que chegam buscando entender padrões e direcionar os esforços do time.

Já em questões técnicas, o core do sistema é o SQL Server, no entanto, com outras inúmeras tecnologias satélites ao redor conectados. As linguagens que mais são utilizadas são C# e Python para fazer automações e implementações em geral, para apoiar no tratamento e importação de grandes massas de dados de dados é também utilizado a linguagem R. As extrações seguem um formato padrão ETL (*Extract, Transform and Load*) e após isso é entregue para o "requisitante" deste trabalho a partir da plataforma. Essa disponibilização pode ser de duas formas. A primeira sendo dentro de um campo de notificações, existe um *pop-up* para quem recebe que a "solicitação chegou", como se fosse uma atualização e internamente na plataforma é possível acessar. Já a segunda forma é disponibilizando via FTP (*File Transfer Protocol*) para o usuário realizar o download. Algumas vezes, pelo volume de dados, pode acabar causando um *overhead* do sistema e é preferível disponibilizar por fora da plataforma.

3) Das análises que são entregues pelos times, do nível de maturidade das mesmas e dos principais desafios e gargalos de projeto de dados.

Já no contexto das análises o entrevistado 2 conta que hoje eles possuem um trabalho maior de consulta e entrega de estudos do que de elaboração de modelos ou processos de descoberta de dados. Portanto, não possuem pessoas que façam análises preditivas ou mesmo utilizem algoritmos de aprendizado de máquina ou de inteligência artificial (IA).

O entrevistado 2 compreende ainda que o nível de maturidade atual do time de dados é baixo quando comparado a um time que utiliza IA nos seus processos. Hoje, dentro da sua realidade, a etapa que se entende ser a mais desafiadora hoje é a parte de limpeza de dados, pois ser um processo robusto dentro que o time realiza para entregar informações.

3.2.3 Empresa 3

1) Do profissional, da empresa, da sua relação com dados, de uma visão mais ampla do seu cenário.

A entrevistada 3 possui já mais de 10 anos com atuação na área de dados, ela é graduada em estatística pela UFRGS e começou sua atuação profissional na área financeira, contexto muito similar ao qual atua hoje em dia. Já atuou em outros segmentos e está atualmente na empresa 3 há 6 anos.

Dentro da empresa onde atua, começou em uma área tratando muito de modelos estatísticos para cuidar da inadimplência. Após esse período, foi atuar em um contexto de dados mais geral com o intuito de apoiar o desenvolvimento de produto de dados. O contexto atual e dos últimos 2 anos e meio é um time focado em cultura analítica e busca ajudar os demais times de crédito nos seus desafios relacionados a dados, avaliando desafios que o time tem, propondo soluções ou até mesmo capacitando os colegas que atuam dentro do time para solucionar esses desafios e ajudar a orientar nas tomadas de decisão.

O objetivo do time de dados hoje é identificar uma forma de ler os dados que existem, captar e conectar com a realidade do negócio para conseguir resolver os problemas de negócio de forma mais precisa e com soluções melhores. Atualmente, o time que a entrevistada atua é de forma mais transversal apoiando o desenvolvimento de 2 produtos da organização e ela conta com 3 cientistas, 2 engenheiros e 5 analistas, estes últimos, tendo diferentes níveis de maturidade dentro do tema.

O tema desenvolvimento de pessoas dentro da área de dados é um aspecto que a entrevista 3 traz com bastante importância e é um ponto de atenção que a empresa 3 tem tido e tem sido feito um investimento considerável em formar novas pessoas.

2) Da literatura, das metodologias, do seu processo de dados atual, de quem são as pessoas que participam desse processo.

No contexto de metodologia, processos e passo-a-passo, a entrevistada 3 conta que a sua área de dados tem uma atuação proativa com o olhar de identificar problemas de negócio e oportunidades e levando para o time de dados validar, no entanto, em torno de 70% das

demandas hoje vêm do time de negócio. De modo geral, um processo natural é construir uma solução com o time de dados em conjunto com o time de negócio.

A entrevistada 3 conta que não existe uma forma padrão de requisitar um apoio da área de dados, ela surge de forma bem diversa e são dos mais diversos tipos de desafios que surgem, desde entender se um produto específico está precificado da forma correta, se o público que está sendo ofertado um determinado produto é o melhor, entre outros.

Dado o problema identificado que se precisa buscar uma solução, então, o time de dados realiza um processo de descoberta em conjunto com o time de negócios. O intuito desse processo é identificar as oportunidades, investigar o problema e até mesmo buscar uma construção de forma colaborativa do que é a visão que se tem do que se está querendo resolver. Em suma, é um grande alinhamento das áreas e do que precisa ser entregue.

O próximo passo é desenhar em conjunto um MVP (Mínimo Produto Viável) do que pode ser entregue, o que é factível o que não é. Então, ao final do processo se tem uma clareza do que seria uma possível solução, o que é o mínimo esperado, o que pode ser adicionado nas interações e validações com o time de negócio.

A demanda que foi gerada pela área de negócio ou pelo próprio time de dados vira um projeto que em geral tem duração de 2 meses, ou 8 semanas, e as interações entre os times de negócio e dados geralmente é feito 1 vez por semana. Durante o período do projeto, em geral, ficam 2 pontos focais do negócio, sendo estes os responsáveis por fazer as interações com o time de dados avaliando se o time está no caminho, se pode seguir, enfim, buscando em conjunto tomar as melhores decisões, caso seja necessário também, pivotar o projeto, será nestas interações que o time tem essa oportunidade.

A metodologia que hoje é inspirada para o time da entrevistada para investigar os seus desafios de negócio da área está baseado no CRISP-DM. Todavia, não é reproduzido de forma idêntica à metodologia. O time de dados em conjunto com o time de negócio adapta e incrementa outros fatores para o seu processo de descoberta de negócio, pois a forma como o CRISP-DM está desenhada não aborda todos os aspectos que entendem necessários. Os outros aspectos que eles investigam, são:

1. O que se quer entregar
2. Qual o problema que se quer resolver

3. Como se pode mensurar o resultado, quais indicadores iremos observar
4. Quais análises que se quer realizar

O time realiza esse processo atualmente com base em um Canvas com todos estes campos citados tendo como esqueleto base o CRISP-DM.

A entrevistada 3 conta que este processo de preenchimento auxilia muito na investigação do problema, na definição do foco do trabalho, no direcionamento dos esforços do time de dados. No entanto, não é um Canvas simples de ser preenchido, é necessário o apoio de todos nesse trabalho de investigação, pois muitas vezes nem o time de negócio tem tanta clareza do problema que está solicitando.

A entrevistada 3 traz um aspecto importante que também é um papel do time de dados em conhecer muito bem o negócio que ele está inserido porque aí ele consegue ser provocativo e direcionar o que está sendo construído, pois assim, dará um norte na tomada de decisão, sabendo qual indicador será analisado e que o time de dados consegue ter esse cuidado desde o início do projeto, de quais informações são relevantes para o problema que se está investigando.

3) Das análises que são entregues pelos times, do nível de maturidade das mesmas e dos principais desafios e gargalos de projeto de dados.

A entrevistada 3 conta que existe hoje uma grande oportunidade de desenvolvimento da cultura de dados na empresa em que atua. O conhecimento dos demais times da área de crédito em dados ainda é bem baixo, do ponto de vista de oportunidades, tomadas de decisão baseadas em dados, monitoramento de performance, ainda são pontos que podem evoluir muito.

Atualmente, a entrevistada entende que o principal gargalo do seu processo de dados é a etapa de entendimento de negócio, pois dentro do time de negócio existe um baixo conhecimento de dados e, por isso, geralmente, eles não possuem muita clareza do que estão buscando resolver. As demandas ainda chegam para o time de dados muito abertas, não se sabe com clareza quais informações são relevantes para o problema que se quer resolver e isso torna mais complexo o processo.

Esse principal desafio tem conexão, do ponto de vista da entrevistada, com uma baixa maturidade para tomada de decisão baseada em dados, ou seja, o processo fim-a-fim de dados,

desde a etapa de investigação até a tomada de decisão e, esta ser baseada em dados. O que não acontece atualmente, comenta a entrevistada, é o acompanhamento destas decisões, o processo de entender se a decisão realmente teve efeito ou não. Na opinião da entrevistada, para uma área de ciência de dados que apoia as áreas do negócio isso é algo que precisaria acontecer. Consequentemente, para o time de dados ainda é um desafio provar o seu valor, ser capaz de mostrar a partir de dados o quanto gerou de receita com o projeto X, quais decisão impactaram os produtos no projeto Y, qual foi o efeito das decisões que foram tomadas e quais foram seus efeitos.

Do ponto de vista de maturidade analítica, algo que é visto pela entrevistada é que o próprio entendimento sobre o que pode ser solicitado de uma área de dados ainda é muito superficial.

Em geral, surgem demandas que qualquer pessoa de qualquer área deveria saber realizar, no entanto, o time de dados assumido essas demandas, mas produzindo entregas mais robustas, com um olhar mais orientado à ciência de dados e não somente à produzir relatórios de análises descritivas, como havia sido solicitado. É dessa forma que a entrevistada conta que o time de dados que ela atua tem apoiado a mudança de mentalidade dos colegas sobre o tema de cultura de dados, aos poucos, conscientizando, mostrando o que pode ser realizado e as possibilidades, todavia, é um processo lento que demanda tempo e investimento da organização.

Atualmente, a capacidade do time de dados é de gerar análises prescritivas, dependendo do problema, mas ainda muito do trabalho do time fica no cenário descritivo. O time de dados já possui *cases* com projetos de inteligência artificial e aprendizado de máquina, algo que outras áreas já estão começando a solicitar para serem desenvolvidas em outros contextos da empresa. Um ponto importante de destacar é, como conta a entrevistada, o time de dados realiza as entregas e a responsabilidade fica por conta da área especialista, o que em algumas vezes, os projetos que atuam com análises prescritivas geram resistência, pois possuem recomendações distintas do padrão esperado, o que ela entende ainda ser um aspecto a ser desenvolvido da cultura de dados da organização.

3.2.4 Empresa 4

1) Do profissional, da empresa, da sua relação com dados, de uma visão mais ampla do seu cenário.

