UNIVERSIDADE FEDERAL DE SERGIPE
CENTRO DE CIÊNCIAS EXATAS E TECNOLÓGICAS
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

UM PROCESSO PARA O DESENVOLVIMENTO EXPERIMENTAL DE APLICAÇÕES DE DATA MINING E DATA SCIENCE ALINHADAS AO PLANEJAMENTO ESTRATÉGICO DA ORGANIZAÇÃO

Dissertação de Mestrado

Rodrigo Fontes Cruz

São Cristóvão - Sergipe
2021

UNIVERSIDADE FEDERAL DE SERGIPE
CENTRO DE CIÊNCIAS EXATAS E TECNOLÓGICAS
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

Rodrigo Fontes Cruz

UM PROCESSO PARA O DESENVOLVIMENTO EXPERIMENTAL DE APLICAÇÕES DE DATA MINING E DATA SCIENCE ALINHADAS AO PLANEJAMENTO ESTRATÉGICO DA ORGANIZAÇÃO

Dissertação de mestrado apresentada ao Programa de Pós-Graduação em Ciência da Computação (PROCC) da Universidade Federal de Sergipe (UFS) como parte de requisito para obtenção do título de Mestre em Ciência da Computação.
Orientador: Prof. Dr. Methanias Colaço Rodrigues Júnior.

São Cristóvão - Sergipe
2021

FICHA CATALOGRÁFICA ELABORADA PELA BIBLIOTECA CENTRAL
UNIVERSIDADE FEDERAL DE SERGIPE

Cruz, Rodrigo Fontes
C957p Um processo para o desenvolvimento experimental de aplicações de *data mining* e *data science* alinhadas ao planejamento estratégico da organização / Rodrigo Fontes Cruz ; orientador Methanias Colaço Rodrigues Júnior. - São Cristóvão, 2021.
112 f.; il.

Dissertação (mestrado em Ciência da Computação) – Universidade Federal de Sergipe, 2021.

1. Mineração de dados (Computação). 2. Estruturas de dados (Computação). 3. Recuperação de dados (Computação). I. Rodrigues Júnior, Methanias Colaço orient. II. Título.

CDU 004

UNIVERSIDADE FEDERAL DE SERGIPE
PRÓ-REITORIA DE PÓS-GRADUAÇÃO E PESQUISA
COORDENAÇÃO DE PÓS-GRADUAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

Ata da Sessão Solene de Defesa da Dissertação do Curso de Mestrado em Ciência da Computação-UFS.
Candidato: RODRIGO FONTES CRUZ

Em 22 dias do mês de junho do ano de dois mil e vinte um, com início às 08h00min, realizou-se na Sala virtual https://meet.google.com/cmk-tkwn-uzx. A Sessão Pública de Defesa de Dissertação de Mestrado do candidato **RODRIGO FONTES CRUZ**, que desenvolveu o trabalho intitulado: **"UM PROCESSO PARA O DESENVOLVIMENTO EXPERIMENTAL DE APLICAÇÕES DE DATA MINING E DATA SCIENCE ALINHADAS AO PLANEJAMENTO ESTRATÉGICO DA ORGANIZAÇÃO",** sob a orientação do Prof. Dr. **Methanias Colaço Rodrigues Júnior**. A Sessão foi presidida pelo Prof. Dr. **Methanias Colaço Rodrigues Júnior** (PROCC/UFS), que após a apresentação da dissertação passou a palavra aos outros membros da Banca Examinadora, Prof. Dr. **Jefferson David Araujo Sales** (UFS) e em seguida, a Prof. Dr. **Hendrik Teixeira Macedo** (PROCC/UFS). Após as discussões, a Banca Examinadora reuniu-se e considerou o mestrando (a)_____aprovado_____ "*(aprovado/reprovado)*". Atendidas as exigências da Instrução Normativa 01/2017/PROCC, do Regimento Interno do PROCC (Resolução 67/2014/CONEPE), Resolução nº 25/2014/CONEPE e da Portaria nº 413 de 27 de maio de 2020 (Banca por videoconferência) que regulamentam a Apresentação e Defesa de Dissertação, e nada mais havendo a tratar, a Banca Examinadora elaborou esta Ata que será assinada pelos seus membros e pelo mestrando.

Cidade Universitária "*Prof. José Aloísio de Campos*", 22 de junho de 2021.

gov.br Documento assinado digitalmente
Methanias Colaco Rodrigues Junior
Data: 22/06/2021 12:08:03-0300
CPF: 693.380.275-20
Verifique em https://verificador.iti.br

Prof. Dr. Methanias Colaço Rodrigues Júnior
(PROCC/UFS)
Presidente

gov.br Documento assinado digitalmente
Hendrik Teixeira Macedo
Data: 22/06/2021 14:59:35-0300
CPF: 928.416.275-00
Verifique em https://verificador.iti.br

Prof. Dr. Hendrik Teixeira Macedo
(PROCC/UFS)
Examinador Interno

gov.br Documento assinado digitalmente
Jefferson David Araujo Sales
Data: 23/06/2021 18:55:30-0300
CPF: 694.934.135-00
Verifique em https://verificador.iti.br

Prof. Dr. Jefferson David Araujo Sales
(UFS)
Examinador Externo

Rodrigo Fontes Cruz

Rodrigo Fontes Cruz
Candidato

AGRADECIMENTOS

Primeiramente quero agradecer à minha familia, principalmente meus pais, por todo apoio, paciencia e compreensão.

A minha esposa Anne, por sempre me incentivar e nunca me deixar desistir, essa conquista é também sua.

Ao meu professor e orientador Methanias, pela confiança, paciência e por todo o aprendizado durante a condução da orientação deste mestrado.

A todos os professores que tive durante toda a jornada da vida, por todo o conhecimento compartilhado.

A todos os membros do Programa 5A que com toda compreensão, apoiaram e me ajudaram.

Por fim, a todos que indiretamente ou diretamente, estiveram envolvidos nessa caminhada e que contribuíram de alguma forma. O meu mais sincero agradecimento.

RESUMO

Contexto: O fenômeno Big Data tem imposto maturidade às empresas na exploração de seus dados, como prerrogativa para obter *insights* valiosos sobre seus clientes e o poder da análise para orientar a tomada de decisão. Desta forma, uma abordagem geral que descreva como extrair conhecimento para a execução da estratégia empresarial precisa ser estabelecida. **Objetivo:** O objetivo deste trabalho é desenvolver e avaliar um processo de desenvolvimento de aplicações de *Data Mining* e *Data Science* dirigidas à estratégia e avaliadas experimentalmente. **Método:** Inicialmente, foi realizada uma Revisão *Quasi*-Sistemática da literatura, com a finalidade de identificar e caracterizar os métodos de desenvolvimento de aplicações de BI (*Business Intelligence*) e de *Data Mining* dirigidos à estratégia, ou que preveem avaliação Experimental. Por fim, foi realizado um estudo de caso em uma instituição de ensino federal, com o objetivo introduzir e avaliar o processo desenvolvido. **Resultados:** A Revisão da Literatura evidenciou a ausência de uma abordagem completa para disciplinar o alinhamento estratégico e a experimentação, prevendo atendimento claro aos objetivos estratégicos e uma fase experimental na validação dos resultados. O estudo de caso trouxe evidências iniciais positivas de que é possível disciplinar e alinhar o desenvolvimento de aplicações de *Data Mining* e *Data Science* ao planejamento estratégico da organização, bem como fomentar o uso do método científico neste contexto. **Conclusão:** Uma metodologia de BI dirigida à estratégia pode ser estendida para a contemplação e o desenvolvimento de aplicações de *Data Mining* e *Data Science* avaliadas experimentalmente.

Palavras-chave: *Data Mining*, Mineração de Dados, *Data Science*, Alinhamento Estratégico, *Data Analytics*, Experimentação.

ABSTRACT

Context: The Big Data phenomenon has imposed maturity on companies in the exploration of their data, as a prerogative to obtain valuable insights about their customers and the power of analysis to guide decision making. In this way, a general approach that describes how to extract knowledge for the execution of the business strategy needs to be established. **Goal:** The objective of this work is to develop and evaluate a process of development of Data Mining and Data Science applications directed to the strategy and evaluated experimentally. **Method:** Initially, a Quasi-Systematic Review of the literature was carried out, with the purpose of identifying and characterizing the methods of developing BI (Business Intelligence) and Data Mining applications directed to the strategy or that provide for Experimental evaluation. Finally, a case study was carried out at a federal educational institution, with the objective of introducing and evaluating the developed process. **Results:** The Literature Review evidenced the absence of a complete approach to discipline strategic alignment and experimentation, providing clear compliance with strategic objectives and an experimental phase in the validation of results. The case study brought positive initial evidence that it is possible to discipline and align the development of Data Mining and Data Science applications to the organization's strategic planning, as well as to encourage the use of the scientific method in this context. **Conclusion:** A BI methodology directed to the strategy can be extended to the contemplation and development of Data Mining and Data Science applications evaluated experimentally.

Keywords: Data Mining, Data Science, Strategic Alignment, Data Analytics, Experimentation.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

ASUM-DM	*Analytics Solutions Unified Method*
BI	*Business Intelligence*
CRISP-DM	*Cross Industry Standard Process for Data Mining*
DM	*Data Mining*
DA	*Data Analytics*
ETL	*Extract Transform Load*
GQM	Goal Question Metric
IBM	*International Business Machines Corporation*
OEC	*Overall Evaluation Criteria*
PEN	Plano Estratégico de Negócio
PETI	Plano Estratégico de Tecnologia de Informação
PICO	População, Intervenção, Comparação e *Outcomes* (Resultado)
ROI	Retorno sobre o investimento
RQSL	Revisão Quasi-Sistemática da Literatura
SAS	*Statistical Analysis System*
TI	Tecnologia da Informação
UFS	Universidade Federal de Sergipe

SUMÁRIO

1.0 INTRODUÇÃO

Este capítulo pretende realizar uma breve contextualização relacionada ao tema da pesquisa, motivação, problemática, questões, objetivos e suposição que se pretende evidenciar. Além disto, são descritas as contribuições que se espera alcançar ao final do trabalho e a metodologia de pesquisa direcionadora.

1.1 CONTEXTUALIZAÇÃO

Com a inevitável mutação dos mercados, no âmbito empresarial, saber como obter resultados positivos em meio a essas mudanças, desenvolvendo processos adequados e eficazes para gerir adequadamente as transformações, tornou-se uma obrigação. Decisões erradas, sejam estratégicas, táticas ou operacionais, podem custar o futuro da empresa, assim como uma correta, definir sua sobrevivência ou sua expansão (Côrte-Real, Oliveira & Ruivo, 2017).

Neste contexto, os dados surgem como uma importante fonte para obter vantagem competitiva (Kubina, Varmus & Kubinova, 2015). Vantagem que se baseia no conhecimento obtido com a análise de dados e tem impulsionado áreas tais como *Business Intelligence* (BI), *Data Mining* (DM) e *Data Science*, sendo esta última um jargão mais recente que busca englobar as duas primeiras e aspectos de validação científica às aplicações. Isto associado à atual era *Big Data*, na qual o progresso tecnológico impulsiona a criação de grandes volumes de dados em alta velocidade, a partir de uma variedade de fontes, tem justificado ainda mais o investimento em DM, cujo poder e automaticidade têm possibilitado lidar com grandes quantidades de dados e extrair valor (Shmueli et al., 2017).

No entanto, antes que qualquer tentativa possa ser feita para realizar a extração desse conhecimento útil, uma abordagem geral que descreva como extrair conhecimento precisa ser estabelecida (Kurgan & Musilek, 2006). Desta forma, vários modelos de processos de *Data Mining* foram propostos por pesquisadores e profissionais. Os exemplos incluem Fayyad, et al. (1996), Cabena et al. (1998), Cios et al. (2000), CRISP-DM (2003), Berry & Linoff (1997), Sharma, Osei-Bryson & Kasper (2012) e Ławrynowicz & Potoniec (2014).

Além disso, como os objetivos estratégicos da organização precisam ser traduzidos para os níveis mais baixos do negócio, esta metodologia deve ser planejada e executada a partir de dentro da organização. Em outras palavras, é preciso preencher a lacuna entre a estratégia de negócios e sua implementação ao nível do projeto de software (Basili, et al., 2010; Basili, et al., 2014).

Nesta mesma linha, Mandic et. al. (2010) enfatizam que para se analisar todos os aspectos relevantes para a tomada de decisão, é necessário implantar métodos para a integração dos dados existentes com as metas estratégicas.

Apesar deste senso comum entre diversos pesquisadores, um *survey* realizado no Brasil (Lima et al., 2017) evidenciou que 72% das empresas não utilizam um método específico para o desenvolvimento de aplicações de BI alinhado ao planejamento estratégico da organização. Outra evidência encontrada pelo *survey*, que também serviu de motivação para apresentação desta dissertação, destaca que 67,50% dos entrevistados não utilizam uma metodologia formal para o desenvolvimento de aplicações BI. Neste caso, vale ressaltar que apesar da literatura e a prática separarem conceitualmente as áreas de *Data Mining* e BI, há uma forte convergência e integração destas, pois o "I", ou Inteligência do BI, só pode ser concretizado com a aplicação de técnicas de *Data Mining*. Isto indica que a ausência destas metodologias de alinhamento também deve atingir os projetos de *Data Mining*, uma vez que existem projetos de BI sem *Data Mining*, *Data Mining* sem BI e, no melhor caso, o qual será considerado nesta dissertação, um BI completo, que utilizada integração de dados analíticos (BI), estatística e inteligência artificial - *Data Mining* -.

Diante destes dados, para complementar e confirmar os resultados do *survey*, foi feita uma revisão *quasi*-sistemática da literatura, apresentada no Capítulo 3, a qual confirmou a escassez de métodos que disciplinem a criação de aplicações de BI alinhadas à estratégia ou que sugerem a combinação de métodos já existentes para o alcance do mesmo propósito de alinhamento. Além disto, foi possível observar que essa deficiência também atinge o desenvolvimento de aplicações de *Data Mining*.

Neste sentido, Colaço Jr et al. (2019) propuseram uma abordagem que mescla a metodologia GQM+*Strategies* com uma metodologia ágil de desenvolvimento de aplicações de *Business Intelligence* proposta pelo autor, visando garantir o alinhamento estratégico e agilidade na entrega das soluções. Esta abordagem foi avaliada por meio de um estudo de caso em uma empresa multinacional latino-americana, o qual apresentou resultados iniciais de conscientização da equipe de desenvolvimento BI sobre as necessidades da transparência dos objetivos estratégicos e da criação de aplicações para o alcance destes.

Consequentemente, o objetivo deste trabalho foi desenvolver e avaliar um processo que estenda a metodologia de BI dirigida à estratégia proposta por Colaço Jr et al. (2019), para contemplar aplicações de *Data Mining* e *Data Science* avaliadas experimentalmente.

1.2 PROBLEMÁTICA E SUPOSIÇÃO

As empresas estão percebendo, cada vez mais, o valor potencial dos dados para obter insights sobre seus clientes e o poder da análise para orientar a tomada de decisões, no entanto, o contexto *Big Data* tem imposto desafios ainda maiores à medida que as empresas lutam para desenvolver capacidades analíticas para os dados existentes (Phillips-Wren & Hoskisson, 2015).

A análise de dados moderna é muito diferente de outros métodos que existiam anteriormente. Além disso, os dados também são muito diferentes. Em outras palavras, a natureza dos dados modernos (maior dimensão, diversos tipos, massa de dados) não autoriza o uso da maioria dos métodos estatísticos convencionais (Sedkaoui, 2018).

Para gerenciar esses conjuntos de dados novos e potencialmente inestimáveis, novos métodos e novas aplicações na forma de análise preditiva estão sendo desenvolvidos. Diante deste contexto, o termo *Data Analytics* (Análise de Dados) tem ganhado força nos últimos anos, o que implica necessidade de profissionais de inteligência com conhecimentos de matemática, mineração de dados, amostragem e estatística inferencial, pois será inviável analisar todos os dados disponíveis (Lalanne, 2016).