O entrevistado 4 começa contando a sua história na área de dados, a primeira memória é no ano de 1996, de uma forma informal, mas já começava o seu primeiro contato com dados. A partir do ano de 2000 começou a ser mais claro com relação ao seu cargo, pois era um analista de dados. No ano de 2009 já havia realizado alguns projetos de consultoria independente e, paralelamente a isso, também ministrava aulas em uma faculdade. Em 2016 já havia formado um time pequeno, entre sócios, gerentes, supervisores e o principal produto era trabalho de mídia para agência e análise de *dashboards* de *Business Intelligence* (BI).

Na pandemia, a empresa precisou se reinventar e para criar produtos a partir dos seus serviços, ela buscou desenvolver uma plataforma de análise de dados. Em resumo, o objetivo era automatizar o máximo de serviços que já realizavam e tornar isso replicável do ponto de vista de negócio.

Após isso, foi possível acoplar também algoritmos de aprendizado de máquina e modelar análises preditivas. Essa plataforma em 2021 ganhou prêmio de inovação em uma pesquisa realizada na área.

A empresa teve uma transformação enorme conforme foram transformando os serviços em uma plataforma. Basicamente o perfil da equipe era de publicitários, administradores e a equipe com a evolução da plataforma foi ficando cada vez mais próxima do perfil de profissionais de tecnologia. Hoje a maior parte da equipe são de desenvolvedores e estes conseguem fazer o relatório de *marketing*, de campanha, de contexto do negócio. A empresa possui grande parte do time de trabalho de forma remota por todo Brasil.

2) Da literatura, das metodologias, do seu processo de dados atual, de quem são as pessoas que participam desse processo.

No contexto de metodologia e processo de atuação nos dados, o entrevistado 4 explica como é realizado hoje a tarefa de começar um serviço do zero, com um cliente novo, por exemplo. Para se começar o processo, a etapa de entendimento de negócio, é importante que sejam feitas perguntas abertas buscando entender o que é importante no contexto do cliente. De fato, uma etapa de descoberta.

O objetivo do cliente é saber se o que a empresa 4 irá entregar irá gerar resultado para ela ou não. Portanto, é uma etapa importante para se ter uma metodologia e mapear possíveis automações, uma etapa de se realizar as perguntas certas.

Após o entendimento do negócio, a equipe técnica irá conectar a partir de APIs (*Application Programming Interface*) as principais ferramentas de mídias sociais, por exemplo, Google Ads, Facebook, Google Analytics, buscando fazer uma importação dos dados do cliente. Após importado, é realizado um tratamento padrão. Inclusive, recomenda uma organização dos dados e realiza o *track* do URL. A saída destes dados é ir para um *Data Warehouse* (DW) que após essa etapa, a saída pode ser direcionada para um painel dos principais indicadores e isso é disponibilizado para o cliente, ou, então, os dados deste DW vão para um serviço de modelagem estatística que alimenta uma ferramenta de aprendizado de máquina que fazem análise de predição.

A partir destes modelos estatísticos é gerado um MVP, pois assim é possível rodar um teste baseado no modelo desenvolvido e outro teste no formato que o cliente sempre realizou. No final será possível entender qual deu mais resultado. O entrevistado conta que, em geral, o tempo de um projeto de dados já sendo possível perceber resultados é de 3 meses. Como inspiração da literatura, o entrevistado, por ser professor e doutor, desenvolveu uma metodologia do índice de maturidade digital, como mostra a sua pesquisa.

3) Das análises que são entregues pelos times, do nível de maturidade das mesmas e dos principais desafios e gargalos de projeto de dados.

No contexto de cultura de dados, o entrevistado comenta que existe ainda uma grande oportunidade de se explorar esse tema nas empresas de modo geral. Os diferentes níveis de maturidade mostram muito sobre sua cultura de dados e a importância de se calibrar as entregas que são feitas. Na opinião do entrevistado, está muito relacionado com o momento de carreira que os líderes das empresas vivem, se o profissional não entendeu que precisa implementar um viés de experimentação dentro da empresa. Então, não adianta se ter uma análise robusta, pois quem consome essa informação não entendeu ainda o que significa na prática.

A etapa mais desafiadora para o entrevistado é a etapa de consolidação dos dados, pois ter um problema de negócio bem definido, ter um escopo claro, ter consistência em quais dados são relevantes é o que faz um projeto de dados dar certo na sua opinião.

As análises em geral que realizam são desde Dashboard de indicadores, há também uma parte de relatoria dentro da plataforma que alguns clientes utilizam. As demais análises são preditivas e modelos estatísticos.

3.2.5 Empresa 5

1) Do profissional, da empresa, da sua relação com dados, de uma visão mais ampla do seu cenário.

O entrevistado 5 conta como começou a sua carreira e sua relação com dados. Entrou de estagiário em um grande banco e posteriormente foi atuar na área de estratégia e prevenção à fraudes do mesmo. Pela oportunidade de interagir com times do banco de fora do Brasil, foi possível ter um profundo conhecimento desde cedo da carreira. Foi aprender sobre prevenção à fraude de cartões aplicado a vários portfólios do banco e por ser *Big Data* eram questões técnicas e relacionadas inclusive à performance.

Depois entrou na empresa 5, em 2019. Foi atuar na área de fraudes da empresa. Era uma movimentação de ir para uma empresa considerada unicórnio no meio de startups e já havia um conhecimento técnico profundo desenvolvido nos tempos de banco.

A empresa 5 já existe há 10 anos, uma empresa que nasceu em Curitiba, possui cerca de 1400 funcionários e é uma processadora de pagamentos, recentemente adquiriu algumas empresas. A estrutura da empresa 5 respeita uma diretoria de dados e se segmenta em áreas de dados nas demais da empresa. O entrevistado conta que o seu time na empresa era formado por três especialistas e um cientista de dados sênior.

2) Da literatura, das metodologias, do seu processo de dados atual, de quem são as pessoas que participam desse processo.

No processo de desenvolvimento de projeto de dados, o entrevistado 5 conta que sempre que se precisava tomar uma decisão macro do desafio, no sentido de qual caminho seguir, qual *feature* priorizar era sempre realizado em conjunto com o time de negócio. O processo que o entrevistado conta está relacionado aos rituais de trabalho dos times de dados. Cada projeto que estava acontecendo estava também sendo acompanhado por pessoas do negócio.

No entanto, algo que o entrevistado afirma é que o processo ideal de fim-a-fim era difícil de acontecer como recomenda a literatura. O início do trabalho do time era feito em conjunto nos rituais de planejamento e priorização do trabalho, no entanto, com o passar do tempo, iam diminuindo as interações com o time de negócio.

Um grande desafio que o entrevistado comenta é a importância do perfil profissional de dados, pois na sua opinião ele precisa ser verdadeiramente técnico, por mais que realize tarefas operacionais ou de desenvolvimento de código, mas, entenda quais perguntas realizar, entenda das relações e impactos da tecnologia e como direcionar os esforços da melhor forma. Um desafio é o da comunicação, conseguir se comunicar com o time de dados, além de conseguir também gerar uma ponte com o time de negócio.

Algo trazido pelo entrevistado como muitas vezes considerado pouco importante em projetos de dados é sobre as métricas. Na sua opinião, é fundamental mensurar desde o dia zero, pois é em um aspecto de experimentação e teste e saber se está sendo feito um trabalho certo ou não.

Na importância de gerar resultados diretos para a empresa com os times de dados, o entrevistado traz uma perspectiva de como mensurar os dados e observar o impacto financeiro do projeto. Na sua opinião, é importante observar se o projeto de dados que está sendo realizado está gerando uma receita ou reduzindo um custo, pois assim fica claro o impacto no negócio.

3) Das análises que são entregues pelos times, do nível de maturidade das mesmas e dos principais desafios e gargalos de projeto de dados.

A cultura de dados da empresa 5 era algo bem profundo, na opinião do entrevistado, apesar de perceber que existia conhecimento de dados, mas onde muitas soluções eram impactadas era no contexto de politicagem, burocracia ou disputas políticas de poder que as áreas de dados geravam.

Já no contexto das análises, possuíam dos mais variados tipos, indo desde acompanhamento de performance por painéis de indicadores até algoritmos de aprendizado de máquina e inteligência artificial. Considerando a experiência do entrevistado 5, ele comenta que o seu entendimento para área de dados que buscam soluções para o negócio em que atuam não devem focar necessariamente nas soluções que os projetos de dados geram e sim em

identificar oportunidades de negócio, se o time de dados não existir, o core do negócio continua a existir.

3.3 Comparação das empresas

Esta Seção apresenta uma análise consolidada das entrevistas buscando destacar o que foi de mais relevante trazido pelas empresas com o objetivo de ter uma visualização geral e poder realizar uma análise comparativa com a metodologia da literatura do *Data Analytics Lifecycle* (DIETRICH *et al.*, 2015).

Tabela 3.1, como mostra abaixo, busca sintetizar as principais informações que foram exploradas pelas empresas entrevistadas. A tabela respeita os mesmos 3 blocos que foram explorados durante as conversas.

O bloco 1, primeiro grupo de perguntas, deixa evidente os segmentos, o porte do tamanho das empresas - este categorizado pelo critério do IBGE, na qual são consideradas empresas grandes com mais de 100 colaboradores, média de 50 até 99 colaboradores, pequena de 11 até 49 colaboradores e micro de 1 até 9 colaboradores.

O bloco 2 sintetiza as principais informações que competem ao processo metodológico das empresas, em geral, as empresas respeitam os rituais dos métodos ágeis para os seus ciclos de desenvolvimento de projetos. Outro ponto que é possível perceber é com relação ao seu tamanho de organização em comparação à sua maturidade em projetos de dados. Nesse sentido, foi possível perceber que, em geral, as empresas realizam uma etapa inicial, descrita nos processos da literatura como a etapa de entendimento de negócio, no entanto, apenas uma empresa não realiza, pois a atuação do time de dados é mais sob demanda. O que evidencia a sua baixa necessidade de explorar soluções mais robustas com os dados. Da mesma forma, a mesma empresa que não realiza a etapa de entendimento de negócio, também não gera hipóteses e MVP de dados, pois, como mencionado, suas soluções são mais específicas e, portanto, não havendo necessidade de tal.