Em contrapartida, da mesma forma que o contexto *Big Data* tem imposto desafios maiores à Análise de Dados, este tem aumentado as oportunidades de extração de conhecimento por meio de projetos de Mineração de Dados. A diversidade dos dados aumentou - em origem, formato e modalidades -, bem como aumentou a variedade de técnicas provenientes de aprendizado de máquina, gerenciamento de dados, visualização, inferência causal e outras áreas. Além disso, tendo como base o que ocorria há vinte anos, existem muitas outras maneiras pelas quais os dados podem ser monetizados, considerando os novos tipos de aplicativos, interfaces e modelos de negócios.

Este crescimento exponencial da área de derivação de valor, em tamanho e complexidade, também a tornou muito mais exploratória, sob a égide da Ciência de Dados. Para esta nova área, os estágios orientados a dados e orientados ao conhecimento interagem, em contraste com o processo tradicional de Mineração de Dados, a partir de objetivos de negócios precisos que se traduzem em uma tarefa clara de garimpagem, a qual, após a validação estatística dos dados, finalmente converte "dados em conhecimento". (Martínez-Plumed et al., 2019).

Nessa tentativa, conforme dito anteriormente, vários modelos de processo de descoberta de conhecimento por Mineração de Dados foram propostos por pesquisadores e profissionais. No entanto, a maioria desses métodos foram propostos em um contexto diferente

do que vivemos hoje. A análise de dados moderna é muito diferente da de outros métodos que já existiam (Sedkaoui, 2018). O CRISP-DM, por exemplo, que é considerada a metodologia mais amplamente utilizada, de acordo com muitas pesquisas de opinião, teve sua origem na segunda metade dos anos 90 e, portanto, tem cerca de duas décadas (Sharma, Osei-Bryson & Kasper, 2012; Schäfer et al., 2018; Martínez-Plumed et al., 2019). Desta forma, por mais que essas mudanças não tenham ocorrido rapidamente e novas metodologias tenham sido propostas para acomodar algumas das mudanças, a essência das metodologias não abrange totalmente a diversidade de projetos de Ciência de Dados (Martínez-Plumed et al., 2019) e não disciplinam explicitamente o alinhamento estratégico. Como exemplos dessas metodologias, a IBM introduziu o ASUM-DM (IBM, 2005) e o SAS introduziu o SEMMA (SAS, 2005).

Uma alternativa para atender os pressupostos da Ciência de Dados é padronizar os projetos de inteligência para o uso de uma abordagem experimental (Wohlin et al., 2012; Juristo & Moreno, 2013; Santos et al., 2018), uma vez que a aplicação de um método científico rigoroso coaduna com a tentativa de tornar a análise de dados uma ciência, com princípios que diminuem as ameaças à validade do conhecimento gerado. Na literatura, alguns projetos, tais como os encontrados em Kohavi et al. (2013) e em Costa et al. (2015), já utilizaram experimentação como forma de selecionar as soluções com maior retorno de valor para o negócio.

Além da exigência de uma Ciência de Dados que averigue a verdade, o valor e a validade das descobertas, em se tratando de Brasil, o *survey* supracitado (Lima et al., 2017) e a revisão *quasi*-sistemática feita nesta pesquisa evidenciaram que as metodologias de desenvolvimento de BI e Data Mining não utilizam um método específico alinhado ao planejamento estratégico da organização. O que faz com que a implementação de um sistema de BI e Data Mining, muita das vezes, não consiga influenciar a tomada de decisão e entregar valor ao negócio.

Segundo Olszak (2012), muitos projetos de BI falham ou não são entregues em sua totalidade. Eliminando as questões políticas e micropolíticas de interesses, muito comum em instituições públicas brasileiras, a principal razão para o fracasso é o conhecimento limitado das organizações sobre as oportunidades geradas e os seus benefícios, além do fato da complexidade e a versatilidade dos sistemas de *Business Intelligence* modernos exigirem uma metodologia sólida em sua implantação. Como exemplo, se considerarmos o ponto de vista tecnológico, são observados como fatores críticos de sucesso: a qualidade dos dados e a maturidade das respostas dos usuários envolvidos, quanto ao processo de negócio da organização.

Diante do exposto e das lacunas identificadas, o problema que será trabalhado dentro desta dissertação pode ser delimitado. Segundo Vergara (2006, p. 21), um problema refere-se “a alguma lacuna epistemológica ou metodológica percebida, a alguma dúvida quanto à sustentação de uma afirmação geralmente aceita, a alguma necessidade de por à prova uma suposição”.

Para Gil (2002, p. 49), o problema, para ser cientificamente válido, deve passar pelo crivo das seguintes questões: pode o problema ser enunciado em forma de pergunta? Corresponde a interesses pessoais (capacidade), sociais e científicos, isto é, de conteúdo e metodológicos? Esses interesses são harmonizados? Constitui-se o problema em questão científica, ou seja, relacionam-se entre si pelo menos duas variáveis? Pode ser objeto de investigação sistemática, controlada e crítica? E, por fim, pode ser empiricamente verificado em suas consequências?

Assim, propõe-se a seguinte indagação central desta pesquisa:

• Um processo de BI dirigido à estratégia pode ser estendido para o desenvolvimento de aplicações de Data Mining e Data Science avaliadas experimentalmente?

A partir desta indagação, duas outras questões subjacentes são colocadas para discussão. Enumeramo-las:

a) Aplicações de BI e DM enfrentam os mesmos problemas de alinhamento estratégico de outras aplicações?

b) Um processo que encapsula experimentação disciplinará o exercício da Ciência de Dados?

Identificado o problema, faz-se necessária a elaboração de uma suposição passível de investigação dentro da proposta desta dissertação. A suposição em questão é: uma metodologia de BI dirigida à estratégia pode ser estendida para a contemplação e o desenvolvimento de aplicações de Data Mining e Data Science avaliadas experimentalmente.

1.3 JUSTIFICATIVA

Uma pesquisa realizada em 2016 pela Deloitte, com 1.200 executivos de TI (CIO), afirma que 78% das empresas entrevistadas veem o alinhamento estratégico de atividades de TI com a estratégia de negócios como fundamental para o sucesso de uma organização (Deloitte, 2017). A transformação digital da maioria dos setores de uma organização requer a integração eficaz e eficiente dos recursos de software, em todos os tipos de produtos e serviços, como um pré-requisito para a construção de modelos de negócios bem-sucedidos, a fim de salvaguardar um futuro lugar no mercado. Como consequência, os gestores das organizações

precisam entender como usar o software e como medir (mensurar) as atividades deste software, aliadas aos objetivos de alto nível da organização, tais como metas de negócios (Münch, et al., 2013).

As medições realizadas no desenvolvimento de um software são definidas nas organizações, entretanto, geralmente não estão vinculadas à estratégia da organização, carregando com isso, dados inúteis ou sem nenhum benefício estratégico (Olszak, 2012). É necessário definir um processo de desenvolvimento de aplicações que possam apoiar a tomada de decisão e que, de fato, estejam vinculadas às demandas estratégicas.

Nesse sentido, visando à entrega de valor para as organizações, as empresas de desenvolvimento de software estão, cada vez mais, orientando-se por dados e tentando experimentar continuamente os produtos utilizados por seus clientes. Muitas tecnologias com alguns pressupostos experimentais, tal como a de teste A / B, já são familiares aos Engenheiros de Software, no entanto, estes raramente conseguem evoluir e adotar uma metodologia deste tipo, perdendo a oportunidade de agregar vantagem competitiva aos projetos (Fagerholm et al., 2017).

Baseando-se neste contexto e na democratização das aplicações de *Data Mining e Data Science*, propõe-se o direcionamento explícito do processo de desenvolvimento dessas aplicações ao planejamento estratégico, agregando também o uso de experimentação. O resultado desse trabalho contribuiu com a criação de um novo processo de desenvolvimento de aplicações de *Data Mining* e *Data Science* voltadas verdadeiramente à estratégia e embasadas experimentalmente.

1.4 OBJETIVO GERAL

Este trabalho teve como objetivo geral desenvolver e avaliar um processo de desenvolvimento de aplicações de *Data Mining* e *Data Science* dirigidas à estratégia e avaliadas experimentalmente, como parte integrante de um projeto de BI.

1.5 OBJETIVOS ESPECÍFICOS

Para possibilitar a realização do objetivo geral, podemos enumerar os seguintes objetivos específicos:

- Revisão *quasi*-sistemática com a finalidade de identificar e caracterizar os métodos de desenvolvimento de aplicações BI e de *Data Mining* dirigidos à estratégia e que preveem avaliação Experimental;

- Processo de alinhamento estratégico para a construção Experimental de aplicações de *Data Mining* e *Data Science* inseridas no contexto BI de uma organização;
- Estudo de caso para avaliar o processo desenvolvido.

1.6 METODOLOGIA

O estudo foi desenvolvido sob uma perspectiva metodológica de natureza aplicada, tendo em vista que o interesse do estudo é a aplicação do conhecimento gerado. A pesquisa aplicada tem como características o interesse na aplicação, utilização e as ações práticas pelo conhecimento (Gil, 2008).

Seu objetivo apresenta caráter exploratório e descritivo. Uma pesquisa exploratória tem como finalidade a familiarização do problema por meio da análise de dados ou de observações empíricas (Marconi & Lakatos, 2003). Quanto ao objetivo descritivo, sua conduta procura a caracterização e a determinação de fenômenos ou populações (Gil, 2008).

Considerando o ponto de vista exploratório, inicialmente, foi realizada uma revisão *quasi*-sistemática da literatura, com a finalidade de identificar e caracterizar as metodologias de desenvolvimento de aplicações de *Business Intelligence* e *Data Mining* dirigidas à estratégia, ou preveem avaliação Experimental. Ato contínuo, foi proposto um processo de alinhamento estratégico para a construção Experimental de aplicações de *Data Mining* e *Data Science*.

Do ponto de vista da pesquisa aplicada, e como forma de avaliar o processo proposto, foi realizado um estudo de caso. Um estudo de caso é uma investigação empírica que investiga um fenômeno contemporâneo em profundidade e em seu contexto de vida real, especialmente quando os limites entre o fenômeno e o contexto não são claramente evidentes (Yin, 2015).

Por questões de escopo e dos riscos inerentes à disponibilidade de pelo menos uma organização para a avaliação, o estudo de caso foi essencialmente qualitativo, com questionários sobre o uso do processo proposto.

1.7 ORGANIZAÇÃO DA DISSERTAÇÃO

Este documento está organizado de acordo com a Instrução Normativa Nº 05/2019/PROCC, a qual permite que a Dissertação seja "uma compilação de artigos científicos submetidos ou publicados em veículos com *Qualis*". São 6 capítulos que fornecem uma base conceitual para o entendimento sistêmico. Os tópicos a seguir descrevem o conteúdo de cada um dos capítulos:

- O Capítulo 1 apresenta esta Introdução, explicando as justificativas, juntamente com os problemas levantados e a suposição de pesquisa;
- O Capítulo 2 apresenta o glossário teórico do trabalho;
- O Capítulo 3 replica parte de uma Revisão *Quasi*-Sistemática submetida ao periódico RIAE - Revista Ibero-Americana de Estratégia;
- O Capítulo 4 apresenta parte de um artigo submetido ao periódico JISTEM - *Journal of Information Systems and Technology Management*, no qual é proposto e avaliado um processo para o desenvolvimento experimental de aplicações de *Data Mining e Data Science* alinhadas ao planejamento estratégico da organização;
- O Capítulo 5 traz uma síntese narrativa da Revisão *Quasi*-Sistemática, juntamente com uma discussão dos resultados obtidos pela aplicação do processo proposto;
- Finalmente, no capítulo 6, é apresentada uma compilação de conclusões, contribuições e sugestões de trabalhos futuros.

2.0 GLOSSÁRIO TEÓRICO

Neste capítulo, é apresentado o glossário teórico necessário para fundamentar a realização deste trabalho. A seguir, será conceituado o Alinhamento Estratégico, bem como os conceitos de BI, *Big Data*, *Data Analytics*, *Data Mining* e Experimentação em Software.

2.1 ALINHAMENTO ESTRATÉGICO

Na literatura, existem muitos conceitos que detalham o significado de alinhamento estratégico, no entanto, buscamos aqueles que possuem maior vínculo com o assunto de nossa proposta. (1) O alinhamento entre o plano estratégico de negócio (PEN) e o plano estratégico de tecnologia de informação (PETI) é alcançado quando o conjunto de estratégias de sistemas (objetivos, obrigações e estratégias) é derivado do conjunto estratégico organizacional (missão, objetivos e estratégias) (King, 1988); (2) O elo entre PEN-PETI corresponde ao grau no qual a missão, os objetivos e os planos de TI refletem, suportam e são suportados pela missão, pelos objetivos e pelos planos de negócio (Reich, et al., 1996); (3) É a forma como os negócios e a TI trabalham em conjunto para alcançar o objetivo comum (Campbell, 2005) e (4) O alinhamento entre PEN-PETI é a adequação da orientação estratégica do negócio com a de TI (Chan, et al., 1997).

O problema das organizações, atualmente, é que nem sempre as empresas conseguem declarar os objetivos estratégicos de forma explícita ou suficientemente clara, para que se possa verificar se tais objetivos têm realmente alcançado as metas e estão alinhados à TI. Este desafio não é ter a área de TI como um suporte, mas sim como parte de uma plataforma de negócio, servindo como elemento essencial à estratégia de negócio. Essa nova visão deve vincular o alinhamento estratégico da TI ao negócio da organização. Estes dois elementos precisam relacionar-se entre si, em busca da melhoria contínua e do sucesso da organização.

Hoje, uma das estratégias de TI mais importantes é a adoção de *Business Intelligence* e *Data Mining*, que tem como principal característica a capacidade de lidar com grandes quantidades de dados e extrair valor. A pergunta que pode ser feita é: as aplicações de BI e DM enfrentam os mesmos problemas de alinhamento estratégico de outras aplicações? A seguir, detalharemos o conceito de BI e Data Mining e sua relação com o Planejamento Estratégico.

2.2 BUSINESS INTELLIGENCE

Business Intelligence – BI – pode ser entendido como um conjunto de metodologias, processos, arquiteturas e tecnologias usadas para apoiar a coleta, análise,

apresentação e disseminação de informações de negócios para permitir tomadas de decisões estratégicas, táticas e operacionais mais efetivas (Hans et al., 2013; Dedić & Stanier, 2017).

O BI ajuda as empresas a pensar melhor sobre a concorrência por meio de um melhor entendimento da base de clientes (Brannon, 2010), o que pode levar à criação de um relacionamento mais próximo e mais forte com os clientes e ao aumento da receita (Alexander, 2014). Além disso, desempenha um papel crítico para os negócios em termos de desenvolvimento organizacional, fornecendo vantagem competitiva, no contexto de alcançar assimetria positiva de informações (Thamir & Poulis, 2015), e contribui para otimizar processos e recursos de negócios, maximizar lucros e melhorar proativamente, bem como para tomada de decisão estratégica (Dedić & Stanier, 2016).

Dessa forma, o *Business Intelligence* nas organizações é entendido como uma vantagem estratégica (Chaudhuri & Dayal, et al., 2011; Kohtamäki & Farmer, 2017), independentemente da área em que a organização atue, seja ela privada ou não, pois, na atualidade, as organizações que utilizam sistemas deste tipo têm facilidade em adquirir conhecimento específico sobre os diversos fatores que a influenciam, podendo posteriormente aplicar tal conhecimento, identificando o potencial de mercado e, com isto, o direcionar na sua estratégia, visão e metas a atingir.

Além de suas vantagens estratégicas e táticas, o *Business Intelligence* também é usado no nível operacional, de forma a permitir que vários tipos de usuários identifiquem tendências emergentes, tomem decisões mais rápidas, tomem ações e lidem com os problemas organizacionais assim que surgirem. Seu objetivo é ajudar as partes interessadas a entender melhor as operações de sua organização, tomar decisões de negócios mais sábias e informadas e gerenciar o desempenho operacional (AICPA, 2015).

Assim, *Business Intelligence* refere-se ao ato de proporcionar aos negócios o apoio necessário para a tomada de decisão, através do uso de um conjunto de técnicas e ferramentas (Gartner, 2015). Para Bologa & Bologa (2011) e Bonel (2015), é possível identificar três principais grupos de atividades para alcançar inteligência nos negócios:

- Acessar, integrar e armazenar dados de diferentes fontes;
- Analisar e transformar dados em informação;
- Apresentar a informação.