Como mencionado anteriormente, foi utilizado como método de comparação a metodologia de *Data Analytics Lifecycle* de Dietrich et al (2015) para comparação com o método das empresas. De modo geral, as empresas respeitam as etapas propostas pela metodologia escolhida, exceto a empresa já citada que não realiza a etapa de entendimento de negócio.

E no contexto do bloco 3, a partir da ótica de maturidade analítica proposta pelo Gartner, fica evidente as diferenças de realidade de dados dentro da organização.

E por fim, um ponto unânime de todas as empresas é com relação à cultura de dados. Apesar de algumas empresas possuírem grande expertise de dados, produzir análises e soluções robustas, a oportunidade de explorar e desenvolver a cultura de dados é algo que todas citaram como grande necessidade de desenvolvimento, pois o entendimento de dados muda muito de pessoa para pessoa e, conseqüentemente, a capacidade de interpretações sobre os dados também.

Tabela 3.1 - Quadro comparativo da análise das tabelas

Empresas	Empresa 1	Empresa 2	Empresa 3	Empresa 4	Empresa 5
<i>Bloco 1</i>					
Porte da Empresa	Grande	Grande	Grande	Pequena	Grande
Segmento	Varejo	Tecnologia	Financeiro	Consultoria	Financeiro
Número de funcionários	~ 24 mil	~ 400	~ 30 mil	-	~ 1400
<i>Bloco 2</i>					
Realiza descoberta de dados?	Sim	Não	Sim	Sim	Sim
Realiza rituais de acompanhamento com negócio?	Sim	Sim	Sim	Sim	Sim
Com qual frequência realiza acompanhamento?	15 dias	15 dias	Semanal	15 dias	15 dias
Utiliza métodos ágeis?	Sim	Sim	Sim	Sim	Sim
Gera MVP de teste?	Sim	Não	Sim	Sim	Sim
Qual metodologia da literatura tem como base?	CRISP-DM	-	CRISP-DM	Índice de Maturidade Digital	-
Quantas etapas do Data Analytics Lifecycle (DAL) realiza?	Todas	5 etapas	Todas	Todas	Todas
Tempo de projeto de dados?	3 meses	-	2 meses	3 meses	3 meses

Número de pessoas na equipe?		5	10	-	4
<i>Bloco 3</i>					
Nível de maturidade das análises Gartner	Nível 4 - Prescritivos	Nível 2 - Diagnóstica	Nível 4 - Prescritivas	Nível 4 - Prescritivas	Nível 4 - Prescritivas
Principal gargalo com base no DAL	Entendimento de negócio	Preparação dos dados	Entendimento de negócio	Entendimento de negócio	-

Fonte: Elaborado pelo autor

4 APLICAÇÃO PRÁTICA EM UM CASO REAL

Este Capítulo tem por objetivo recomendar um Canvas de Descoberta de Dados, sua criação foi baseada em um estudo prévio realizado pela autora Karina Moura (2018), em experiências profissionais, além do mapeamento realizado da literatura e na atual pesquisa feita com as empresas.

A aplicação deste Canvas tem por objetivo ser um protótipo e busca validar o processo que se recomenda neste trabalho aplicando em uma empresa real buscando entregar, ao final, um produto de dados para a empresa.

Este Capítulo está organizado da seguinte maneira: a primeira Seção será apresentado o Canvas de Descoberta de Dados que foi desenvolvido neste estudo. A segunda Seção será apresentado desde a aplicação dele na empresa até o desenvolvimento de um produto de dados, que será um painel de BI. Por fim, a terceira Seção, apresentará os principais achados de toda a pesquisa como ideias, pontos importantes e conclusões do estudo.

4.1 Canvas de Descoberta de Dados

O Canvas de Descoberta de Dados foi baseado no *Canvas Data Analytics*, como mencionado no Capítulo 2.2 deste estudo. Este Canvas foi publicado pela colega Karina Moura (2018) em seu Trabalho de Conclusão de Curso orientado pela professora Dra. Daniela F. Brauner.

O Canvas de Descoberta agrega diferentes fatores para ser sugerido. Ele possui uma sequência lógica sugerida e sendo primeiro passo é definir o problema de negócio, insumo principal vindo das entrevistas e da literatura. Ele também considera que existem duas

dimensões na forma como ele está elaborado. O centro sendo a parte dinâmica e já em uma perspectiva cíclica. As laterais, sendo como pilares, trazendo uma perspectiva de aspectos mais estruturantes, do lado esquerdo pilares estratégicos como investimentos, prazos, riscos. Do lado direito, uma perspectiva de pessoas, sendo o campo de Cultura Analítica um incremento importante que o atua Canvas de Descoberta de dados trás. Como resultado desta análise, a Figura 4.1, este é o Canvas de Descoberta de Dados que é sugerido neste estudo.

Figura 4.1 - Canvas de Descoberta de Dados



Fonte: Elaborado pelo autor

A fim de aplicar na prática um protótipo que possa ser usado de maneira genérica por qualquer tipo de empresa, que ajude a explorar de forma mais assertiva os problemas de negócio e apoio o desenvolvimento de uma solução de dados, foi elaborado um Canvas com onze etapas distintas.

Estas etapas seguem uma linha lógica e ordenada, que deve ser preferencialmente conduzida pelo cientista de dados (ou por outro facilitador, preferencialmente uma pessoa neutra ao desafio que se visa resolver). Na Tabela 4.1, é possível visualizar todas as etapas dessa metodologia, bem como as perguntas norteadoras que servem para instigar o time de trabalho na etapa de *brainstorm* das possibilidades, bem como exemplos que podem auxiliar na condução do processo.

Tabela 4.1 - Descrição das etapas do Canvas de Descoberta Dados

Nº da etapa	Etapa	Perguntas norteadoras	Exemplos
1	Desafios	Qual o problema de negócio a ser resolvido?	Aumentar a conversão de clientes, elaborar uma campanha de marketing para um produto, identificar o tipo de projeto mais rentável, prospectar o mercado de forma ativa.
2	Captação de dados	Quais são as fontes de dados que nos ajudam a resolver esse problema?	E-mails, históricos de projetos, bases de dados disponíveis, sistemas internos, Google Analytics.
3	Perguntas de Negócio	Quais são as perguntas que temos para esse problema que queremos resolver?	Quem é o público ideal, sazonalidade das compras do produto, qual o valor investidos em anúncio para cada produto,
4	Piloto	Como rodar um MVP e partir dessas hipóteses?	Dashboard simplificado
5	Métricas de sucesso	Como eu conseguiria mensurar o resultado?	Aumento de conversão em 10% para um produto, alcance de 20% maior,
6	Riscos	Com o que devemos ter atenção que pode atrapalhar nosso projeto?	O aumento da demanda ser maior que a empresa pode atender,
7	Prazo	Em quanto tempo conseguimos desenvolver o ciclo deste projeto?	3 meses de campanha,

8	Investimentos	No que será necessário investir para que essa solução ocorra?	Campanhas de anúncios pagos, investimento em treinamento,
9	Cultura de dados	O quão familiarizado estão com dados?	Familiaridade com dados e capacidade analítica alta, média ou baixa.
10	Público-alvo	Quem são as personas que irão consumir essa informação?	Time de marketing, gerentes, equipe de TI...
11	Pessoas	Quem pode nos ajudar (Stakeholders, usuários, colaboradores) ?	Time de marketing, gerentes, equipe de TI...

Fonte: Elaborado pelo autor

O Canvas consiste em uma parte central, que abrange as etapas de 1 a 5, que são etapas fundamentais e podem ser repetidas de maneira cíclica. As etapas 6, 7 e 8, bem como 9, 10, 11 servem como “pilares” que sustentam o ciclo central, conforme mostrado Tabela 4.1.

Essa configuração das etapas pretende mostrar que os elementos fundamentais estão contidos nas etapas de 1 a 5, pois é a partir delas que é possível aplicar um MVP (*Minimum Viable Product*) com o apoio dos elementos tangentes (pilares).

4.1.1 Ciclo de dados – Etapas 1 a 5

A Etapa 1 é o momento em que se elencam **os desafios** que podem ser melhor identificados e resolvidos a partir de uma solução de dados. Essa etapa pode compreender um momento de *brainstorm*, porém é importante escolher apenas um desafio para seguir o preenchimento do Canvas. A pergunta norteadora para essa etapa é “Qual o problema de negócio que precisa ser resolvido?” e, dessa forma, a questão mais latente a ser resolvida no time pode ser votada ou selecionada em comum acordo. Para instigar o time no *brainstorm* de cada uma das etapas do Canvas, exemplos são trazidos na Tabela 4.1 que podem servir como complemento às perguntas.

Na Etapa 2, é o momento de pensar quais **fontes** podem fornecer **dados** para o time resolver o desafio com base neles. Alguns exemplos são: bases de dados externas disponíveis, bases de dados internas, e-mails, planilhas com históricos de projetos etc.

A Etapa 3 tem o intuito de levantar **as perguntas de negócio** que podem ser úteis para resolução do problema, ou seja, quais as perguntas que se tem – ou se deve ter – para solucionar o desafio em questão. Alguns exemplos de perguntas de negócio podem ser: quem será o público, a sazonalidade de um produto, o investimento que se tem em marketing para determinado produto etc. Uma perspectiva que pode ajudar nessa etapa é pensar quais são os filtros ou balizadores que podem ser levantados para contribuir na manipulação destes dados, visando sempre uma análise que possa indicar uma solução. Para avançar para a próxima etapa, é importante selecionar qual(is) pergunta(s) de negócio pode ser mais facilmente testada.

Na Etapa 4, a ideia é que se vislumbre como pode ser a versão piloto ou o protótipo para essa solução. Um dashboard com dados simples relacionados ao desafio pode ser o piloto ideal, que pode ser designado também como um produto mínimo viável, ou seja, um MVP.

Por fim, o ciclo de dados termina com a determinação das **métricas de sucesso** para o desafio que se pretende resolver – o que compreende a Etapa 5. A pergunta norteadora dessa etapa é "Como é possível mensurar o resultado para o desafio?". Alguns exemplos de métricas de sucesso são: aumento da conversão de clientes em 10%, aumento do alcance de campanhas de marketing de um produto em 15%, etc.

Como dito anteriormente, essas são as etapas fundamentais do Canvas e que podem ser repetidas de maneira cíclica dentro de um ciclo de projeto, ou seja: define-se o desafio (Etapa 1), e segue-se nas demais etapas até a obtenção de métricas de sucesso que se relacionem de forma satisfatório com o desafio. Por isso, também, que o desafio selecionado pelo time de trabalho na Etapa 1 não necessita de muita especificidade, pois isto se atinge na Etapa 5, tornando o desafio uma meta específica.

4.1.2 Pilares de sustentação – Etapas 6 a 11

Na Etapa 6, é o momento em que o time elenca **os riscos** do projeto de dados, que abrangem qualquer ponto de atenção que possa influenciar o projeto. O aumento da demanda de um determinado produto decorrido pelo aumento em marketing, por exemplo, pode ser um risco em que a empresa não está preparada para atender.

A Etapa 7 compreende o **prazo** total em que se deseja rodar o ciclo de projeto. É comum que as empresas optem por ciclos trimestrais, em que é possível rodar diferentes pilotos (Etapas 1 a 5) até se chegar em uma solução mais refinada e aderente ao desafio, com base em tentativa e erro. O prazo que se define nesta etapa é, portanto, o prazo total (macro) do ciclo de projeto, e este pode variar conforme disponibilidade e necessidade dos times. A pergunta norteadora dessa etapa pode ser "Por quanto tempo nosso time pode trabalhar nesse desafio?".

Na Etapa 8, definem-se **os investimentos** necessários para que se encontre uma solução de dados. Exemplos comuns são os investimentos em campanhas de marketing e em treinamento de colaboradores para que possam aumentar a sua familiaridade com dados, entendendo contexto de negócios e quais dados se relacionam a ele.

A Etapa 9 é o momento de se avaliar, de maneira mais subjetiva, como está a **familiaridade**, experiência ou mesmo capacidade analítica que o time tem em relação a dados. Uma pergunta norteadora possível é: o time sabe quais dados são fornecidos continuamente e cuja análise podem ajudar a resolver desafios de negócio? Para ajudar nessa avaliação, sugere-se uma lógica de semáforo, em que a cor o verde representa um nível alto de maturidade de dados; amarela, um nível médio; e a cor vermelha, um nível baixo.

Na Etapa 10, é o momento de pensar quem é o **público-alvo** desse projeto, ou seja, quase são as pessoas que irão receber esses dados e que, a partir deles, podem tomar melhores decisões de negócio. Alguns exemplos são o time responsável pelo desafio em questão, papéis de liderança (como gerentes), etc.

Por fim, a Etapa 11 compreende uma avaliação das **pessoas** que se relacionam de alguma forma com o projeto, como stakeholders, usuários, clientes ou colaboradores. Uma pergunta norteadora para essa etapa é "Como envolver as pessoas que se relacionam com o desafio?"

4.2 Protótipo

O processo realizado para aplicar o protótipo foi o seguinte: como primeiro passo foi a definição do modelo do Canvas, exemplificado na Seção anterior. Como segundo passo, foi definido a empresa que seria a parceira no experimento. Em reunião online, foi conversado com o gerente comercial comentando do desafio do trabalho de conclusão de curso e já realizando um primeiro convite para fazerem parte do processo. O passo seguinte foi realizar a condução e preenchimento do Canvas com a empresa. Logo após, foi recebido as bases de

dados, aplicado o ETL e por fim, aplicado visualizações de dados no *PowerBI*. A proposta agora é detalhar cada um destes passos que foram realizados.

Antes de realizar a aplicação do Canvas com o time de projeto, é importante que o cientista de dados – ou facilitador – tenha um momento de *briefing* sobre o(s) desafio(s) que estão emergentes e podem ser resolvidos a partir de análise de dados, para que se tenha um maior entendimento de contexto: dos desafios que se apresentam e, também, de oportunidades de melhoria. Outro ponto fundamental é salientar a importância de gerar e analisar dados para melhorar a tomada de decisão dos times, processo conhecido como *Data-Driven*, ou seja, orientado por dados. Desta forma, é possível tomar decisões mais assertivas e aderentes ao contexto de negócio. A fim de testar e validar a metodologia de Canvas de Descoberta de Dados desenvolvida neste trabalho, foi realizada um encontro facilitado pelo autor para aplicação do Canvas.

A empresa convidada para participar do estudo é do ramo de Consultoria Ambiental, e o time selecionado foi o time comercial, com o total de três colaboradores participando no encontro. O time atualmente gera dados a partir de sistemas internos, planilhas, e-mails, porém não tem um direcionamento de como sistematizar e analisar os dados que gera. Em termos de Cultura de Dados, que é a Etapa 9 do Canvas, o time se considera com um nível médio de familiaridade com dados, pois gera dados, entende a importância deles para tornar o dia a dia do trabalho mais assertivo, porém não consegue ter uma visualização clara do que esses dados podem estar apontando em termos de tomada de decisão estratégicas.

Na Tabela 4.2 foram sintetizados os principais *insights* do encontro, abrangendo desde o desafio escolhido pelo time para ser resolvido pelo projeto de dados, bem como os principais pontos relevantes das outras etapas.

Tabela 4.2 - Resultados obtidos no encontro facilitado do Canvas de Descoberta de Dados

Nº da etapa	Etapa	Brainstorm do time
1	Desafios	Ser mais assertivo no processo de venda de projetos, aumentando o número de propostas fechadas
2	Captação de dados	Dados públicos, CNPJ de empresas cadastradas, e-mails, sistema interno da empresa, formulário

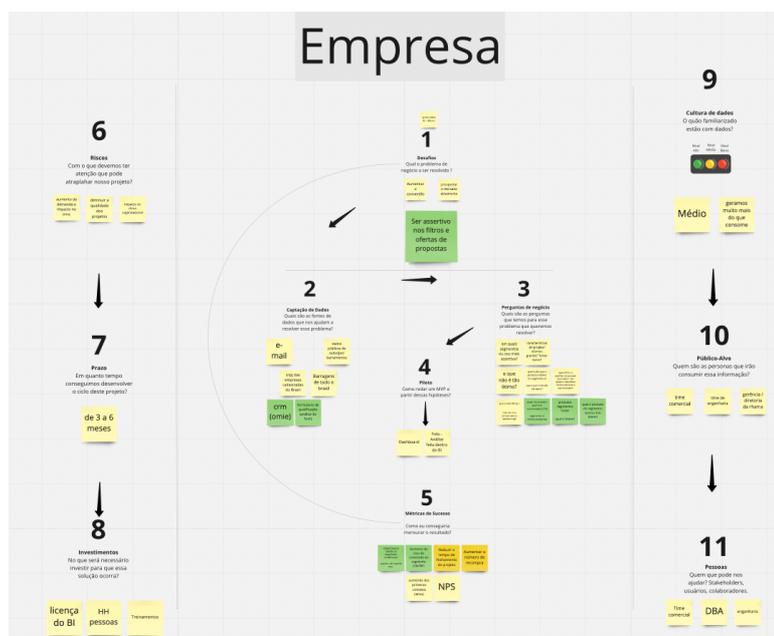
		de qualificação dos clientes.
3	Perguntas de negócio	Quais os tipos de projeto de maior recorrência? Quais produtos/segmentos vendemos mais? Qual o produto/segmento no qual a empresa tem um diferencial/se torna referência?
4	Piloto	Dashboard, análise feita dentro do BI
5	Métricas de sucesso	Reduzir taxa de rejeição, aumento da taxa de conversão no segmento X, NPS, taxa de recompra.
6	Riscos	Aumento excessivo da demanda, diminuição da qualidade dos projetos, impacto no clima organizacional
7	Prazo	De 3 a 6 meses
8	Investimentos	Licença do BI, horas de trabalho do time, treinamentos.
9	Cultura de dados	Nível médio (gera-se muito mais dados do que se analisa)
10	Público-alvo	Time comercial, gerência/diretoria
11	Pessoas	Time comercial, DBA

Fonte: Desenvolvido pelo autor

Ao final do encontro, foi feita uma rodada de *feedbacks* com o time em relação a como foi a experiência com o Canvas. Em termos gerais, o grupo achou bastante clara e assertiva a lógica proposta pelo Canvas, com a um passo a passo coerente dos pontos-chaves que devem ser considerados para buscar uma solução de dados. Um dos colaboradores pontuou que a metodologia do Canvas propõe não somente uma abordagem técnica para se iniciar um trabalho de dados com o time, mas também propõe uma mudança de mentalidade que fomenta a cultura de dados de, por exemplo, "**Como os dados podem nos ajudar a solucionar desafios de negócio?**". Isso auxilia em uma mudança de cultura, tornando o time mais analítico e orientado a buscar soluções com base no que realmente acontece no dia a dia.

Outro ponto importante que surgiu foi a importância de se ter um facilitador para conduzir as etapas do Canvas. O time achou de extrema importância o papel de um facilitador que entenda sobre o tema de dados para fazer as provocações certas, conduzindo o time com as perguntas pertinentes para se encontrar uma solução de dados que faça sentido, ou mesmo que sirva como um primeiro piloto. Nesse sentido, fica evidente a importância do papel de um cientista de dados para atuar nesse tipo de projeto: não basta analisar, refinar e entregar dados, e sim conversar com o time e saber conduzi-lo na melhor experiência que contemple os pontos mais sensíveis do contexto de projeto. Além disso, é importante levar em consideração que um cientista de dados alinhado com o negócio torna a solução dos desafios muito mais assertiva. A Figura 4.2 mostra o resultado do Canvas preenchido em conjunto com a empresa.