O BI convencional se concentrou em atividades como ETL, *Data Warehousing* e *Reporting*, cobrindo assim áreas de pesquisa de manipulação, propagação e visualização de dados. No entanto, a nova geração de BI tem um foco adicional de pesquisa em áreas como exploração e visualização de dados (Obeidat et al., 2015). Há também evidências de mudança

de relatórios estáticos para visualizações interativas, o que estende questões de pesquisa da visão geral das métricas à descoberta de causas e efeitos dos fenômenos expressos pelas métricas. Além disso, a pressão da concorrência nos negócios causa novas tendências em BI e pesquisas relacionadas, tais como BI quase em tempo real, Mineração de dados e Análise de Texto, BI de autoatendimento, BI na nuvem (Dedić & Stanier, 2016) e *Big Data*.

2.3 BIG DATA

Nas últimas décadas, os avanços nos dispositivos de coletas de dados digitais e nas tecnologias de armazenamento impulsionou a criação de grandes volumes de dados. Estamos vivendo uma era de dilúvio de dados e, como resultado, o termo "*Big Data*" está aparecendo em muitos contextos (Sowmya & Suneetha, 2017).

O *Big Data* pode ser conceituado ou representado por meio de 5 "V"s: Volume, Variedade, Velocidade, Veracidade e Valor (Taurion, 2013; Cielen et al., 2016).

- Volume: Diariamente são geradas quantidades enormes de dados. Sejam em sistemas empresariais ou mesmo por meio de usuários em seus computadores pessoais ou dispositivos móveis.
- Variedade: Esses dados gerados são provenientes de sistemas estruturados e não estruturados. Os dados de sistemas estruturados são minoria, enquanto os não estruturados fazem parte da imensa maioria desses dados gerados através de envio de e-mails, redes sociais (Twitter, Facebook, Blog's, Youtube e outros), documentos eletrônicos, câmeras de vídeo, fotos, mensagens de voz e etc.
- Velocidade: Muitas vezes, para se fazer uso de uma maneira eficaz, é necessário que essas informações sejam analisadas praticamente em tempo real.
- Veracidade: É necessário ter a certeza de que os dados analisados são autênticos e tenham segurança, para que as conclusões e tomadas de decisões provenientes das informações geradas por meio da análise desses dados sejam assertivas.
- Valor: Para se implementar um projeto relacionado a *Big Data* em uma organização (empresa, comércios, área acadêmica, saúde e etc), é absolutamente necessário que a solução implementada traga retorno dos investimentos feitos, do contrário isso resultaria em um grande prejuízo e desperdício de recursos da organização. Um exemplo se aplica na área de

> seguros, na qual uma análise de fraudes pode se tornar mais ágil e imensamente melhor, minimizando riscos e fazendo o uso de análise de dados que não se encontram nas fontes de dados estruturadas que as seguradoras fornecem, tais como, por exemplo, dados que estão circulando em diversas mídias sociais (Disner, 2015).

Em suma, dados maciços e sempre crescentes são úteis apenas se puderem ser analisados (Henry & Venkatraman, 2015). Visto no passado como um problema técnico, o *Big Data* é hoje visto como uma oportunidade de negócios, que pode oferecer novas oportunidades com base na análise de dados (Rosemary Williams, 2015). O principal desafio do *Big Data* é explorar grandes dados com o objetivo de extrair informações úteis e conhecimento competitivo. Desbloquear o valor do Big Data em mercados complexos e em rápida mudança pode trazer vantagem competitiva e permitir uma melhor resposta das empresas (Martínez-Plumed et al., 2019).

2.4 DATA ANALYTICS

Conforme dito na seção anterior, a proliferação do *Big Data*, por si só, não é útil. Os benefícios reais estão na análise dos dados e no uso dos padrões que estes revelam na tomada de decisões. Diante deste contexto, o termo *Data Analytics* (DA) tem ganhado força nos últimos anos. Isto implica necessidade de profissionais de inteligência com conhecimentos de matemática, Mineração de Dados, amostragem e estatística inferencial, diante da inviabilidade de se analisar todos os dados disponíveis (Lalanne, 2016).

Nesse âmbito, enquanto técnicas de *Data Mining* identificam padrões e informações ocultas que auxiliam a consecução procedimental no âmbito investigativo, em paralelo, técnicas de *Data Analytics* podem se concentrar apenas na seleção das amostras verídicas e valorosas do *Big Data*, bem como nas inferências, derivando conclusões baseadas no que já foi descoberto pelo investigador (Nunes et al., 2019).

A análise fornecerá informações úteis sobre os problemas de negócios e talvez até faça sugestões sobre quando e onde os problemas futuros ocorrerão (análise preditiva), para que os problemas possam ser evitados ou pelo menos atenuados. A maioria das grandes organizações do mundo, tais como Apple, GE, Walmart, Exxon e Samsung, possui operações globais (fábricas, armazéns, transportadores e clientes) e atende a vários clientes com uma ampla variedade de produtos e serviços. É difícil desvendar a complexidade dessas redes vastas e altamente conectadas, o que exige o auxílio de algoritmos baseados em modelos matemáticos

e estatísticos para descobrir onde e por que os problemas ocorrem (Henry & Venkatraman, 2015).

2.5 EXPERIMENTAÇÃO EM SOFTWARE

A rápida entrega de valor aos clientes é uma das principais prioridades das empresas de software (Fagerholm et al., 2017). Com esse objetivo em mente, as empresas geralmente desenvolvem suas práticas de desenvolvimento. Inicialmente, estas herdam os princípios *Agile* na parte de desenvolvimento da organização (Martin, 2002) e os expandem para outros departamentos (Olsson, Alahyari & Bosch, 2012). Em seguida, as empresas se concentram em vários conceitos enxutos, tais como eliminar o desperdício, remover restrições no *pipeline* de desenvolvimento (Goldratt & Cox, 2016) e avançar para a integração (Dittrich et al., 2018) e a implantação contínuas da funcionalidade do software (Rodríguez et al., 2017). No entanto, a implantação contínua é caracterizada por um canal bidirecional que permite às empresas não apenas enviar dados aos seus clientes para prototipar rapidamente (Singh, 2016), mas também receber dados de feedback dos produtos em campo.

Neste contexto, a intuição das empresas de desenvolvimento de software sobre as preferências do cliente pode estar errada em até 90% das vezes (Clancy, 1995; Castellion, 2008; Manzi, 2019). Para mitigar isso, os dados atuais de uso de um produto têm o potencial de tornar o processo de priorização no desenvolvimento de novos produtos mais preciso, pois permitem a concentração no que os clientes fazem e não no que dizem (Bosch-Sijtsema & Bosch, 2015). Neste sentido, a experimentação está se tornando a norma nas empresas de software avançadas, para a avaliação confiável de ideias com os clientes, a fim de priorizar corretamente as atividades de desenvolvimento de produtos (Olsson & Bosch, 2014; Kohavi & Longbotham, 2017).

Desta forma, existe uma crescente compreensão na comunidade de que estudos empíricos são necessários para desenvolver ou melhorar processos, métodos e ferramentas para desenvolvimento e manutenção de software (Basili, 1996; Endres & Rombach, 2003; Fagerholm et al., 2017).

A pesquisa em engenharia empírica de software deve ter como objetivo adquirir conhecimentos gerais sobre qual tecnologia (processo, método, técnica, linguagem ou ferramenta) é útil, para quem é útil, na realização de quais tarefas (engenharia de software), em quais ambientes. Portanto, essa pesquisa se concentra no tipo de tecnologia que está sendo estudada nos experimentos investigados (que reflete os tópicos dos experimentos), nos sujeitos que participaram, nas tarefas que eles executaram, no tipo de sistemas de aplicativos nos quais

essas tarefas foram executadas e os ambientes em que os experimentos foram conduzidos. Além disso, também deve inclur dados sobre replicação de experimentos e até que ponto a validade interna e externa é discutida (Sjøberg et al., 2005).

Uma categoria importante de estudo empírico é a do experimento controlado, cuja condução é regida pelo método científico clássico, para identificar relações de causa-efeito. Em um experimento controlado, os usuários são divididos aleatoriamente entre as variantes (por exemplo, os dois designs diferentes de uma interface de produto) de maneira persistente (um usuário recebe a mesma experiência várias vezes). As interações dos usuários com o produto são instrumentadas e as principais métricas são computadas (Kohavi & Longbotham, 2017).

Um dos principais desafios das métricas é decidir sobre o que incluir em um Critério de Avaliação Geral (Overall Evaluation Criteria - OEC). Um OEC é uma medida quantitativa do objetivo de um experimento controlado (Roy, 2001) e orienta a direção do desenvolvimento de negócios. Na experimentação controlada, é intuitivo medir o efeito a curto prazo, ou seja, o impacto observado durante o experimento (Hohnhold, O'brien & Tang, 2014). Fornecer mais peso às métricas de publicidade, por exemplo, torna as empresas mais lucrativas no curto prazo. No entanto, o efeito a curto prazo nem sempre é preditivo do efeito a longo prazo e, consequentemente, não deve ser o único componente de uma OEC (Kohavi et al., 2014). Definir um OEC não é trivial e deve ser realizado com muito cuidado. Kohavi et al. (2009, 2014, 2017), em seus trabalhos, apresentam armadilhas comuns no processo de estabelecimento de um sistema de experimentação controlado e orientações sobre como definir de forma confiável um OEC.

Apresentado o Glossário Teórico, será explanada, no próximo capítulo, a Revisão *Quasi*-Sistemática da Literatura.

3.0 REVISÃO *QUASI*-SISTEMÁTICA

Neste capitulo, será apresentada parte do artigo intitulado: Quão Experimentais e Estratégicas são as Aplicações de *Business Intelligence* (BI) e *Data Mining*?

Objetivo do Trabalho: Identificar e caracterizar as metodologias utilizadas para o desenvolvimento experimental de aplicações inteligentes alinhadas ao planejamento estratégico.

Metodologia: Uma revisão *quasi*-sistemática foi realizada, para caracterizar a pesquisa na área, considerando os últimos dez anos.

Originalidade: Não foram encontrados trabalhos científicos com o mesmo objeto de pesquisa deste artigo, de identificar e caracterizar as metodologias para o desenvolvimento experimental de aplicações inteligentes alinhadas ao planejamento estratégico, o que aumenta a importância dos resultados aqui apresentados.

Principais Resultados: Como resultados, não foram encontrados trabalhos que apresentassem alguma abordagem completa para disciplinar o alinhamento estratégico e a experimentação, prevendo atendimento claro aos objetivos estratégicos e uma fase experimental na validação dos resultados. No entanto, alguns ensaios de partes dessas características puderam ser mapeados, como, por exemplo, a experimentação, encontrada em 28,57% dos trabalhos. Entre os países, a China, os Estados Unidos e o Brasil lideraram o ranking de publicações sobre o tema. Quanto ao meio de publicação, o *Journal* foi a opção mais utilizada para publicação. Além disso, a conferência "*IEEE International Conference on Advanced Communications, Control and Computing Technologies*" e o periódico "*Expert Systems with Applications*", destacaram-se como maiores publicadores.

Contribuições Teóricas: Esta pesquisa apresenta resultados relevantes à academia e aos empreendedores, fornecendo evidências de que não há um método formal de desenvolvimento experimental de aplicações de BI e *Data Mining* voltado ao planejamento estratégico de uma organização. Além disso, este trabalho apresenta-se como uma fonte de consulta aos padrões de métodos existentes para o desenvolvimento de aplicações inteligentes, bem como pode ser replicado e estendido, pela sistematização aplicada. Por fim, há o direcionamento para

pesquisas que proponham métodos de criação de aplicações inteligentes validadas experimentalmente e alinhadas à estratégia.

Palavras-chave: Alinhamento Estratégico. *Business Intelligence*. *Data Mining*. Mineração de Dados. *Data Science*. Ciência de Dados.

3.1 TRABALHOS RELACIONADOS

Não foram encontrados trabalhos científicos com o mesmo objeto de pesquisa deste artigo, de identificar e caracterizar as metodologias para o desenvolvimento experimental de aplicações inteligentes alinhadas ao planejamento estratégico, o que aumenta a importância dos resultados aqui apresentados. No entanto, alguns trabalhos mencionam a importância desse alinhamento estratégico no desenvolvimento de tais aplicações.

Com relação aos trabalhos avaliados por esta revisão, ou seja, trabalhos que lidaram especificamente com *Business Intelligence* e *Data Mining*, alguns trabalhos pontuaram a importância do alinhamento estratégico, tais como os trabalhos de Sharma, Osei-Bryson & Kasper (2012) e Kohavi et al. (2013). Além disso, muitos trabalhos na literatura destacam a importância do alinhamento estratégico de aplicações de tecnologia da informação em geral. São exemplos os trabalhos de: Isaca (2018), Weber & Klein (2013), Medeiros Júnior et al. (2017) e Araújo & Dornelas (2017). A seguir, serão sintetizados outros trabalhos relevantes à temática específica aqui abordada.

Mola et al. (2015) fizeram um estudo exploratório que analisa os efeitos das características técnicas e organizacionais dos sistemas de BI nos processos de compartilhamento de conhecimento, colaboração e tomada de decisão. Em média, as características técnicas e organizacionais dos Sistemas de BI estão positivamente associadas a um aumento no compartilhamento de conhecimento, levando a uma melhoria na colaboração interna e na qualidade de tomada de decisão. Estas melhorias dependem da forma como o BI é projetado.

Em um *survey* realizado no Brasil por Lima et al. (2017), foi constatado que 67,50% das empresas não utilizam uma metodologia experimentada para o desenvolvimento de BI, o que contribui para que os projetos não obtenham sucesso. Associado a este resultado, um percentual de 72,00% das empresas não utiliza uma metodologia de alinhamento estratégico. A ausência de uma metodologia alinhada à estratégia da empresa evidencia que os gestores podem estar tomando decisões com base em informações não relevantes à instituição ou desalinhadas às estratégias de negócio.

Em outro *survey* similar sobre *Business Intelligence*, Duan e Xu (2012) apresentam uma introdução sobre BI, com ênfase em algoritmos fundamentais para o uso de BI em ambientes empresariais, destacando os desafios e oportunidades encontrados nesses ambientes.

Colaço Júnior et al. (2019) apresentaram um processo que mescla a abordagem GQM+*Strategies* com uma metodologia de desenvolvimento ágil de aplicações de *Business Intelligence* proposta pelo autor, visando garantir o alinhamento estratégico. O processo proposto foi avaliado por meio de um estudo de caso, em uma empresa multinacional latino-americana do mercado de varejo, no qual foi evidenciado que é possível integrar a abordagem de alinhamento estratégico adotada com uma metodologia de desenvolvimento de aplicações de BI. Com as boas evidências iniciais, os pesquisadores poderão evoluir o processo e prever o uso de experimentação para validação dos modelos inteligentes que poderão ser criados com técnicas de *Data Mining* e IA.

3.2 MÉTODO

Com o objetivo de identificar e caracterizar as metodologias de desenvolvimento de aplicações de *Business Intelligence* e *Data Mining* dirigidas à estratégia e/ou que preveem avaliação experimental, foi adotada para este trabalho a metodologia de Revisão *quasi*-Sistemática da Literatura (*quasi*-RSL), a qual foi conduzida valendo-se das definições apresentadas no capítulo de introdução e das diretrizes para realizar revisões sistemáticas definidas por Kitchenham et al. (2009).

Este método foi adotado por se pautar nos conceitos da Medicina baseada em Evidência (*Evidence-based Medicine*), área madura no processo de revisões sistemáticas, e por propor uma mudança de paradigma em como as pesquisas na área de Software deveriam ser conduzidas. Segundo Kitchenham et al. (2009), a pesquisa em Medicina mudou de forma drástica com o paradigma baseado em evidências, possibilitando uma organização mais efetiva da pesquisa médica e embasando o julgamento clínico de especialistas. O sucesso deste novo paradigma influenciou fortemente a adoção da abordagem baseada em evidências em outras áreas do conhecimento tais como psicologia, enfermagem, ciências sociais, educação e computação.

As próximas seções detalham os passos para replicação desta revisão, englobando as *strings* e comandos utilizados, as bases pesquisadas, bem como os critérios de seleção dos artigos e extração dos dados. Para extração, foi explicitamente descrito como os trabalhos foram classificados como experimentais e/ou estratégicos, bem como as variáveis a serem identificadas e seus valores. A seguir, são listadas as fases resumidas do método.