Figura 4.2 - A representação do Canvas preenchido



Fonte: Desenvolvido pelo autor

4.3 Aplicação do ETL (*Extract, Transform, Load*)

Após a aplicação e facilitação do Canvas de Descoberta de Dados, para ser possível realizar o estudo de dados mapeado na aplicação, a empresa disponibilizou um base de dados CRM, do período de 04/01/2021 até 28/03/2022, com dados da área comercial. Esta data final foi utilizada visto que houve uma migração de plataforma de gestão comercial na empresa. A base não era muito robusta, a tabela continha 240 linhas de oportunidades, como é

chamado internamente, e 41 colunas de informações diversas do seu processo comercial. O arquivo foi disponibilizado via arquivo Excel.

Após o recebimento da base (arquivo Excel), foi necessário passar por um processo padrão de ETL (Extração, Transformação e Carregamento), que significa, em termos técnicos, preparar os dados para poder fazer as análises. Esse processo de ETL finaliza na construção do *Data Warehouse*, que é o local que fica todo o armazenamento dos dados, e ele possui duas qualidades fundamentais: ter dados formatados consistentes e serem facilmente acessíveis.

Essa etapa do ETL é uma das partes mais importantes de todo o processo de manipulação dos dados. Uma ETL bem realizada faz com tenhamos uma boa *Data Warehouse*, e possibilita que tenhamos um bom *Business Intelligence* (BI).

Para facilitar a compreensão da base de dados que foi fornecida pela empresa, a Tabela 4.3 mostra o dicionário de dados utilizado.

Tabela 4.3 - Dicionário de dados do CRM utilizado nas análises do protótipo

Dicionário de Dados			
Nome do campo	Descrição	Tipo de dado	Valor
Solução	Campo que descreve o serviço que a empresa realiza para os seus clientes	Texto	-
Valor dos serviços	Campo que representa o valor financeiro em reais dos produtos e serviços da empresa	Numeral - Moeda	-
Descrição da Vertical	Campo que descreve o segmento de atuação que o cliente atua	Texto	-
<i>Status</i>	Campo que simboliza o <i>status</i> da oportunidade para a empresa.	Texto	Ativo: Oportunidade aberta Conquistado: Oportunidade ganha (vendida) Perdido: Oportunidade perdida (não vendida)

Fonte: Desenvolvido pelo autor

O processo de extração foi realizado ao obter a base de dados fornecida pela empresa. Em outros exemplos, a parte de extração poderia contar com sites da internet, diferentes arquivos

de texto, diferentes tabelas etc. A etapa de extração busca trazer para o mesmo lugar toda e qualquer base de dados que tenhamos mapeado no Canvas.

Na parte de transformações dos dados, em virtude de CRM não ser tão robusta, foi necessário realizar poucas transformações e não foi necessário normalizar nenhuma coluna. Foram feitas 4 transformações a partir da ferramenta do *PowerBI*. As transformações foram acertar as colunas da base de dados para serem os campos principais da tabela e as demais foram nas 3 colunas específicas que foram utilizadas para as análises deste estudo. Para este estudo era o suficiente, no entanto, este processo sendo estruturado dentro da organização, possivelmente outras transformações e demais normalizações seriam necessárias a fim de deixar uma solução robusta e preparada para demais análises e usos futuros.

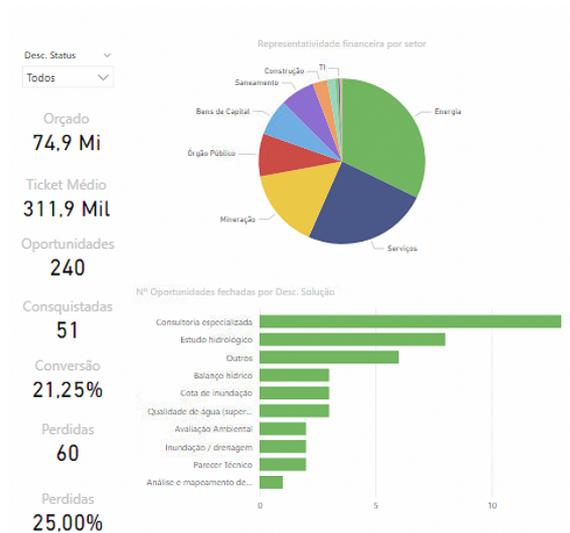
Ressalvas:

- 1) A base fornecida possuía alguns dados não preenchidos. Esse seria o exemplo de uma base que possui um grau de ruído (sujeira). Ademais, foi comentado pelo gerente comercial da empresa que no início do 2021 havia sido feito uma migração de banco de dados interno e esse processo causou inserção de alguns dados não consistentes.

4.4 Organização do BI

A ferramenta que foi utilizada para realizar as análises foi o *PowerBI* da Microsoft (MICROSOFT). O BI será o produto de dados fornecido como entrega para a empresa. A última etapa do ETL é a o momento de *Load*. Para isso, foi desenvolvido para o estudo um painel de visualização de dados que possui duas partes e está representado na Figura 4.3 e Figura 4.4. A organização do BI está disposta da seguinte maneira: na parte 1 (à esquerda), é demonstrada uma lista dos principais indicadores do time comercial. Alguns destes indicadores foram levantados em conjunto com todos na aplicação do Canvas.

Figura 4.3 - Painel Parte 1 das análises da Empresa



Fonte: Desenvolvido pelo autor

A Tabela 4.4 explica o significado dos principais indicadores e como foram calculados.

Tabela 4.4 - Principais indicadores da área comercial calculados

Organização do Painel			
Nome do campo	Descrição	Como é calculado	Valor
Descrição Status	Campo que simboliza o status da oportunidade para a empresa.	Campo texto	Ativo: Oportunidade aberta Conquistado: Oportunidade ganha (vendida) Perdido: Oportunidade perdida (não vendida)
Valor Orçado	Valor total dos serviços orçados pelo time comercial para clientes	$\text{Orçado} = (\text{SUM}(\text{'Oportunidades'}[\text{Vlr. Serviços}]))$	Numeral - Moeda
Ticket Médio	Valor médio dos projetos vendidos pela empresa	$\text{Ticket Médio} = (\text{SUM}(\text{'Oportunidades'}[\text{Vlr. Serviços}])) / ([\text{Qtd. Oportunidades}])$	Numeral - Média
Oportunidades	Número total de oportunidades do time comercial	$\text{Qtd. Oportunidades} = \text{COUNTA}(\text{Oportunidades}[\text{Cód. Oportunidade}])$	Número inteiro
Conquistadas	Número de oportunidades vendidas pelo time	$\text{Nº Oportunidades fechadas} = \text{CALCULATE}([\text{Qtd. Oportunidades}], \text{'Oportunidades'}[\text{Desc.}])$	Número inteiro

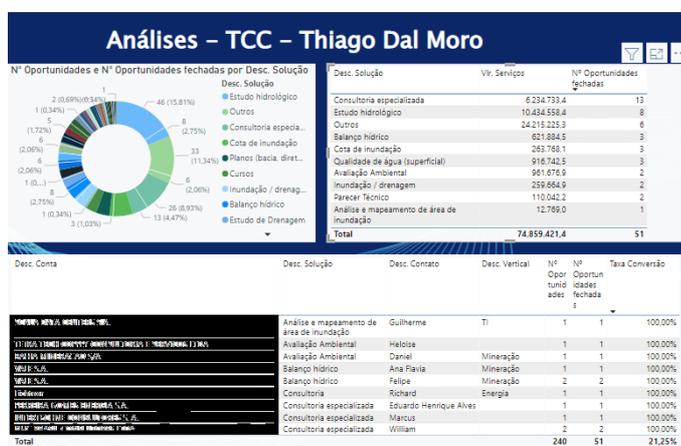
	comercial	Status]="Conquistado")	
% Conversão	Número que representa a taxa de conversão de conquistas da empresa	Taxa Conversão = $[\text{N}^{\circ} \text{ Oportunidades fechadas}] / [\text{N}^{\circ} \text{ Oportunidades}]$	Percentual
Perdidas	Número de oportunidades perdidas pelo time comercial	$\text{N}^{\circ} \text{ Oportunidades fechadas perdida} = \text{CALCULATE}([\text{Qtd. Oportunidades}], 'Oportunidades'[\text{Desc. Status}] = "Perdido")$	Número inteiro
% Perdido	Número que representa a taxa de conversão de oportunidades perdidas da empresa	Taxa de perdas = $[\text{N}^{\circ} \text{ Oportunidades fechadas perdida}] / [\text{N}^{\circ} \text{ Oportunidades}]$	Percentual

Fonte: Elaborador pelo autor

Além da tabela de indicadores no canto esquerdo, existem outras duas visualizações principais: O primeiro quadro representa o valor em volume orçado dentro de cada segmento de atuação da empresa. O quadro abaixo representa dentro do segmento, o número de oportunidades abertas em comparação com o número de vendas realizadas. O quadro inferior representa o valor total orçado por tipo de serviço prestado pela empresa.

Na parte 2 do painel, conforme mostra a Figura 4.4, temos um complemento de visualizações, que são: No canto superior esquerdo, uma visualização do número de oportunidades totais e oportunidades vendidas em relação os serviços prestados pela empresa.

Figura 4.4 - Painel Parte 2 das análises da Empresa



Fonte: Desenvolvido pelo autor

No canto superior direito está a tabela de serviços prestados com os valores orçados. Por fim, na parte inferior, a tabela que busca relacionar os dados para análise de recorrência. Além disso, a última tabela possui uma rasura para que seja preservado os dados sensíveis da empresa.

Entende-se que com este painel já é possível responder as perguntas de negócios com base no desafio levantados no Canvas.