1. **Planejamento da Revisão:** os objetivos da pesquisa são listados e o protocolo da revisão é definido;
2. **Condução da Revisão:** nesta atividade, as fontes para a revisão sistemática são selecionadas, os estudos primários são além de identificados, selecionados, avaliados de acordo com os critérios de inclusão, exclusão, e de qualidade estabelecidos durante o protocolo da revisão;
3. **Análise e Publicação dos Resultados:** os dados dos estudos são extraídos e sintetizados para serem publicados.

3.3 PLANEJAMENTO DA REVISÃO *QUASI*-SISTEMÁTICA

3.3.1 OBJETIVO

Como dito anteriormente, esta revisão tem como objetivo identificar e caracterizar as metodologias de desenvolvimento de aplicações de *Business Intelligence* e *Data Mining* dirigidas à estratégia e/ou que preveem avaliação Experimental.

3.3.2 QUESTÕES DE PESQUISA

As questões de pesquisa foram elaboradas tendo como base a abordagem PICO (Bergin & Wraight, 2006; Costa et al., 2007). Este modelo, que surgiu na execução de estudos clínicos, estrutura a pesquisa em quatro elementos básicos: População, Intervenção, Comparação (ou controle) e "*Outcomes*" (Resultados). Apesar de ter surgido na medicina, essa estratégia pode ser adaptada para outras áreas, como demonstra o trabalho de Kitchenham (2004), usado como referência para a produção da Tabela 1. A tabela também apresenta alguns artigos de controle, os quais foram selecionados com uma consulta preliminar nas bases de pesquisa, com a seleção dos dois primeiros artigos encontrados, alinhados com a intervenção de interesse. Essa busca preliminar serviu de base para encontrar outras palavras-chave importantes, presentes em artigos verdadeiros positivos que serão resultados finais da revisão, e refinar a *string* de busca. Um artigo de controle pode inclusive já ser de conhecimento dos pesquisadores e é um elemento para ajudar a criar e validar a *string*. Neste sentido, foi também usado um artigo de controle que é efetivamente de comparação, ou seja, não se enquadrava totalmente na intervenção de interesse e não está indexado nas bases pesquisadas, mas apresenta uma proposta de alinhamento estratégico. Artigos de controle deste tipo também ajudam a estabelecer uma *string* de busca mais assertiva.

Tabela 1 - Modelo PICO para conformidade das questões de pesquisa

Acrônimo	Definição	Descrição

P	População	Publicações de pesquisadores e desenvolvedores, tendo em vista o desenvolvimento de aplicações de *Business Intelligence* com ou sem o apoio de *Data Mining*.
I	Intervenção	Métodos de desenvolvimento de aplicações de BI e *Data Mining* dirigidos à estratégia e/ou que usaram avaliação experimental. A direção estratégica pressupõe métodos que partem da identificação de objetivo(s) estratégico(s) que será(ão) alavancado(s) pelo produto final, evitando a criação de soluções que não estão alinhadas com a estratégia organizacional e não ajudam diretamente a realização dos objetivos traçados.
C	Controle	Métodos de criação de aplicações de BI ou de descoberta de conhecimento, como, por exemplo, o modelo de processo de Data Mining: *Cross Industry Standard Process for Data Mining* (CRISP-DM). Artigos de Controle Artigos que se enquadram na intervenção: • *Evaluation of an integrated knowledge discovery and data mining process model*; • *Pattern based feature construction in semantic data mining.* Artigo de comparação: • Proposta e Avaliação de um Processo para o Desenvolvimento de Aplicações de Business Intelligence Dirigido à Estratégia
O	Resultado	Metodologias de BI e Data Mining que validam suas conclusões por meio de experimentos controlados e/ou definem explicitamente a direção da aplicação para um ou mais objetivos estratégicos.

Assim, a partir da definição do PICO, as seguintes questões de pesquisa foram elaboradas:

- Q1: Quais as metodologias dirigidas à estratégia utilizadas no desenvolvimento de aplicações de BI e de *Data Mining*?
- Q2: Como é feito este alinhamento entre o Planejamento Estratégico e o desenvolvimento de aplicações de BI e *Data Mining*?
- Q3: As metodologias de desenvolvimento de BI e *Data Mining* preveem uma fase de avaliação experimental com validação de conclusões por meio de testes estatísticos apropriados?

- Q4: Quais países possuem mais pesquisadores publicando sobre esse tema?
- Q5: Quais os anos que tiveram mais publicações nessa área?
- Q6: Quais os principais periódicos e conferências sobre o tema?
- Q7: Quais os meios de publicações?
- Q8: Quais os tipos de estudos?

Estas questões foram formuladas tendo como base as diretrizes do protocolo de Revisão Sistemática da Literatura (Kitchenham, 2004; Petersen et al., 2008; Petersen et al., 2015). Além da visão geral, pretendeu-se averiguar se as metodologias usadas relacionavam explicitamente os requisitos da aplicação com objetivos estratégicos e se possuíam uma fase para validar as soluções experimentalmente.

3.3.3 ESTRATÉGIA DE BUSCA E DE SELEÇÃO

Para a execução da busca, foram consultadas as bases de dados: ACM Digital Library (ACM), IEEE Xplore (IEEE) e SCOPUS. As buscas foram realizadas utilizando as ferramentas de filtragem disponibilizadas em cada base de dados citada anteriormente, considerando nas buscas: título, resumo/abstract e palavras-chave dos respectivos artigos. A respeito do idioma, foram selecionados apenas trabalhos em inglês. A respeito da área, foram selecionados apenas trabalhos referentes à Ciência da Computação. E a respeito do tempo de publicação, foram selecionados apenas trabalhos publicados a partir de 2008.

Para realizar a pesquisa nas bases digitais, foi definida uma *string* de busca com a utilização de termos em inglês e do uso de vários sinônimos, associados ao pressuposto de que os estudos estariam contidos nas áreas da computação que lidam com metodologias de desenvolvimento de aplicações de BI e *Data Mining*. Tais termos foram identificados com auxílio dos artigos de controles do modelo PICO, descritos na seção anterior (Tabela 1), e posteriormente, refinados e adaptados para o maior aproveitamento da *string*. A Tabela 2 mostra os termos, antes de refiná-los, que foram selecionados.

Tabela 2 - Categorias do modelo PICO e termos identificados para pesquisa bibliográfica antes de refiná-los

Categoria	Descrição
População	*Development, Implementation, Construction, Deployment, Creation, Business Intelligence, Data Mining*.
Intervenção	*Methodology, Method, Approach, Process, Experiment*.

Controle	Métodos de criação de aplicações de BI e de descoberta de conhecimento, como, por exemplo, o CRISP-DM (sem *strings*).
Resultado	*Strategy Oriented, Strategy Driven, Strategic Alignment, Hypothesis Testing, Statistical Validation, Statistical Analysis, Control Experiment, Controlled Experiment, Experimental Analysis, Experimental Evaluation, Statistical Test, Formal Experiment, Null Hypothesis, Primary Hypothesis, Statistical Significance*.

Após o refinamento, os termos ajustados foram utilizados para construir a *string* de busca, os quais estão descritos na Tabela 3.

Tabela 3 - *Strings* eleitas após o refinamento

Termos da *string* de busca		
Development, Implementation, Construction, Deployment, Creation, Business Intelligence, Data Mining.	*Methodology, Method, Approach, Process, Experiment*.	*Strategy Oriented, Strategy Driven, Strategic Alignment, Hypothesis Testing, Statistical Validation, Statistical Analysis, Control Experiment, Controlled Experiment, Experimental Analysis, Experimental Evaluation, Statistical Test, Formal Experiment, Null Hypothesis, Primary Hypothesis, Statistical Significance*.

A string de pesquisa gerada com os termos evidenciados acima foi:

- *TITLE-ABS-KEY(("Method*" OR "Approach" OR "Process" OR "Experiment*") AND ("Development" OR "Implementation" OR "Construction" OR "Deployment" OR "Creation") AND ("Business Intelligence" OR "Mining") AND ("Strategy Oriented" OR "Strategy-Oriented" OR "Strategy Driven" OR "Strategy-Driven" OR "Strategic Alignment" OR "Hypothes* Test*" OR "Statistic* Valid*" OR "Statistic* Analy*" OR "Contro* Experiment*" OR "Experiment* Analy*" OR "Experimen* Evaluation" OR "Statisti* Test*" OR "Formal Experiment*" OR "Null Hypothes*" OR "Primary Hypothes*" OR "Statisti* Significan*")) AND SUBJAREA(COMP) AND (PUBYEAR > 2008).*

A busca por pesquisas em computação foi considerada com base em artigos de controle que abrangiam mais de uma área, uma vez que a computação é um meio para a gestão

e tomada de decisão, e, além disso, considerou-se que o objetivo foi explorar a área que desenvolve a parte técnica da Tecnologia da Informação, a qual já deveria ter como padrão o alinhamento estratégico e a experimentação, ou seja: Isso já é uma realidade?

3.3.4 CRITÉRIOS DE SELEÇÃO DE FONTES

Para filtrar os artigos relevantes para esta revisão, foram estabelecidos os critérios de inclusão e exclusão dos mesmos. Depois de todas as fases, o estudo contabilizou os artigos que focassem no uso de alguma metodologia para o desenvolvimento de aplicações de BI e DM voltada ao planejamento estratégico ou que usaram avaliação experimental, utilizando os seguintes critérios preliminares de inclusão:

1. Resultado deve conter o tema deste estudo, já automaticamente limitado pela *string*, no título, resumo ou palavras-chave;
2. O resultado deve datar entre os anos de 2009 a 2019;
3. O resultado deve apresentar uma avaliação experimental ou explorar alguma metodologia, método, processo ou abordagem de desenvolvimento de aplicações de BI ou *Data Mining*. Isso foi feito para gerir o risco de não haver metodologias dirigidas à estratégia ou experimentais e pelo menos serem listados trabalhos que abordaram metodologias de desenvolvimento de BI ou *Data Mining* ou usaram alguma avaliação experimental;
4. O resultado deve estar disponível para consulta online.

A confirmação dos critérios de inclusão foi dada após análise do resumo de cada um dos artigos encontrados, na primeira filtragem, seguida de uma leitura completa, para segunda filtragem. Antes da primeira leitura de cada artigo, foi realizada a análise quanto aos critérios de exclusão. Foram eliminados:

1. Estudos secundários, pois eles tratam de abordagens de terceiros;
2. Publicações duplicadas;
3. Artigos Curtos (*Short Papers*);
4. *Surveys*.

3.3.5 ESTRATÉGIA DE EXTRAÇÃO DE INFORMAÇÕES

A estratégia de extração de dados é projetada para reunir as informações necessárias e responder às questões de pesquisa, avaliando a qualidade do trabalho. Isto posto, após o

término da etapa de seleção, os trabalhos definidos serão lidos na íntegra, logo em seguida, um formulário será respondido, conforme Tabela 4, contendo informações sobre o conteúdo abordado em cada trabalho. Sendo assim, este formulário nos permite avaliar os trabalhos de forma mais detalhada e precisa, realizando a classificação baseado nos critérios de inclusão e exclusão.

No que concerne à identificação de alinhamento estratégico, não foi suficiente a citação do termo e/ou o destaque para sua importância em algum ponto do artigo. Desta forma, para ser classificado como um trabalho que usou alguma abordagem de alinhamento, foi necessário que a solução proposta pelos autores considerasse como premissa ou base pelo menos um objetivo estratégico explícito.

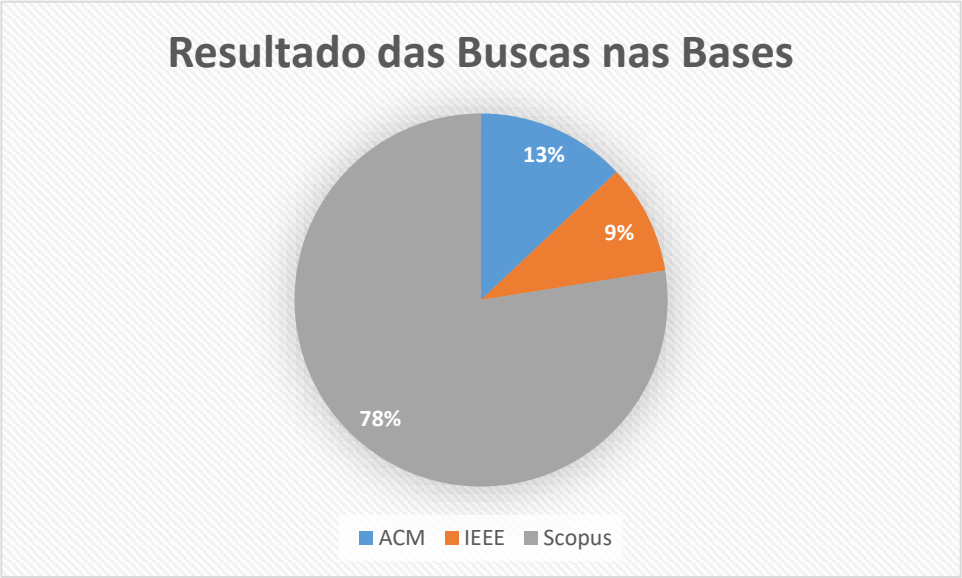
Tabela 4: Formulário de Extração

1.	Qual a metodologia utilizada?	
2.	Como é feito o alinhamento estratégico?	
3.	O artigo se referia a *Business Intelligence*?	[Sim, Não]
4.	O artigo se referia a *Data Mining*?	[Sim, Não]
5.	Qual o tipo de estudo utilizado?	[Aplicação Prática, Estudo de Caso, ...]
6.	O estudo foi dirigido à estratégia?	[Sim, Não]
7.	O estudo possui alguma avaliação experimental?	[Sim, Não]
8.	Foi feito o cálculo do tamanho da amostra?	[Sim, Não]
9.	Foi feito o teste de normalidade?	[Sim, Não]
10.	Foi declarada hipótese formalmente?	[Sim, Não]
11.	Foi calculado o intervalo de confiança?	[Sim, Não]
12.	Foram declaradas as ameaças à validade?	[Sim, Não]

3.4 CONDUÇÃO DA REVISÃO *QUASI*-SISTEMÁTICA

O planejamento da revisão *quasi*-sistemática foi elaborado entre os meses de fevereiro a março de 2019, já a execução, ocorreu em abril do mesmo ano. Para a obtenção dos estudos primários, foi necessária a formação da *string* de busca a partir das combinações das palavras-chave em inglês. Assim, a *string* base foi definida na máquina de busca Scopus, refinada e, quando julgado que a *string* era adequada, foi traduzida para as máquinas de busca ACM e IEEE. Logo em seguida, as buscas foram realizadas. No total, foram retornados 841 trabalhos, sendo 652 (78%) do Scopus, 109 (13%) da ACM e 80 (10%) do IEEE, como mostra a Figura 1.

Figura 1 - Resultado da Execução das Strings de Busca nas Bibliotecas Digitais

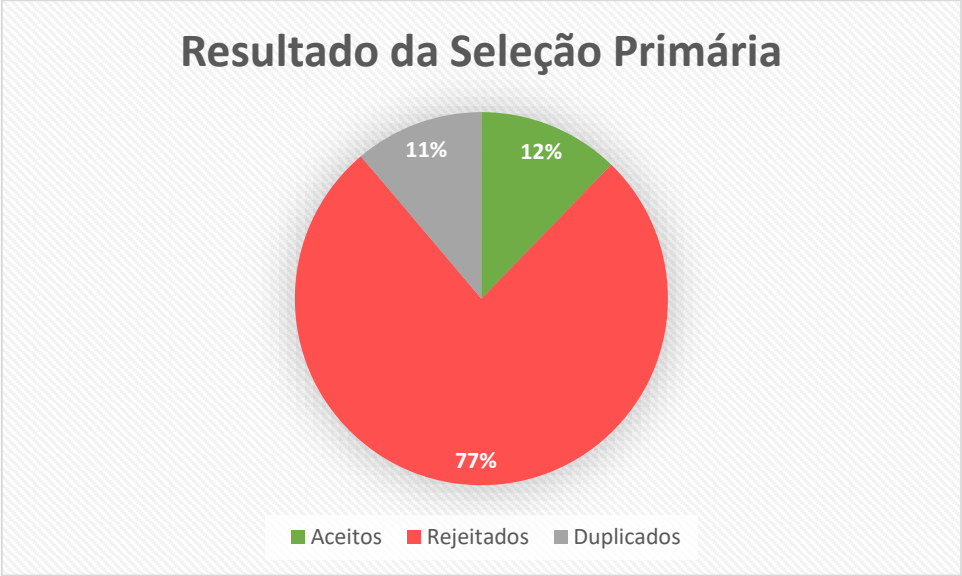


Com a finalização da busca, houve o início do processo de filtragem dos artigos encontrados, com base nos critérios de seleção, definidos na seção 3.3.4. Nesta fase, os trabalhos foram classificados em Aceito, Rejeitado e Duplicado.

Neste sentido, do total de 841 publicações analisadas, 94 (11% do total) eram duplicadas e, consequentemente, as duplicatas acabaram sendo eliminadas. Assim, 747 trabalhos foram selecionados, para uma avaliação superficial, na qual foi realizada a leitura de todos os títulos e *abstract*, aplicando os critérios de inclusão e exclusão, definidos previamente no protocolo. Ao final desta etapa, foram identificados 644 (77%) trabalhos que estavam fora do escopo desta revisão e por isso foram rejeitados. Estes trabalhos eram artigos curtos, revisões de literatura, *surveys* ou artigos fora dos critérios. Os aceitos mencionavam a realização de um experimento ou mencionavam o uso ou proposta de uma abordagem, processo, metodologia ou

método. Como resultado final, foram aceitos 103 (12%) trabalhos para serem avaliados de forma mais precisa e detalhada. A Figura 2 ilustra os resultados dessa primeira seleção.

Figura 2 – Resultado da Primeira Etapa de Seleção dos Trabalhos



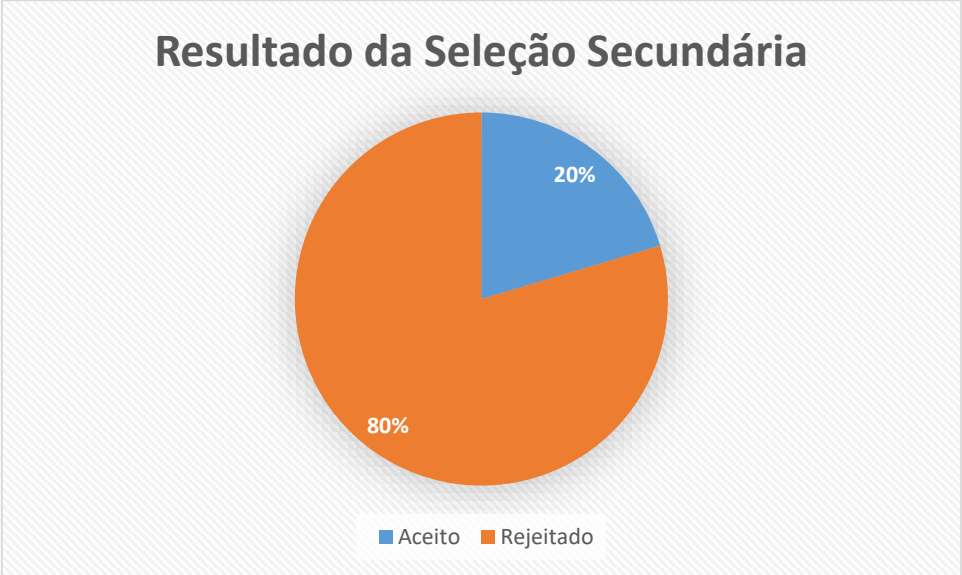
Após essa primeira etapa de seleção, foi feita uma avaliação precisa e detalhada a partir da leitura completa dos trabalhos, de forma a confirmar o indício da presença de alinhamento estratégico ou de avaliação experimental. No entanto, não foi possível acessar todas as publicações, mesmo após contatar os autores por e-mail. Dos 103 artigos que deveriam ter sido avaliados nessa etapa, foram recuperados 100 (97,09%). Dessa forma, não foi possível acessar o texto completo de 3 (2,91%) artigos, que consequentemente acabaram sendo rejeitados.

Para enquadrar os estudos como experimentos controlados e classificar a presença de uma avaliação experimental, foram considerados os trabalhos nos quais a base do método científico foi verificada, ou seja, aqueles que formalizaram a definição das hipóteses e efetuaram a validação estatística necessária para teste.

Durante esta etapa, foi identificado que 82 (80%) trabalhos estavam fora do escopo desta revisão e por isso foram rejeitados. Todos os artigos que confirmaram a realização de um experimento foram novamente aceitos. Fora estes, só foram aceitos artigos que o método, metodologia, processo ou abordagem mencionados estavam relacionados à concepção de aplicações de BI ou *Data Mining*, ou seja, BI ou *Data Mining* não eram assuntos transversais do artigo. Como resultado final, foram aceitos 21 (20%) trabalhos. Para estes artigos, finalmente, uma nova leitura norteou o preenchimento do formulário de extração, descrito na

seção 4.1.5. A Figura 3 ilustra o resultado final dessa segunda fase de seleção e extração dos dados.

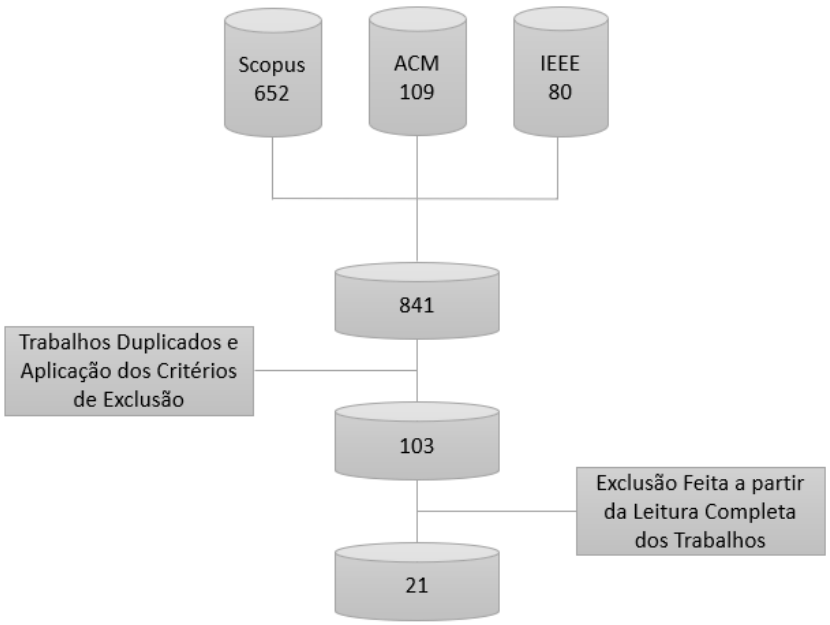
Figura 3 – Resultado da Segunda Etapa de Seleção dos Trabalhos



3.5 SÍNTESE DOS DADOS

Nesta seção, os resultados da revisão *quasi*-sistemática são apresentados. A Figura 4 mostra um resumo do número de trabalhos obtidos em cada etapa do processo de busca e em seguida as questões de pesquisa são respondidas de acordo com os dados extraídos.

Figura 4 – Resultados Obtidos Durante o Processo de Busca



3.5.1 QUAIS AS METODOLOGIAS DIRIGIDAS À ESTRATÉGIA UTILIZADAS NO DESENVOLVIMENTO DE APLICAÇÕES BI E DE DATA MINING?

Não foram encontrados trabalhos que apresentassem alguma abordagem para disciplinar o alinhamento estratégico. No entanto, alguns trabalhos mencionam a importância deste alinhamento ou propõem alguma metodologia para o desenvolvimento eficiente deste tipo de aplicação, a exemplo de Sharma, Osei-Bryson & Kasper (2012), Lin et al. (2017), Cheng et al. (2009), Ju et al. (2018), Kohavi et al. (2013), Manigandan et al. (2019), Ławrynowicz & Potoniec (2014) e Wang & Sun (2013).

3.5.2 COMO É FEITO ESTE ALINHAMENTO ENTRE O PLANEJAMENTO ESTRATÉGICO E O DESENVOLVIMENTO DE APLICAÇÕES DE BI E DATA MINING?

Conforme dito anteriormente, não foi possível identificar métodos de desenvolvimento de ambas as aplicações, BI e *Data Mining*, com previsão de alinhamento estratégico, excetuando o trabalho de Colaço et al. (2019), cujo escopo é apenas para BI. No entanto, podemos destacar alguns trabalhos que propõem alguma metodologia para o desenvolvimento eficiente deste tipo de aplicação.

Sharma, Osei-Bryson & Kasper (2012) abordam em seu trabalho as várias limitações identificadas nos modelos de processos existentes de *Data Mining* e propõem resolvê-las por meio da proposta de um novo modelo melhorado, denominado, Modelo Integrado de Descoberta de Conhecimento e *Data Mining* (IKDDM), que apresenta uma visão integrada do processo KDDM (Knowledge Discovery and Data Mining – Descoberta de Conhecimento e Mineração de Dados) e fornece suporte explícito para a execução de cada uma das tarefas descritas no modelo. Também foi avaliada a eficácia e a eficiência oferecidas pelo modelo IKDDM contra o CRISP-DM, um modelo líder no processo de KDDM. Os resultados dos testes estatísticos indicaram que o modelo IKDDM supera o modelo CRISP-DM em termos de eficiência e eficácia. O modelo IKDM também superou o CRISP-DM em termos de qualidade do próprio modelo de processo.

Cheng et al. (2009) apresentam uma abordagem baseada em ontologias para aplicações de BI, especificamente em Análise Estatística e Data Mining. Implementando a abordagem em um Sistema de Gestão do Conhecimento Financeiro (FKMS), que é capaz de: (i) extração, transformação e carregamento de dados, (ii) criação e recuperação de cubos de dados, (iii) análise estatística e mineração de dados, (iv) gerenciamento de metadados de experimentos; (v) recuperação de experimentos para nova resolução de problemas. O conhecimento resultante de cada experimento, definido como um conjunto de conhecimento

que consiste em sequências de dados, modelo, parâmetros e relatórios, é armazenado, compartilhado, disseminado e, portanto, útil para apoiar a tomada de decisões.

Assim como Cheng et al. (2009), Ławrynowicz & Potoniec (2014) propõem uma abordagem de Data Mining, na qual as ontologias de domínio são usadas como conhecimento de fundo. Ao invés de usarem apenas dados puramente empíricos, os autores também desenvolveram uma ferramenta que implementa essa abordagem. Desta forma, foi conduzida uma avaliação experimental, comparando o método proposto com abordagens de ponta para a classificação de dados semânticos.

Ju et al. (2018) propõem uma estrutura para o uso de análise de big data centrada no cidadão, para impulsionar a inteligência de governança em cidades inteligentes, por meio de duas perspectivas: questões de governança urbana e algoritmos de análise de dados. A estrutura consiste em três camadas: 1) A camada de mesclagem de dados, que constrói os dados panorâmicos centrados no cidadão, para cada cidadão, mesclando dados relacionados a cidadãos de várias fontes na governança urbana colaborativa, por meio de cálculos de similaridade e resolução de conflitos; 2) uma camada de descoberta de conhecimento, que traça o perfil do cidadão, em nível individual e de grupo, em termos de prestação de serviços públicos urbanos e participação do cidadão por meio de técnicas simples de análise estatística, aprendizado de máquina e métodos econométricos; e 3) uma camada de tomada de decisão, que usa modelos de ontologia para padronizar atributos relacionados à governança, pessoas e associações para apoiar a governança de tomada de decisão, por meio de mineração de dados e técnicas de Rede Bayesiana. A estrutura proposta é validada em um estudo de caso sobre a governança da doação de sangue na China.

Manigandan et al. (2019) propõem o algoritmo M-Clustering, o qual fornece uma solução para a mineração de dados usando clusters. O algoritmo proposto foi avaliado, comparando a eficiência de processamento de dados experimentais em relação ao K-Means.

Wang & Sun (2013) propõem, com base na arquitetura orientada a serviços e computação na nuvem, uma plataforma de Sistema de Informação Geográfica de recursos hídricos e de energia elétrica. O objetivo da plataforma é gerenciar os diversos e massivos dados de forma eficiente, com base na construção da estrutura fundamental de big data de pesquisa, projeto, construção, ambiente, imigração, equipamentos e suprimentos.

3.5.3 AS METODOLOGIAS DE DESENVOLVIMENTO DE BI E DATA MINING PREVEEM UMA FASE DE AVALIAÇÃO EXPERIMENTAL COM VALIDAÇÃO DE CONCLUSÕES POR MEIO DE TESTES ESTATÍSTICOS APROPRIADOS?

A Tabela 5 apresenta uma sintetize dos trabalhos em que foi observado algum tipo de avaliação experimental.

Tabela 5: Avaliação Experimental dos Trabalhos

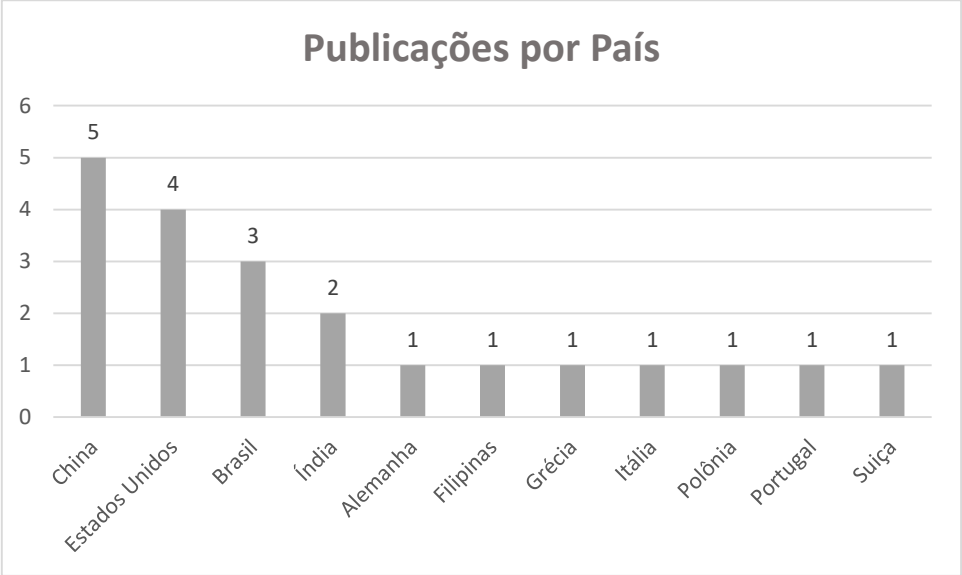
Questão	***Número de Artigos***	***Trabalhos***
O estudo possui alguma avaliação experimental?	06	Bock et al. (2018); Costa et al. (2015); Costa et al. (2016); Ławrynowicz & Potoniec (2014); Santos et al. (2017); Sharma, Osei-Bryson & Kasper (2012)
Foi feito o cálculo do tamanho da amostra?	0	-
Foi feito o teste de normalidade?	04	Costa et al. (2016); Costa et al. (2015); Ławrynowicz & Potoniec (2014); Santos et al. (2017);
Foi declarada hipótese formalmente?	06	Bock et al. (2018); Costa et al. (2015); Costa et al. (2016); Ławrynowicz & Potoniec (2014); Santos et al. (2017); Sharma, Osei-Bryson & Kasper (2012)
Foi calculado o intervalo de confiança?	02	Santos et al. (2017); Sharma, Osei-Bryson & Kasper (2012)
Foram declaradas as ameaças à validade?	04	Costa et al. (2015); Costa et al. (2016); Santos et al. (2017); Sharma, Osei-Bryson & Kasper (2012)

Dos 21 trabalhos selecionados, apenas 6 (28,5%) eram validados experimentalmente. Além disso, das metodologias observadas nos trabalhos, não identificamos nenhuma, seja de BI ou de *Data Mining*, que fosse dirigida à experimentação, ou seja, que prevê uma fase experimental na validação dos resultados. Assim, fica evidenciado que a maioria das pesquisas que envolvem aplicações de BI e *Data Mining* não conduzem um processo experimental.

3.5.4 Quais Países Possuem mais Pesquisadores Publicando sobre Esse Tema?

A Figura 5 mostra os trabalhos selecionados por país, entre os quais predominou a China como o país que possui mais pesquisadores sobre esse tema, seguida por Estados Unidos e Brasil.