4.5 Análises do BI

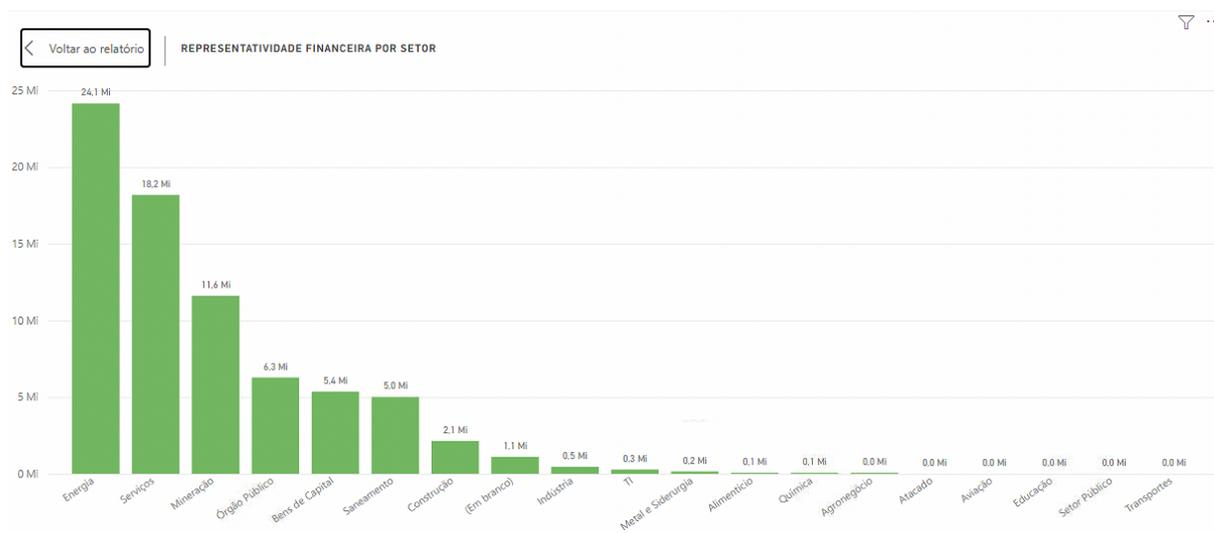
Buscando responder as perguntas de negócio mapeadas na aplicação do Canvas de Descoberta de Dados, seguem as análises abaixo.

**Qual o produto/segmento no qual a empresa tem um diferencial/se torna referência?
Quais produtos/segmentos vendemos mais?**

Para o entendimento desta hipótese, foi realizado um estudo dos melhores segmentos e posteriormente relacionado com os melhores produtos da empresa.

Conforme mostra a Figura 4.5, é possível perceber que existem segmentos mais relevantes dentro da empresa. Do total de R\$ 74,9 Mi orçado no período de análise, os setores mais representativos foram: Energia com R\$ 24,1 Mi, representando 30,1% dos projetos orçados. O setor de Serviços com R\$ 18,2 Mi, representando 22,7% e Mineração com R\$ 11,6 Mi, representando 14,5%. O total destes três segmentos de atuação da empresa representam 67,3% de projetos orçados. Portanto, os três melhores segmentos da empresa são **Energia, Mineração e Serviços**.

Figura 4.5 - **Representatividade financeira** do Segmento



Fonte: Elaborado pelo o autor

Para uma análise dos principais produtos vendidos no período, como mostra a Figura 4.6, foi constatado a partir dos dados que das 51 vendas realizadas, 13 vendas, ou seja, 25,4% do total vendido foi de **Consultoria Especializada** (considerada um tipo de produto); 8 vendas foram para o tipo de produto de **Estudo Hidrológico**, representando 15,6% das vendas; e 6 vendas foram para a categoria de produto **Outros**, representando 11,7%. Estes três produtos representam juntos 52,7 % dos serviços prestados.

Figura 4.6 - Principais Produtos Vendidos no período

Desc. Solução	Vlr. Serviços	Nº Oportunidades	Nº Oportunidades fechadas	Taxa Conversão
Consultoria especializada	6.234.733,4	26	13	50,00%
Estudo hidrológico	10.434.558,4	46	8	17,39%
Outros	24.215.225,3	33	6	18,18%
Balanço hídrico	621.884,5	8	3	37,50%
Cota de inundação	263.768,1	16	3	18,75%
Qualidade de água (superficial)	916.742,5	5	3	60,00%

Fonte:

Elaborado pelo autor

Correlacionando as duas análises realizadas anteriormente (Segmentos e tipo de produto), pretende-se descobrir quais tipos de produtos se enquadram em cada segmento.

No caso do segmento de **Energia**, conforme mostra a Figura 4.7, os produtos principais são Consultoria Especializada e a categoria **outros**. Neste caso, a categoria **outros** poderia estar melhor discriminada nos dados da empresa, de forma a auxiliar o time comercial a tomar

ações mais assertivas sobre esse produto, dado que esta categoria de produto tem uma grande representação no faturamento da empresa, sendo no segmento de Energia R\$2.88 Mi, representando 91% do total faturado no período, conforme mostra a Figura 4.8.

Figura 4.7 - Análise produtos Segmento **Energia**



Fonte: Elaborado pelo autor

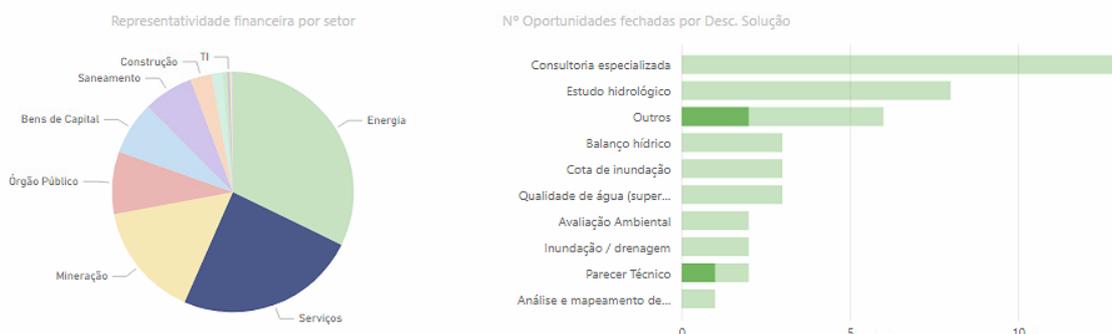
Figura 4.8 - Relevância da Categoria de Produtos Outros no Setor **Energia**

Desc. Solução	Vlr. Serviços	Nº Oportunidades fechadas
Consultoria especializada	193.113,0	4
Outros	2.885.000,0	2
Consultoria	63.800,0	1
Modelagem hidro-climático	28.811,7	1
Total	3.170.724,7	8

Fonte: Elaborado pelo autor

Já para o setor de **Serviços**, o principal produto como mostra a Figura 4.9, é a categoria **outros**, novamente, sendo aproximadamente 99% do faturamento no setor de serviços, representando R\$ 7.97 Mi de faturamento no período como mostra a Figura 4.10. Dessa forma, fica evidente a necessidade de melhor detalhamento dessa categoria de produto **outros**, dado a representatividade do faturamento. Como iniciativa para o time comercial, criar subcategorias dentro deste grupo, ou seja, dividir em grupos menores e mapear todas as oportunidades ainda não vendidas e tomar ações e buscando aproximação dos potenciais negócios.

Figura 4.9 - Análise produtos Segmento **Serviços**



Fonte: Elaborado pelo autor

Figura 4.10 - Categoria de Produto Outros no setor de **Serviços**

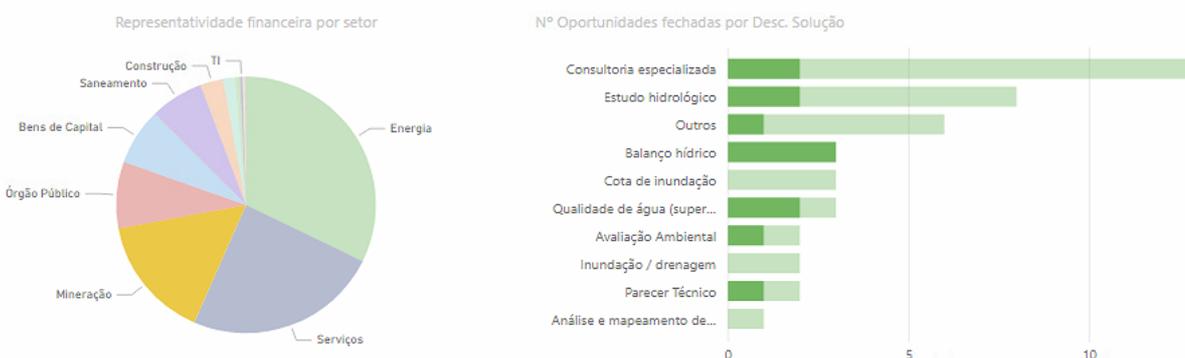
Desc. Solução	Vlr. Serviços	N° Oportunidades fechadas
Outros	7.975.462,0	2
Cursos	33.733,1	1
Parecer Técnico	36.000,0	1
Total	8.045.195,1	4

Fonte: Elaborado pelo autor

Por fim, o setor de **Mineração**, como mostra a Figura 4.11, é de forma isolada o melhor segmento de atuação da empresa, com 27 oportunidades abertas no período, tendo 15 convertidas em vendas, representando uma taxa de conversão no setor de 55,56%. Já, para os produtos do setor, temos três principais: a categoria **outros** com 45,40%, **Dam Break** com 26,45% e **Estudo Hidrológico** com 12,01%, representando 83,9% do total faturado no setor, conforme mostra

Figura 4.12.

Figura 4.11 - Análise produtos Segmento **Mineração**



Fonte: Elaborado pelo autor

Figura 4.12 - Produtos vendidos no setor de **Mineração**

Desc. Solução	Vir. Serviços	Nº Oportunidades fechadas
Outros	3.438.250,0	1
Dam Break	2.000.000,0	1
Estudo hidrológico	908.920,4	2
Drenagem pluvial	415.000,0	1
Qualidade de água (superficial)	311.141,1	2
Consultoria especializada	172.888,4	2
Balanço hídrico	163.038,2	3
Previsões meteorológicas	73.453,6	1
Parecer Técnico	50.000,0	1
Avaliação Ambiental	28.000,0	1
Total	7.560.691,7	15

Fonte: Elaborado pelo autor

Portanto, pode-se afirmar, baseado em dados, que o segmento de melhor atuação da empresa é o **setor de Mineração**, que a categoria de **serviços outros** é a mais **rentável** para a empresa.

Para a outra pergunta de negócio levantada na aplicação do Canvas, temos as seguintes análises.

Quais os tipos de projeto de maior recorrência?

Para responder essa análise, foi necessário montar a parte 2 do painel do BI. A análise de recorrência de projetos foi entendida através do número de vezes que o mesmo CNPJ e o mesmo Ponto focal (pessoa dentro da empresa que conduz projeto ou serviço) compram um produto da empresa. Dado os três produtos mais relevantes dentro da empresa (Consultoria Especializada Balanço Hídrico e Outros), a análise de recorrência é apresentada abaixo.