Figura 5 – Publicações por País



3.5.5 Quais os Anos que Tiveram mais Publicações nessa Área?

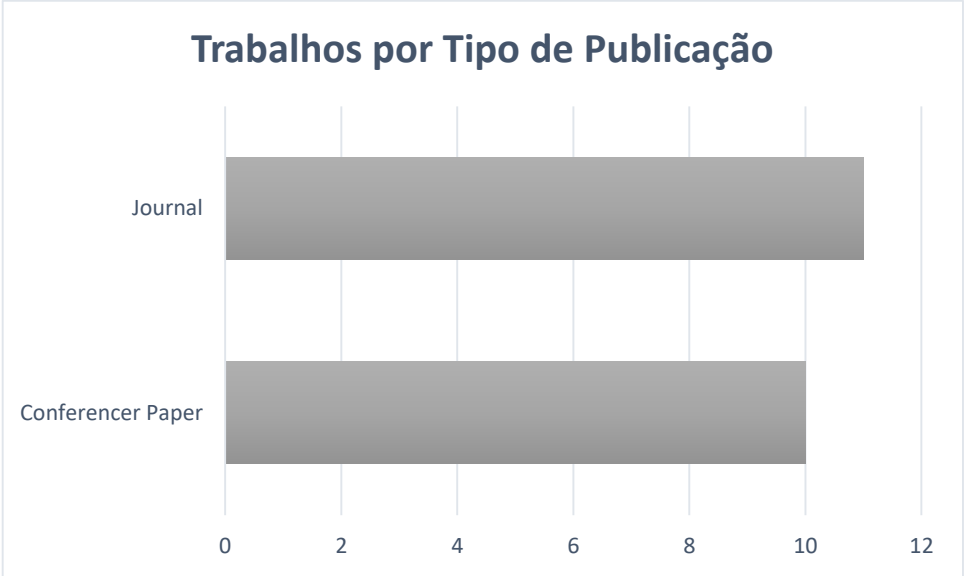
A Figura 6 mostra os artigos selecionados por ano de publicação. No ano de 2019, os artigos foram selecionados até o mês de março (mês em que a primeira etapa da execução da revisão foi encerrada). Pode-se observar que os maiores números de estudos foram publicados nos anos de 2015 e 2017.

Figura 6 – Artigos selecionados por Ano de Publicação

3.5.6 QUAIS OS MEIOS DE PUBLICAÇÕES MAIS POPULARES?

A Figura 7 apresenta o quantitativo dos trabalhos selecionados por tipo de publicação.

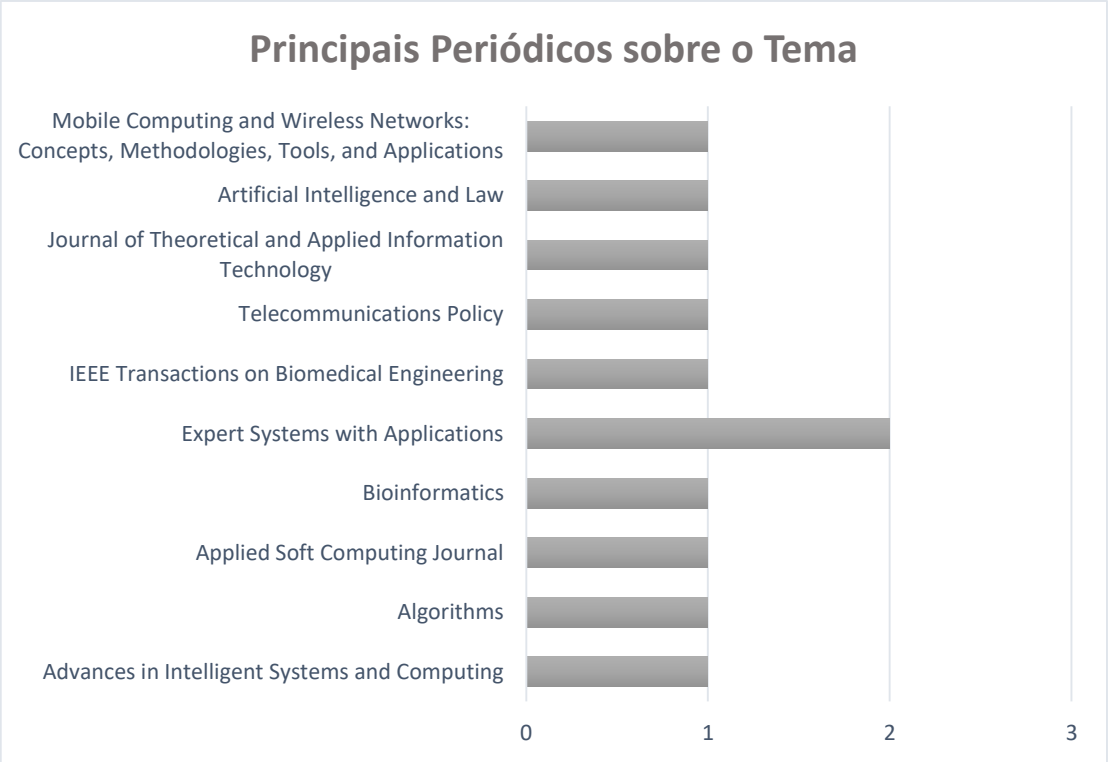
Figura 7 – Trabalhos Selecionados por Tipo de Publicação



3.5.7 QUAIS OS PRINCIPAIS PERIÓDICOS E CONFERÊNCIAS SOBRE O TEMA?

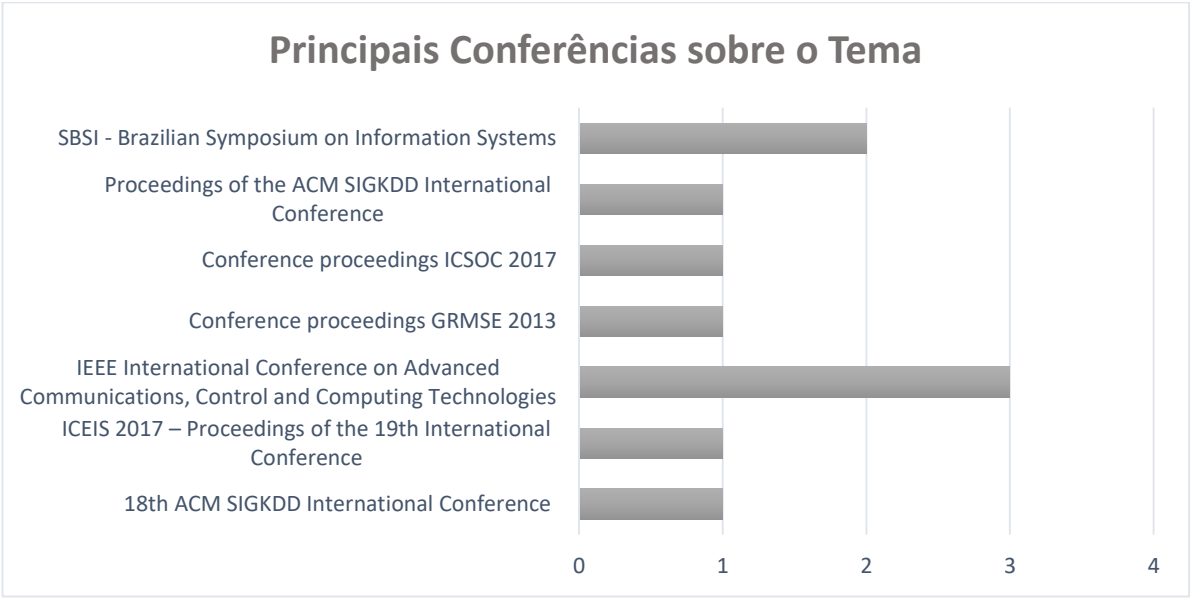
A Figura 8 apresenta os principais periódicos sobre o tema.

Figura 8 – Principais Periódicos sobre o Tema



Na Figura 9, é possível visualizar as principais conferências. Entre as quais a "*IEEE International Conference on Advanced Communications, Control and Computing Technologies*" foi a proeminente, com 3 publicações consecutivas, nos anos 2013, 2014 e 2015.

Figura 9 – Principais Conferências sobre o Tema



3.5.8 QUAIS OS TIPOS DE ESTUDOS EXECUTADOS?

A Figura 10 apresenta os tipos de estudos executados nos artigos primários, sendo que a "Aplicação Prática" foi o tipo de estudo mais utilizado.

Figura 10 – Tipos de Estudos

Para esta classificação, foram utilizadas as definições:

- Experimento Controlado: Forma de estudo experimental na qual o investigador tem controle sobre os principais aspectos do estudo e as variáveis independentes que estão sendo estudadas. Este tipo de estudo é caracterizado pelo controle sistemático das variáveis e do processo, tendo como objetivo confirmar teorias, conhecimento convencional, explorar relacionamentos, avaliar a predição de modelos ou validar medidas. Além disso, envolve a formulação de hipóteses, que precisam ser verificadas em relação aos resultados obtidos (Delamaro, Jino, & Maldonado, 2017; Wohlin et al., 2012).
- Estudo de Caso: Baseia-se na utilização de um ou mais métodos qualitativos ou não segue uma linha rígida de investigação. Consiste geralmente no estudo aprofundado de um único “caso” ou de “casos relacionados”, sendo executado em condições típicas, por exemplo, a partir de alguns projetos típicos representativos (Read, 2003).
- Aplicação Prática: Consideramos esta classificação ao encontrarmos trabalhos que são semelhantes a um estudo de caso, no entanto, com alguma lacuna no método de avaliação. Em outras palavras, aplicações que podem não ter usado dados reais, ou não foram executadas em ambiente real. Além disso, pode não ter existido uma avaliação, minimamente, qualitativa.

- Prova de Conceito: Termo utilizado para denominar um modelo prático que possa provar o conceito (teórico) estabelecido por uma pesquisa ou artigo técnico. Pode ser considerado também uma implementação, em geral resumida ou incompleta, de um método ou de uma ideia, realizada com o propósito de verificar que o conceito ou teoria em questão é suscetível de ser explorado de uma maneira útil (Farias et al., 2019).
- Estudo Exploratório: Caracterizado pela flexibilidade, criatividade e informalidade que esse tipo de estudo permite ao pesquisador na busca em obter um maior conhecimento sobre um determinado tema ou problema de pesquisa. Muitos autores consideram os estudos exploratórios como um estágio preliminar no processo de pesquisa como um todo, servindo para coletar dados e informações (Hall & Rist, 1999).

3.6 AMEAÇAS À VALIDADE

As ameaças à validade podem limitar a habilidade de interpretar e/ou descrever resultados dos dados obtidos. Portanto, não há como desconsiderar as seguintes ameaças encontradas nesse estudo.

Validade de Construção: A *string* de busca e as questões de pesquisa utilizadas podem não cobrir a área de metodologias de BI e *Data Mining* dirigidas à estratégia e avaliadas experimentalmente. Para mitigar essa ameaça, tentou-se elaborar uma *string* mais abrangente possível, quanto aos termos que pudessem ser usados na área, utilizando vários sinônimos. Tais termos foram identificados e refinados, com auxílio dos artigos de controle norteados pelo modelo PICO, utilizando trabalhos que interessavam à pesquisa (intervenção) e falsos positivos, com o objetivo de calibrar a *string* de busca. Além disso, foram consideradas as opiniões de três pesquisadores.

Validade Interna: (Extração de dados): Três pesquisadores foram responsáveis por extrair e classificar os dados de cada publicação. Logo, vieses ou problemas na extração dos dados podem ameaçar a validade da caracterização dos dados. (Viés de Seleção): Inicialmente, os artigos foram incluídos ou excluídos de acordo com julgamento dos próprios pesquisadores. Consequentemente, alguns estudos podem ter sido categorizados incorretamente. Para mitigar estas ameaças, as revisões da seleção e extração foram feitas por todos os pesquisadores envolvidos e as discordâncias encontradas foram resolvidas em uma votação final.

Validade Externa: O uso da língua inglesa pode ter contribuído para a não inclusão de possíveis documentos relevantes em outras línguas.

3.7 CONCLUSÃO

Neste trabalho, foi realizada uma revisão *quasi*-sistemática, visando identificar e caracterizar as metodologias de desenvolvimento de aplicações de *Business Intelligence* e *Data Mining* dirigidas à estratégia e/ou que preveem avaliação Experimental, avaliando e evidenciando os trabalhos mais relevantes na área. Nesse caminho, a revisão compreendeu trabalhos publicados no período de 2009 a 2019. Sendo assim, a busca inicial retornou 841 estudos, dos quais, após serem avaliados, segundo critérios de inclusão e exclusão estabelecidos no protocolo de seleção, foram aceitos apenas 21 trabalhos.

Como resultados, não foram encontrados trabalhos que apresentassem alguma abordagem completa para disciplinar o alinhamento estratégico e a experimentação, prevendo atendimento claro aos objetivos estratégicos e uma fase experimental na validação dos resultados. No entanto, alguns ensaios de partes dessas características puderam ser mapeados, como, por exemplo, a experimentação, encontrada em 28,57% dos trabalhos. Entre os países, a China, os Estados Unidos e o Brasil lideraram o ranking de publicações sobre o tema. Quanto ao meio de publicação, o *Journal* foi a opção mais utilizada para publicação. Além disso, a conferência "*IEEE International Conference on Advanced Communications, Control and Computing Technologies*" e o periódico "*Expert Systems with Applications*", destacaram-se como maiores publicadores.

Assim, acredita-se que esta pesquisa apresenta resultados relevantes à academia e aos empreendedores, fornecendo evidências de que não há um método formal de desenvolvimento experimental de aplicações de BI e *Data Mining*, voltado ao planejamento estratégico de uma organização.

Por fim, este trabalho apresenta-se como uma fonte de consulta aos padrões de métodos existentes para o desenvolvimento de aplicações inteligentes, bem como pode ser replicado e estendido. Como trabalhos futuros, podem ser propostos métodos para criação de aplicações inteligentes alinhadas à estratégia e com validação experimental.

Uma vez apresentada a Revisão *Quasi*-Sistemática, serão explanadas, no próximo capítulo, a proposta e a avaliação de um processo para o desenvolvimento experimental de aplicações de *Data Mining* e *Data Science* alinhadas ao planejamento estratégico da organização.

4.0 UM PROCESSO PARA O DESENVOLVIMENTO EXPERIMENTAL DE APLICAÇÕES DE DATA MINING E DATA SCIENCE ALINHADAS AO PLANEJAMENTO ESTRATÉGICO DA ORGANIZAÇÃO

Este capítulo apresenta parte do artigo intitulado: Proposta e Avaliação de um Processo para o Desenvolvimento Experimental de Aplicações de *Data Mining* e *Data Science* Alinhadas ao Planejamento Estratégico da Organização.

Resumo: Contexto: O fenômeno *Big Data* tem imposto maturidade às empresas na exploração de seus dados, como prerrogativa para obter insights valiosos sobre seus clientes e o poder da análise para orientar a tomada de decisão. Desta forma, uma abordagem geral que descreva como extrair conhecimento para a execução da estratégia empresarial precisa ser estabelecida. **Objetivo:** O objetivo deste trabalho é propor, introduzir e avaliar a implantação de um processo para o desenvolvimento experimental de aplicações de *Data Mining* (DM) e *Data Science*, alinhadas ao planejamento estratégico da organização **Método:** Foi realizado um estudo de caso com o processo proposto em uma instituição de ensino federal. **Resultados:** Os resultados trouxeram evidências de que é possível integrar uma abordagem de alinhamento estratégico, método científico e uma metodologia de desenvolvimento de aplicações de DM. **Conclusão:** Aplicações de *Data Mining* e *Data Science* também possuem os riscos de outros Sistemas de Informação e as adoções de processos dirigidos à estratégia e de método científico são fatores críticos de sucesso. Além disso, foi possível concluir que a aplicação do método científico é facilitada, além de ser uma importante ferramenta para garantia da qualidade de aplicações inteligentes. Por fim, para fomentar e disciplinar o alinhamento estratégico, o processo precisa estar mapeado.

Palavras-chave: *Data Mining*, Mineração de Dados, *Data Science*, Alinhamento Estratégico, Experimentação.

4.1 METODOLOGIA

Inicialmente, foi feita uma revisão *quasi*-sistemática da literatura, publicada em (Cruz, Colaço Júnior & Gois, 2021), com a finalidade de identificar e caracterizar as metodologias de desenvolvimento de aplicações de *Business Intelligence* e *Data Mining* dirigidas à estratégia e que preveem avaliação experimental.