No produto de **Consultoria Especializada** é possível perceber, conforme mostra Figura 4.13, que durante todo o período analisado, somente uma vez ocorreu uma recompra da mesma pessoa ponto focal e do mesmo CNPJ.

Um ponto interessante nesse tipo análise é que o produto Consultoria Especializada possui alta taxa de conversão e baixa taxa de recorrência, portanto, o time comercial pensar em ações de relacionamento com os outros demais clientes que já compraram esse tipo de serviço pode

ser uma forma de ser mais assertivo nas propostas para clientes e ampliar os projetos vendidos para a empresa.

Figura 4.13 - Análise de Recorrência de **Consultoria Especializada**

Desc. Conta	Desc. Solução	Desc. Contato	Desc. Vertical	Nº Oportunidades	Nº Oportunidades fechadas	Taxa Conversão
	Consultoria especializada	Eduardo Henrique Alves		1	1	100,00%
	Consultoria especializada	Marcus		1	1	100,00%
	Consultoria especializada	William		2	2	100,00%
	Consultoria especializada	Carlos	Construção	1	1	100,00%
	Consultoria especializada	Gabriel	Construção	1	1	100,00%
	Consultoria especializada	Carine	Energia	1	1	100,00%
	Consultoria especializada	Karen	Energia	1	1	100,00%
	Consultoria especializada	Poliana	Energia	1	1	100,00%
	Consultoria especializada	Renata	Energia	1	1	100,00%
	Consultoria especializada	Ana Flavia	Mineração	1	1	100,00%
Total				26	13	50,00%

Fonte: Elaborado pelo autor

No produto de **Estudos Hidrológicos** é possível perceber, conforme Figura 4.14, que a recorrência ocorreu também em uma ocasião somente no período. No entanto, nesta ocasião da recorrência, ou seja, a recompra de mesmo CNPJ e mesmo ponto focal, foram cotados três projetos e somente dois foram vendidos. Um encaminhamento dessa análise poderia ser do time comercial revisar qual motivo da terceira oportunidade ter sido perdida e estreitar os relacionamentos com todos os demais pontos focais dos outros CNPJ, pois, baseado em dados, é possível constatar que existem oportunidades de recompra para os projetos de **Estudos Hidrológicos**.

Figura 4.14 - Análise de Recorrência de **Estudo Hidrológico**

Desc. Conta	Desc. Solução	Desc. Contato	Desc. Vertical	Nº Oportunidades	Nº Oportunidades fechadas	Taxa Conversão
	Estudo hidrológico	Willyan	Mineração	3	2	66,67%
	Estudo hidrológico	Fabio		1	1	100,00%
	Estudo hidrológico	Karine		1	1	100,00%
	Estudo hidrológico	Guilherme	Construção	1	1	100,00%
	Estudo hidrológico	Ian	Química	1	1	100,00%
	Estudo hidrológico	Rafael	Saneamento	1	1	100,00%
	Estudo hidrológico	Rafael	Saneamento	1	1	100,00%
	Estudo hidrológico	Giancarlo		1		
	Estudo hidrológico	Henrique	Agronegócio	1		
	Estudo hidrológico	Alejandro	Bens de Capital	1		
				46	8	17,39%

Fonte: Elaborado pelo Autor

Por fim, na categoria de produtos **outros** é possível perceber, conforme Figura 4.15, que a recorrência não se concretizou no período analisado. Um encaminhamento dessa análise seria de o time comercial retomar o contato com o Reinaldo afim de compreender os motivos pelos quais o projeto não foi fechado e se colocara disposição para outros projetos possíveis. Outro encaminhamento é investigar os demais pontos focais, inclusive, começando pelo segmento de Energia para entender as possibilidades de novos projetos na área.

Figura 4.15 - Análise de **Recorrência de Outros**

Desc. Conta	Desc. Solução	Desc. Contato	Desc. Vertical	Nº Oportunidades	Nº Oportunidades fechadas	Taxa Conversão
	Outros	Guilherme		1	1	100,00%
	Outros	Daniel	Energia	1	1	100,00%
	Outros	Reinaldo Bomfim da	Energia	2	1	50,00%
	Outros	Maria de Fátima	Mineração	1	1	100,00%
	Outros	Daila Aparecida Ferreira	Serviços	1	1	100,00%
	Outros	Paulo Henrique	Serviços	1	1	100,00%
	Outros	Andréa		1		
	Outros	Bontje		1		
	Outros	EPE		1		
	Outros	Tiaoo		1		
Total				33	6	18,18%

Fonte: Elaborado pelo autor

De modo geral, as informações no contexto de recorrência é algo bem frágil dentro da empresa de consultoria ambiental. Esse acompanhamento da recorrência é uma grande oportunidade de ser cada vez mais assertivo nas ofertas de propostas para outros clientes. Existem muitas outras possíveis análises para se investigar o tema recorrência. No entanto, é primordial que seja revisitado o Canvas pelo time comercial da empresa e faça novas reflexões sobre estes dados e suas aplicações.

Como sugestão de aprofundamento das análises futuras: mapear a partir dos principais segmentos de atuação (Energia, Mineração e Serviços) quais foram as oportunidades que foram abertas e que não foram vendidas há mais de 3 meses e propor uma ação com eles, de forma a realizar o giro do CRM e se manter próximo dos potenciais clientes.

4.6 Feedback da empresa

Buscando avaliar a efetividade das análises realizadas neste estudo, foi proposta uma reunião de apresentação das análises do BI realizadas neste trabalho. Para compartilhar os

feedbacks no presente estudo, foi solicitado a gravação da reunião. Os feedbacks foram bem positivos, como mostra trecho abaixo.

“As análises nos levaram a ter mais clareza daquilo que a gente faz e mais certeza para quem a gente tem relevância. O estudo nos deu mais clareza do nosso negócio. Bem positivo, claro, com algumas oportunidades de melhoria, mas para uma primeira aplicação está ótimo.”

Além disso, também tiveram pontos de melhoria, conforme abaixo:

- A conclusão da análise do melhor segmento poderia estar mais bem descrita, no sentido de quais foram as informações que fizeram chegar nesta conclusão, mais concatenada de sumarizar e se tornar didáticas, pois as informações já haviam sido apresentadas anteriormente.
- Na etapa do Canvas de perguntas de negócio, seria importante deixar mais especificado sob qual prisma seria a análise de melhor segmento.
- Sobre as análises de recorrência uma oportunidade futura para o time comercial seria aprofundar no contexto de CNPJ.

4.7 Principais achados da pesquisa

Esta Seção se propõe a trazer de forma sintetizada os principais achados da pesquisa realizada.

- É fundamental que o processo de descoberta de dados seja um processo cíclico, ser constantemente revisitado e acompanhado ao longo do desenvolvimento do projeto é de suma importância. O tempo recomendado seria pelo menos a cada quinze dias ou até mesmo uma vez por semana;
- Não são todos os times que realizam o processo de descoberta de dados, ou seja, não se soluciona todos os problemas de negócio com dados, muito menos, realizando um processo de descoberta. Em geral, para esse processo de descoberta a maturidade analítica do time é mais robusta;

- Importante que a aplicação do processo de descoberta de dados seja feita em conjunto com o time técnico e o time de negócio e que seja acompanhado pelos dois times ao longo do desenvolvimento do projeto;
- É fundamental que o time de dados tenha um bom conhecimento do negócio para provocar e direcionar a conversa de forma assertiva com o negócio;
- O processo de descoberta de dados não serve somente para resolver problemas, mas também para identificar oportunidades de melhoria e de novas possibilidades de receitas para empresa;
- As metodologias da literatura não são suficientes para as empresas como são apresentadas, todas realizam algum tipo de adaptação. A metodologia mais referenciada pelas empresas foi o CRISP-DM;
- Grande dificuldade dentro das empresas é no que tange a cultura de dados, as pessoas, de modo geral, que não são time técnico, não sabem sobre dados, não sabem o que podem solicitar, muito menos o que pode ser feito. Isso causa um grande *gap* entre time técnico e time de negócio;
- Ao definir que será utilizado um Canvas de dados é importante que o *output* desse trabalho de descoberta seja um produto de dados bem definido resolvendo problema de negócio claro. Em geral, nesses momentos o time de negócio quer resolver todos os problemas da empresa.
- A delimitação do tempo de projeto de dados para ser mensurado resultados em geral é de três meses. A definição dos problemas de negócio ou oportunidades que se querem desenvolver são prototipadas dentro deste tempo definido.
- A parte mais demorada das análises é saber quais são as informações importantes para a análise e como se pode combinar os dados que se possui para ter uma visualização.
- Um desafio de pessoas nesse contexto, hoje a competição por profissionais é global, pois com o trabalho remoto a competição começa a ser com empresas estrangeiras.
- O recomendado de próximo passo para o time comercial é de receber estas estas análises e realizar um acompanhamento por 3 meses para fazer uma primeira medida, do que está mudando da atual visualização.

- As análises levantadas no primeiro protótipo não devem ser consideradas esgotamento das possibilidades com os dados. No entanto, propor ações e mensurar de forma consistente irá fazer a grande diferença.
- É importante considerar o nível de entendimento das pessoas com dados quando se desenvolve um produto de dados.
- Um grande propulsor do projeto de dados está associado ao processo de mensuração dos indicadores e entendimento se está dando certo ou não.
- No que tange a desafios de negócio a serem resolvidos com dados, é preciso existir, pelo menos, um indicador que seja na perspectiva financeira. Este indicador está gerando receita ou diminuindo custo?
 - O Canvas de Descoberta de dados não esgota as possibilidades das informações fundamentais para descoberta de soluções em dados. Ainda seria possível explorar outros aspectos que o Canvas não considera. Exemplo: Quais tomadas de decisão queremos tomar com esses dados?