A revisão identificou a ausência de métodos ou processos integrados ao desenvolvimento de aplicações de DM que apresentassem alguma abordagem completa para disciplinar o alinhamento estratégico e a experimentação, prevendo atendimento claro aos

objetivos estratégicos e uma fase experimental na validação dos resultados. Ato contínuo, diante da lacuna identificada, foi desenvolvido um processo para o alinhamento estratégico e para a construção Experimental de aplicações de *Data Mining*.

Por fim, para avaliar o processo proposto, foi planejado e realizado um estudo de caso com a área de inteligência de uma instituição federal de ensino. De acordo com Yin (2015), um estudo de caso é uma investigação empírica de algum fenômeno contemporâneo, em profundidade e em seu contexto de vida real, especialmente quando os limites entre o fenômeno e o contexto não são claramente evidentes. Refere-se a uma análise detalhada de um caso específico, supondo que é possível o conhecimento deste fenômeno a partir do estudo minucioso de um único caso (Costa, et al., 2013).

Na seção 4.5, de forma autocontida, o estudo de caso e sua metodologia são detalhados.

4.2 TRABALHOS RELACIONADOS

Não foram encontrados trabalhos científicos com o mesmo objeto de pesquisa deste artigo, considerando o uso de alguma abordagem para disciplinar o alinhamento estratégico ao desenvolvimento experimental de aplicações de DM, o que aumenta a importância dos resultados aqui apresentados. No entanto, alguns trabalhos mencionam a importância deste alinhamento ou propõem alguma metodologia para o desenvolvimento experimental eficiente deste tipo de aplicação.

Em um *survey* realizado no Brasil (Lima et al., 2017), foi constatado que 67,50% das empresas não utilizam uma metodologia experimentada para o desenvolvimento de BI, o que contribui para que os projetos não obtenham sucesso. Associado a este resultado, um percentual de 72,00% das empresas não utilizam uma metodologia de alinhamento estratégico. A ausência de uma metodologia alinhada à estratégia da empresa evidencia que os gestores podem estar tomando decisões com base em informações não relevantes à instituição ou desalinhadas às estratégias de negócio.

Neste caso, vale ressaltar que apesar da literatura e a prática separarem conceitualmente as áreas de *Data Mining* e BI, há uma forte convergência e integração destas, pois o "I", ou Inteligência do BI, só pode ser concretizado com a aplicação de técnicas de *Data Mining*. Isto indica que a ausência destas metodologias de alinhamento também deve atingir os projetos de *Data Mining*, uma vez que existem projetos de BI sem *Data Mining*, *Data Mining* sem BI e, no melhor caso, um BI completo, que utilizada integração de dados analíticos (BI), estatística e inteligência artificial - *Data Mining* -.

Sharma, Osei-Bryson & Kasper (2012) abordam em seu trabalho as várias limitações identificadas nos modelos de processos existentes de *Data Mining* e propõem resolvê-las por meio da proposta de um novo modelo melhorado, denominado, Modelo Integrado de Descoberta de Conhecimento e Data Mining (IKDDM), que apresenta uma visão integrada do processo KDDM (*Knowledge Discovery and Data Mining* – Descoberta de Conhecimento e Mineração de Dados) e fornece suporte explícito para a execução de cada uma das tarefas descritas no modelo. Também foi avaliada a eficácia e a eficiência oferecidas pelo modelo IKDDM contra o CRISP-DM, um modelo líder no processo de KDDM. Os resultados dos testes estatísticos indicaram que o modelo IKDDM supera o modelo CRISP-DM em termos de eficiência e eficácia. O modelo IKDM também superou o CRISP-DM em termos de qualidade do próprio modelo de processo.

Kohavi et al. (2013) apresentam o *Bing Experimentation System*, sistema que busca orientar o desenvolvimento de produtos e permitir que a organização avalie o ROI dos projetos através da experimentação. O Sistema permite a execução simultânea de mais de 200 experimentos, expondo cerca de 100 milhões de clientes ativos mensais a bilhões de variantes do Bing que incluem implementações de novas ideias e variações das existentes.

Cheng et al. (2009) apresentam uma abordagem baseada em ontologias para aplicações de BI, especificamente em Análise Estatística e *Data Mining*, implementando a abordagem em um Sistema de Gestão do Conhecimento Financeiro. O conhecimento resultante de cada experimento, o qual consiste em sequências de dados, modelo, parâmetros e relatórios, é armazenado, compartilhado, disseminado e, portanto, útil para apoiar a tomada de decisões.

Por fim, Colaço Júnior et al. (2019) apresentaram um processo que mescla a abordagem GQM+*Strategies* com uma metodologia de desenvolvimento ágil de aplicações de *Business Intelligence* proposta pelo autor, visando garantir o alinhamento estratégico. O processo proposto foi avaliado por meio de um estudo de caso, em uma empresa multinacional latino-americana do mercado de varejo, no qual foi evidenciado que é possível integrar a abordagem de alinhamento estratégico adotada com uma metodologia de desenvolvimento de aplicações de BI. Com as boas evidências iniciais, os pesquisadores deste artigo evoluíram o processo e previram o uso de experimentação para validação dos modelos inteligentes que podem ser criados com técnicas de Data Mining e IA, a partir de uma aplicação de BI já pronta ou não.

4.3 BASE CONCEITUAL

Nesta seção, são apresentados alguns conceitos necessários para o entendimento deste trabalho.

4.3.1 GQM+STRATEGIES

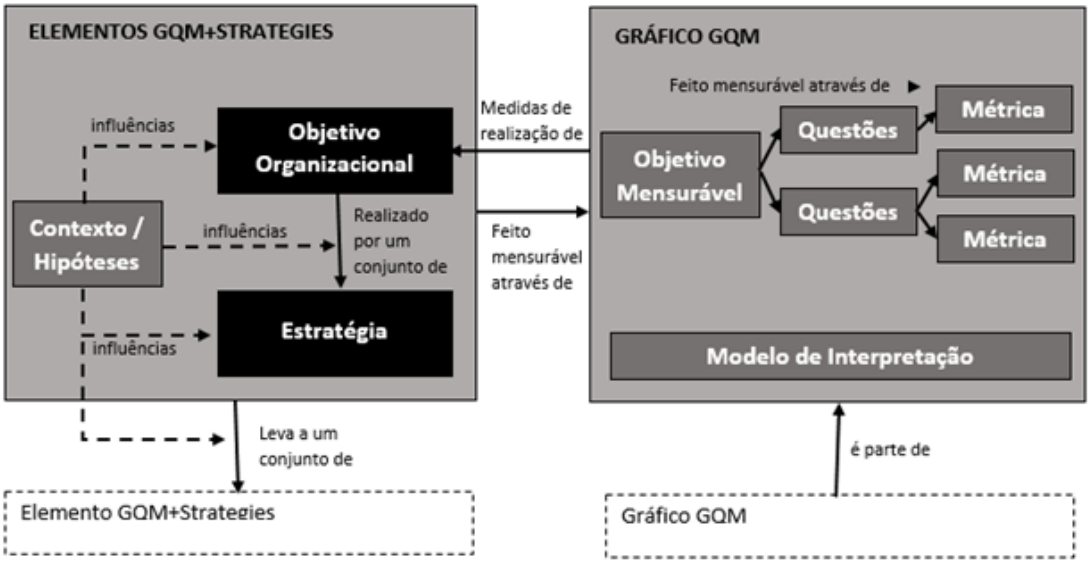
O GQM + Strategies ampliou o modelo GQM, este último conceituado como uma abordagem sistemática que integra os objetivos de negócio, adaptando-os aos modelos de processos de software, produtos e perspectivas de interesse de qualidade, com base nas necessidades específicas do projeto (Basili et al., 2007).

Segundo Basili et. al. 2010, GQM+Strategies é uma abordagem para alinhar as organizações através de medição. Isto permite a uma organização integrar o alinhamento estratégico e suas metas de forma consistente em diferentes unidades, para tomar decisões com base em métricas identificadas, comunicar objetivos e estratégias da organização, bem como monitorar a realização do objetivo e o sucesso/fracasso das estratégias definidas.

O principal resultado dessa abordagem é um programa de medição estratégica que permite decisões baseadas em dados (Basili et al., 2010). Para poder vincular uma metodologia de desenvolvimento de software ao alinhamento estratégico, o GQM+*Strategies* possui dois componentes principais (Basili et al., 2010):

- Grade (*Grid*) - Documenta os objetivos estratégicos nos quais a organização deseja focar, suas justificativas de vinculação das metas a unidades organizacionais diferentes e um método de medição para avaliar e interpretar os dados a serem medidos para tomada de decisão (Figura 11);
- Processo (*Process*) - Define como criar o modelo, a implementação de suas estratégias, coleta e análise dos dados, além de como iniciar as ações de melhoria dentro do processo.

Figura 11 – GQM+*Strategies* Grid (Basili et al., 2010)



O processo GQM+*Strategies* é composto por seis fases repetitivas, mais uma fase de inicialização. As seis fases são organizadas como um ciclo de melhoria contínua e baseiam-se no *Quality Improvement Paradigm* (QIP) proposto por Victor Basili et al. (2014). Essas fases são agrupadas em 3 macro etapas, cada uma contendo 2 fases específicas. As 3 macro etapas são:

- Desenvolvimento (*Develop*): desenvolve o modelo hierárquico (grid) que alinha as metas, estratégias e dados de medição;
- Implementação (*Implement*): executa as estratégias e medições definidas no processo anterior e, assim, verifica a consecução das metas e a eficácia das estratégias;
- Aprendizagem (*Learn*): envolve o conhecimento a partir do que foi feito, por meio da análise dos resultados, para melhorar o processo de geração de novas metas e estratégias.

Tais elementos auxiliam o levantamento de dados a ser realizado na implantação da metodologia, do fluxo e da relação entre eles. São definidos pelas 6 fases exibidas na Figura 12: Inicialização; Caracterização de Ambiente; Definição de Metas, Estratégias e Medições; Planejamento da Implementação do Modelo; Execução do Planejamento; Análise dos resultados e Pacotes de Melhorias (Basili et al., 2014).

Figura 12 – Processos básicos do GQM+Strategies (Basili et al., 2010)

4.4 PROCESSO PARA O DESENVOLVIMENTO EXPERIMENTAL DE APLICAÇÕES DE DATA MINING E DATA SCIENCE ALINHADO AO PLANEJAMENTO ESTRATÉGICO DA ORGANIZAÇÃO

Esta seção apresenta um processo de desenvolvimento de aplicações de Data Mining dirigidas à estratégia e avaliadas experimentalmente, como parte integrante de um projeto de BI. O principal objetivo deste processo é tratar uma lacuna identificada a partir da revisão sobre o uso de metodologias de desenvolvimento de aplicações de Data Mining voltadas ao planejamento estratégico e à experimentação.

Antes da apresentação do novo processo, será revisado e introduzido, a seguir, um outro processo que também visa o alinhamento estratégico, mas, de aplicações de BI, sem considerar, explicitamente, a Mineração de Dados.

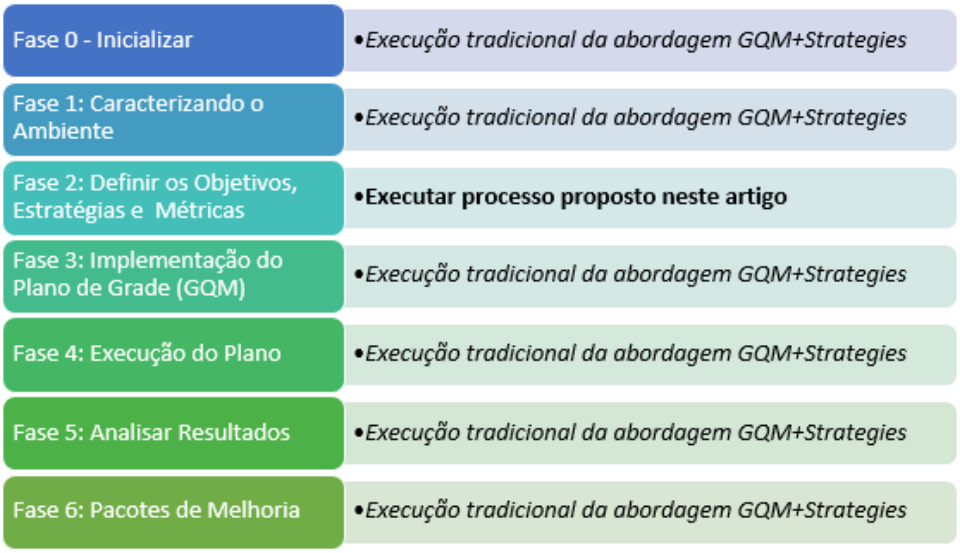
4.4.1 GQM+STRATEGIES E UMA METODOLOGIA ÁGIL, NA ELICITAÇÃO DE REQUISITOS PARA PROJETOS DE BI

Colaço Jr et al. (2019), diante da lacuna identificada por Lima el al., 2017, sobre a inexistência de métodos que disciplinassem o desenvolvimento de aplicações de BI alinhadas ao planejamento estratégico da organização, propuseram uma abordagem que mescla a metodologia GQM+*Strategies* com uma metodologia ágil de desenvolvimento de aplicações de

Business Intelligence proposta pelo autor, visando garantir o alinhamento estratégico e agilidade na entrega das soluções.

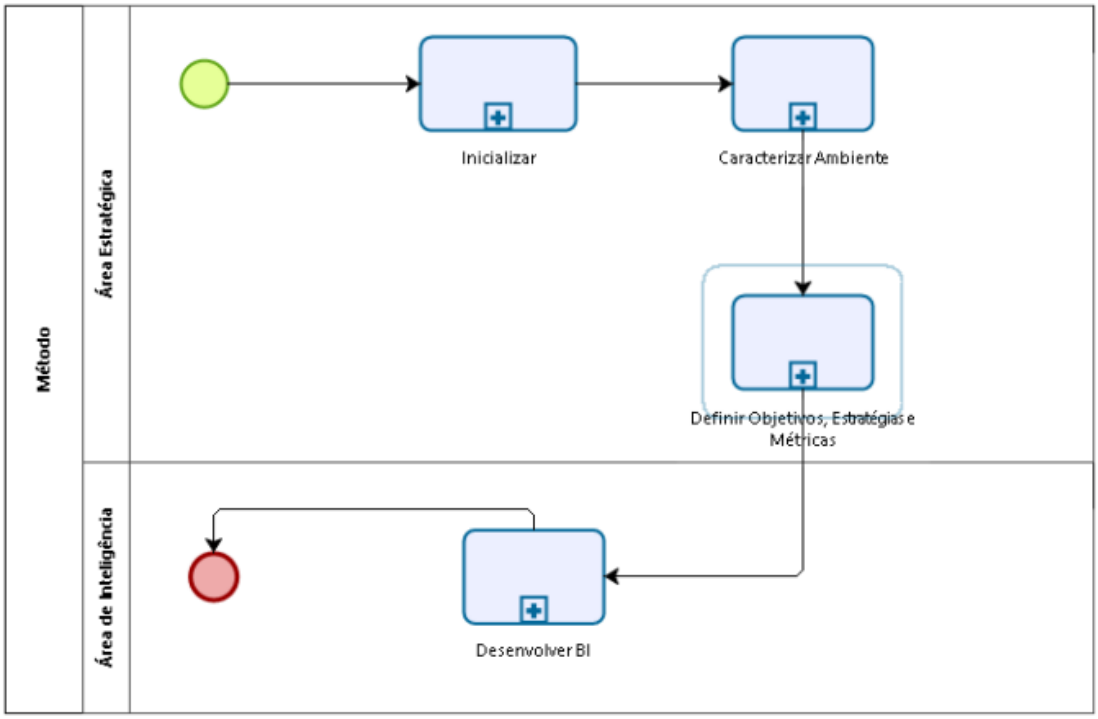
Desta forma, o processo proposto por Colaço Jr et al. (2019) também é uma adaptação da abordagem GQM+*Strategies*, acrescentando uma nova metodologia de concepção de aplicações de BI, sem a extensão para o detalhamento de funcionalidades que usam Mineração de Dados e Inteligência Artificial (vide Figura 13).