5 CONCLUSÃO

O presente trabalho teve por objetivo identificar os métodos de dados mais utilizados pelas empresas, como estes se relacionavam com a literatura e por fim identificar as informações fundamentais para se transformar um problema de negócio em uma solução de dados. Para que isso fosse possível, o presente estudo foi compreender as principais metodologias da literatura, identificando os processos e os passo-a-passo que se recomendam para que se gere valor a partir dos dados. Além disso, para entender a realidade prática das empresas com dados, foi feita uma pesquisa exploratória com roteiro semiestruturado com cinco organizações com as mais diferentes realidades e setores de atuação. Os dados sensíveis dos entrevistados foram preservados como compromisso firmado do autor com os presentes entrevistados. Na pesquisa exploratória com as empresas, o objetivo foi dividido em três blocos. O primeiro, foi entender o seu contexto e sua realidade com dados. O segundo foi mapear o seu processo e metodologia de dados, aprofundando, principalmente, no seu processo fim-a-fim, de fato, de como é transformado os problemas de negócio em soluções de dados, desde a geração do problema de negócio até a entrega final de dados para empresa ou time. Por fim, o terceiro foi entender sobre a cultura de dados, maturidade analítica e

principais gargalos e desafios dos projetos de dados nas suas realidades. Para tangibilizar na prática todas as informações levantadas, foi recomendado um Canvas de Descoberta de Dados, inspirado na literatura mapeada no presente estudo, nas entrevistas com as empresas e também em um *Canvas Analytics* citado pela colega Karina Moura (2018) orientada pela pesquisa da professora Dra. Daniela F. Brauner. Com isso, foi executado um protótipo do Canvas de Descoberta de Dados, recomendado neste estudo, com uma empresa do ramo de Consultoria Ambiental com seu time comercial. Um processo facilitado e conduzido de forma que fosse possível passar pelo processo fim-a-fim. Como entrega analítica deste protótipo, foram realizadas análises e buscou-se responder as hipóteses levantadas no processo. As análises foram possíveis de serem feitas a partir da base CRM disponibilizada pela empresa e foi utilizado a plataforma *PowerBI* da Microsoft. Portanto, a partir do presente estudo e processo de pesquisa realizado, cumprem-se os objetivos específicos levantados neste trabalho.

Como gargalos e dificuldades com este trabalho, foi considerado a parte do levantamento da literatura um desafio, pois há muita informação disponível e todas as metodologias da literatura possuem uma vasta explicação e detalhamento, além disso, as mais variadas metodologias que existem atualmente se tornavam um desafio de qual ser escolhida, como trazer de forma didática para o presente estudo, como relacionar estas informações com os demais passos da pesquisa, enfim, alguns pontos de desafios levantados. Ademais, um outro desafio foi a parte da pesquisa exploratória com as empresas, pois suas realidades sempre muito intensas, era um desafio de conseguir tempo com estes profissionais gabaritados. Todavia, as cinco empresas que foram entrevistadas foram de altíssima qualidade e cumpriram com o objetivo da pesquisa.

Por fim, a maior dificuldade e desafio foi relacionado a aplicação prática do Canvas de Descoberta de Dados, pois recomendar um processo, campos a serem preenchidos, a ordem que se é preenchida e por fim, a própria condução do preenchimento foi complexo. Além disso, ser possível gerar análises relevantes para a empresa em questão também foi um grande desafio. Um aspecto frágil deste trabalho foi a aplicação do Canvas desenvolvido neste estudo somente com a condução do próprio autor. No entanto, após ser possível visualizar os dados no *PowerBI*, após ter sido feito um mergulho na base fornecida pela empresa e poder fornecer análises que havia, de fato, uma necessidade real foi gratificante.

Para trabalhos futuros seria interessante também aprofundar:

1. Como seria a aplicação do Canvas de Descoberta de Dados em outros problemas de negócio e com bases de dados mais robustas;
2. Aplicação do Canvas pelo próprio time comercial e comparação dos resultados com o presente estudo.
3. Aplicação de algoritmos de aprendizado de máquina e inteligência artificial para bases de dados mais robustas;
4. Acompanhamento de, no mínimo, três meses desenvolvendo um projeto de dados com apoio contínuo do Canvas sendo possível mensurar desde o dia de início até o fim do período os avanços do trabalho;
5. Aplicação do Canvas para todas as empresas entrevistadas e entender os impactos, gargalos e oportunidades de melhoria no método;
6. Explorar mais o campo do Canvas de Cultura Analítica, aprofundando seu impacto nas pessoas e forma de aproximar ainda mais o tema de dados com a realidade da empresa.

Por fim, é compreendido que o presente estudo teve seus objetivos atingidos, aprofundando no tema de Ciência de Dados, entendendo como a literatura recomenda que seja trabalhado essa temática, investigando o processo de dados e a realidade das empresas no contexto de Ciência de Dados, e, por fim, aplicando de forma prática um protótipo empresa real e se extraíndo análises relevantes para o negócio em questão.

REFERÊNCIAS

A VISUAL GUIDE TO CRISP-DM METHODOLOGY. *Em: DESIGN FOR EXPERIENCES*. 13 mar. 2009. Disponível em: <https://exde.wordpress.com/2009/03/13/a-visual-guide-to-crisp-dm-methodology/>. Acesso em: 26 mar. 2022.

BOWNE-ANDERSON, H. What Data Scientists Really Do, According to 35 Data Scientists. **Harvard Business Review**, [s. l.], 2018. Disponível em: <https://hbr.org/2018/08/what-data-scientists-really-do-according-to-35-data-scientists>. Acesso em: 27 mar. 2022.

CHAPMAN, P. *et al.* CRISP-DM 1.0: Step-by-step data mining guide. [s. l.], 2000. Disponível em: <https://www.semanticscholar.org/paper/CRISP-DM-1.0%3A-Step-by-step-data-mining-guide-Chapman-Clinton/54bad20bbc7938991bf34f86dde0babfbd2d5a72>. Acesso em: 1 maio 2022.

DAVENPORT, T. H.; PATIL, D. J. Data Scientist: The Sexiest Job of the 21st Century. **Harvard Business Review**, [s. l.], 2012. Disponível em: <https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century>. Acesso em: 13 mar. 2022.

DHAR, V. **Data Science and Prediction**. [S. l.], 2013. Disponível em: <https://cacm.acm.org/magazines/2013/12/169933-data-science-and-prediction/fulltext>. Acesso em: 21 mar. 2022.

DIETRICH, D. *et al.* (org.). **Data science & big data analytics: discovering, analyzing, visualizing and presenting data**. Indianapolis, IN: Wiley, 2015.

EL ARASS, M.; SOUISSI, N. Data Lifecycle: From Big Data to Smart Data. *Em: DATA LIFECYCLE*, 2018. **Anais [...]**. [S. l.: s. n.], 2018. Disponível em: da.

GARTNER. **Data and Analytics: Everything You Need to Know**. [S. l.], 2019. Disponível em: <https://www.gartner.com/en/topics/data-and-analytics>. Acesso em: 2 abr. 2022.

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

MASON, H. **A Taxonomy of Data Science**. [S. l.], 2010. Disponível em: <https://web.archive.org/web/20211219192027/http://www.dataists.com/2010/09/a-taxonomy-of-data-science/>. Acesso em: 27 mar. 2022.

MICROSOFT. [S. l.], [s. d.]. Disponível em: <https://powerbi.microsoft.com/en-au/>. Acesso em: 1 maio 2022.

MOURA, K. V. de. Data Science : um estudo dos métodos no mercado e na academia. [s. l.], 2018. Disponível em: <https://lume.ufrgs.br/handle/10183/195011>. Acesso em: 1 maio 2022.

NIST. **Big Data Interoperability Framework: Volume 1, Definitions**. [S. l.]: National Institute of Standards and Technology, 2015. Disponível em: <https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.1500-1.pdf>. Acesso em: 26 abr. 2022.

PROVOST, F.; FAWCETT, T. Data Science and its Relationship to Big Data and Data-Driven Decision Making. **Big Data**, [s. l.], v. 1, n. 1, p. 51–59, 2013. Disponível em: <https://www.liebertpub.com/doi/10.1089/big.2013.1508>. Acesso em: 13 mar. 2022.

ROSSETTI, A.; MORALES, A. B. The role of information technology in knowledge management. **Ciência da Informação**, [s. l.], v. 36, p. 124–135, 2007. Disponível em: <http://www.scielo.br/j/ci/a/FzcdzsLpNJ43cXj5RcRWg5v/abstract/?lang=en>. Acesso em: 26 fev. 2022.

SALTZ, J. CRISP-DM is Still the Most Popular Framework for Executing Data Science Projects. *Em: DATA SCIENCE PROCESS ALLIANCE*. 30 nov. 2020. Disponível em: <https://www.datascience-pm.com/crisp-dm-still-most-popular/>. Acesso em: 24 mar. 2022.

SCHREIBER, A. T. *et al.* **Knowledge Engineering and Management: The CommonKADS Methodology**. [S. l.]: MIT Press, 2000.

TAYLOR, D. Battle of the Data Science Venn Diagrams. *Em*: KDNUGGETS. 2016. Disponível em: <https://www.kdnuggets.com/battle-of-the-data-science-venn-diagrams.html/>. Acesso em: 1 maio 2022.

ANEXO A – ROTEIRO SEMI-ESTRUTURADO DAS ENTREVISTAS

Roteiro das Entrevistas

Contexto Profissional, Empresa e Atuação com dados

Nome, Cargo e Formação

Nome da Empresa

Tempo de atuação na área?

Quanto tempo a empresa atua com a área de dados?

Processo, metodologia e passo-a-passo

Como é feito hoje o desenvolvimento de um projeto de dados do zero ? Qual é o processo, passo-a-passo para desenvolver um projeto ?

Existe alguma metodologia de base para o desenvolvimento do projeto de dados? Existe algum framework para embasar os projetos de dados?

Conhece algum dos métodos sugeridos pela literatura para o desenvolvimento de projetos de dados? (KDD, CRISP-DM, Data Analytics Lifecycle)

Quem são as pessoas envolvidas nesse trabalho?

Cultura, Maturidade e Análises

Qual a maturidade de análises que são feitas?

Quais análises são feitas?

Quais são os principais gargalos e desafios no desenvolvimento do projeto de dados? Em qual fase é mais desafiador?