Figura 13 – Proposta de desenvolvimento de BI adaptada ao GQM+Strategies (COLAÇO JR et al., 2019)



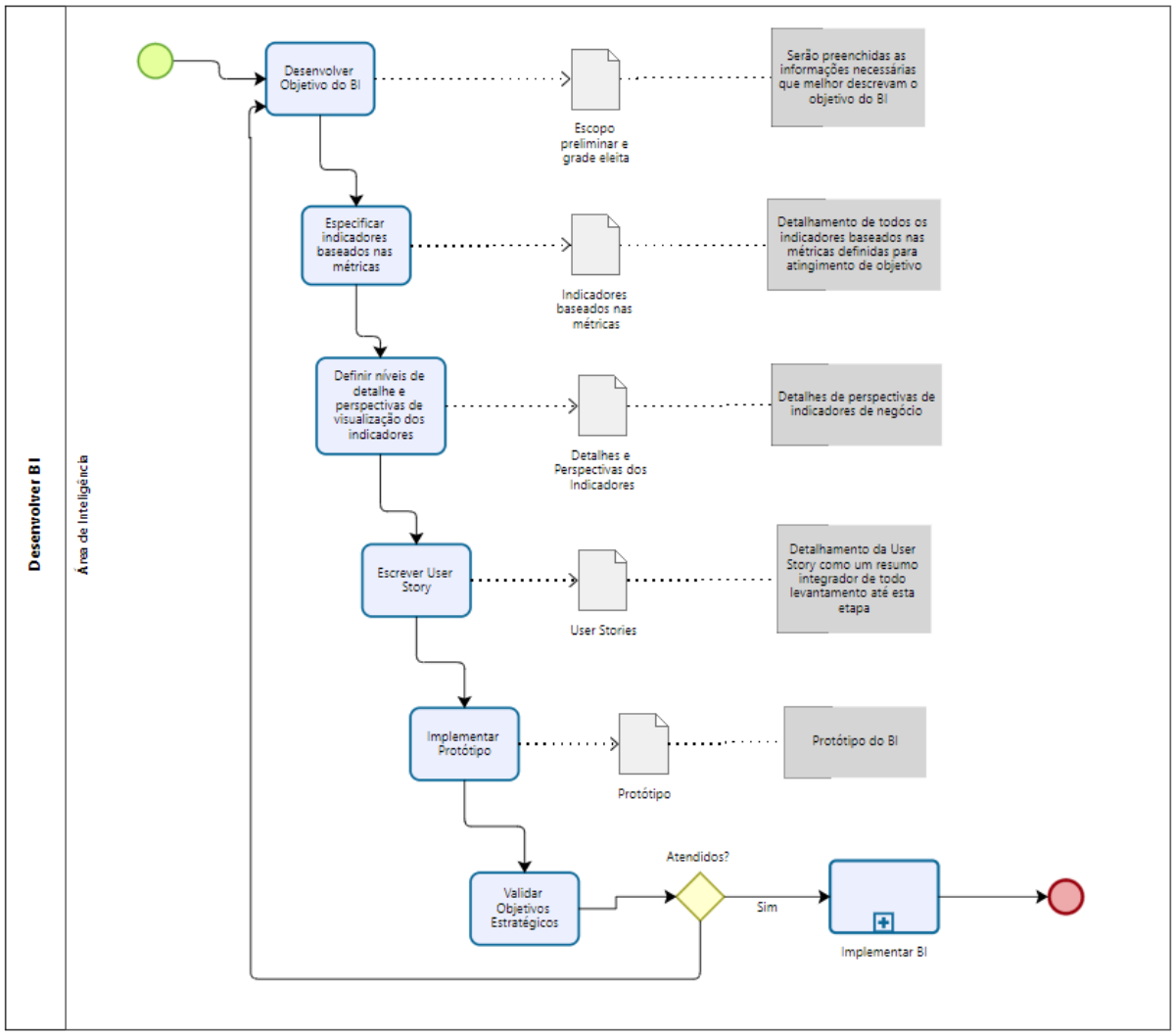
A Figura 14 exibe a proposta macro como um processo. Nas Fases 0 e 1, a área Estratégica seguiria o modelo tradicional GQM+*Strategies*. Na fase 2, foi adaptado o GQM+*Strategies*, criando uma nova proposta para o desenvolvimento de aplicações de BI.

Figura 14 – Macro Processo Proposto por Colaço Jr. Et al. (2019)



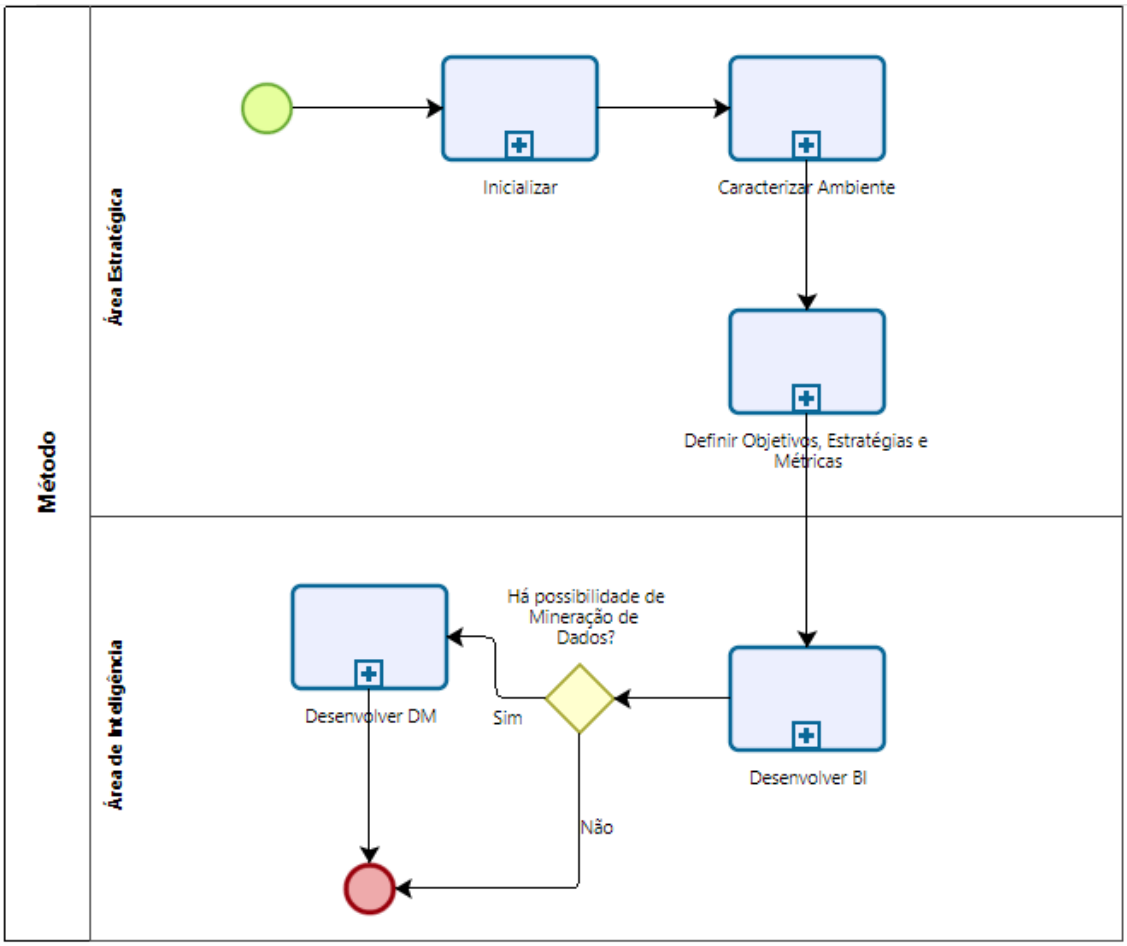
Após execução das etapas preliminares do GQM+Strategies, a Área de Inteligência é envolvida na fase 2, para executar a etapa: Desenvolver o BI. Esta etapa é dividida em 6 (seis) atividades (Figura 15): 1.0 - Definir objetivo do BI; 2.0 - Especificar indicadores baseados nas métricas; 3.0 - Definir níveis de detalhe e perspectivas de visualização dos indicadores; 4.0 - Escrever UserStories; 5.0 - Implementar Protótipo e 6.0 - Validar Objetivos Estratégicos.

Figura 15 – Atividades do Processo Desenvolver BI (Colaço Jr et al., 2019)



Neste sentido, o objetivo deste trabalho é estender a metodologia de BI dirigida à estratégia proposta por Colaço Jr et al. (2019), para contemplar aplicações de *Data Mining* avaliadas experimentalmente. A Figura 16 exibe a proposta macro como um processo unificado e o processo em si será discutido na próxima seção.

Figura 16 – Macro Processo Proposto

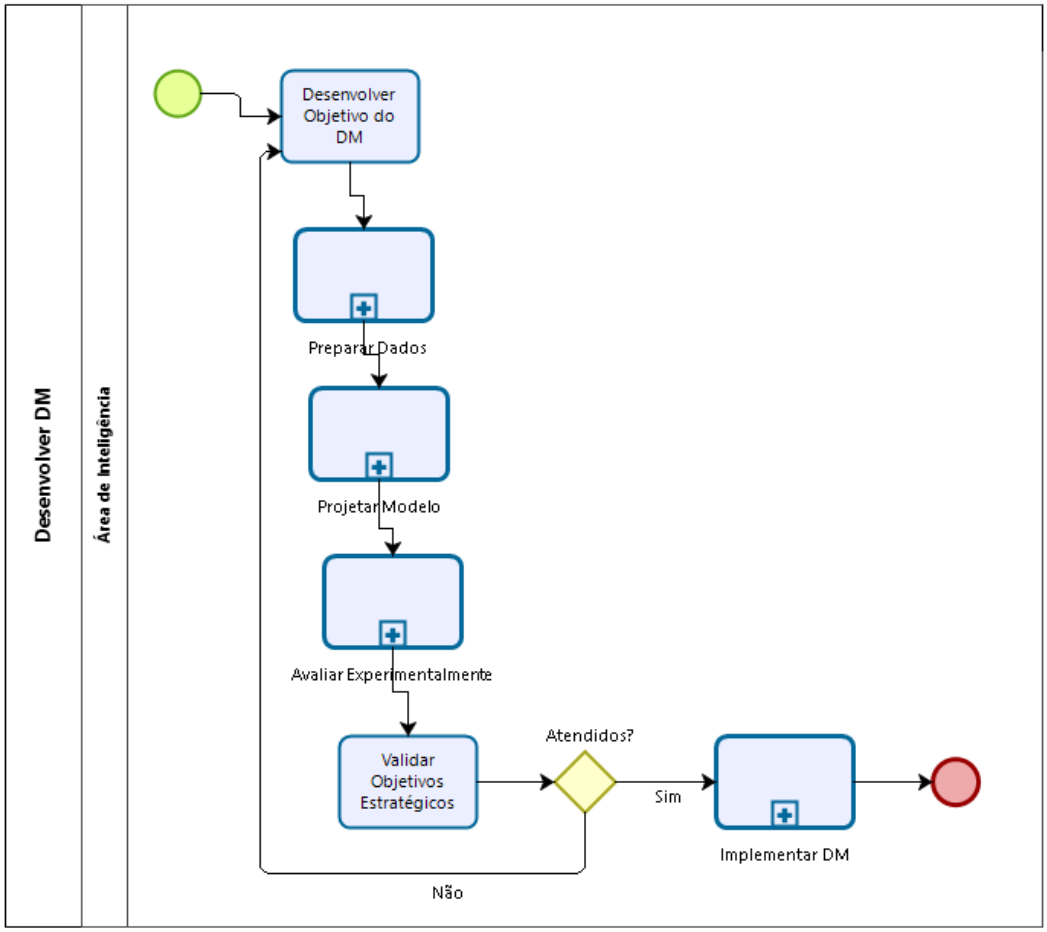


4.4.2 DESENVOLVIMENTO EXPERIMENTAL DE APLICAÇÕES DE DATA MINING ALINHADO AO PLANEJAMENTO ESTRATÉGICO DA ORGANIZAÇÃO

Diante das lacunas identificadas e dos resultados obtidos por Colaço Jr et al. (2019), decidiu-se desenvolver um processo de desenvolvimento de aplicações *Data Mining* dirigidas à estratégia e avaliadas experimentalmente, como parte integrante de um projeto de BI.

O processo proposto é composto de 6 atividades (Figura 17): (i) Desenvolver Objetivo do Processo de DM, (ii) Preparar Dados, (iii) Projetar Modelo, (iv) Avaliar Experimentalmente, (v) Validar Objetivos Estratégicos e (vi) Implementar DM.

Figura 17 – Atividade do Processo Desenvolver DM



A seguir, serão descritas todas as atividades que compõem cada etapa do processo, bem como serão detalhados Objetivo, Entrada, Subatividades e Resultados (Saídas) de cada atividade.

4.4.2.1 DESENVOLVER OBJETIVO DO PROCESSO DE DM

Esta atividade tem como foco compreender os objetivos e requisitos do negócio, convertendo este conhecimento num problema de mineração de dados. O objetivo, entradas, subatividades e resultados são mostrados na Tabela 6, que segue abaixo.

Tabela 6 - Descritivo da atividade: Definir Objetivo da Mineração de Dados

Atividade: Definir Objetivo da Mineração	
Objetivo	Definir objetivo geral da mineração, vinculado ao planejamento estratégico.
Entradas	• Grade do *GQM+Strategies* e Objetivo prioritário do cliente.
Subatividades	1. Identificar os objetivos organizacionais; 2. Selecionar o objetivo estratégico na grade; 3. Selecionar questões e métricas referentes ao objetivo estratégico; 4. Escrever objetivo geral do incremento, baseado nas questões GQM; 5. Revisar e Ajustar.

Resultados (Saídas)	• Escopo preliminar e grade eleita.

Para definir o objetivo da mineração para uma organização, a motivação básica precisa ser determinada e a lógica que leva à definição do objetivo precisa ser descrita.

4.4.2.1.1 ENTRADAS

A Grade do GQM+*Strategies* é um elemento da abordagem GQM+*Strategies* já descrita na base conceitual deste trabalho. Neste elemento, são documentados os objetivos estratégicos que a organização deseja focar, suas justificativas de vinculação das metas, bem como um método de medição para avaliar e interpretar os dados a serem medidos para tomada de decisão.

Nesta fase, são identificados os objetivos e requisitos do projeto sob a perspectiva do negócio, para então converter estas metas em um projeto de mineração de dados. Vale ressaltar que o cliente pode usar outra metodologia e elencar os objetivos estratégicos e prioridades de outra forma.

4.4.2.1.2 SUBATIVIDADES

É todo esforço necessário para execução da atividade macro denominada: *Definir Objetivo da Mineração de Dados*. Para essa atividade, foram definidas 5 subatividades:

1. ***Identificar os objetivos organizacionais:*** Para que seja possível realizar essa subatividade, é necessário ter disponível a Grade do *GQM+Strategies* e o Objetivo prioritário do cliente. Neles ou em artefatos de outra metodologia, devem estar contidos todos os objetivos estratégicos definidos pela organização. Essa lista é necessária para que seja possível identificar qual objetivo estratégico será selecionado.
2. ***Selecionar o objetivo estratégico na grade:*** Nesta subatividade, são consideradas metas promissoras em relação à viabilidade, benefício e custo. Recomenda-se a concentração em metas que tenham o maior impacto de sucesso para o negócio. O processo de seleção dessas metas é um processo muito interativo, exigindo, pois, a participação de várias unidades da organização.
3. ***Selecionar questões e métricas referentes ao objetivo estratégico:*** Uma vez que já foram definidas e identificadas as questões e métricas em seu nível estratégico, torna-se necessário selecionar agora as questões e

métricas que irão auxiliar a definir o escopo, limitá-lo e constituir um raciocínio para os objetivos e suas estratégias, selecionados da grade GQM + *Strategies*. Por exemplo, para o objetivo de um negócio de nível superior, as questões e métricas geralmente se referem às restrições e oportunidades externas e estarão relacionadas à visão e à missão da organização. As restrições e oportunidades externas incluem aspectos como os produtos competitivos, estratégia comercial com os fornecedores e as tendências do mercado. As restrições e oportunidades internas incluem aspectos como nível de competência da equipe, satisfação de cliente interno, avanços tecnológicos e infraestrutura existente.

4. ***Escrever objetivo geral do incremento, baseado nas questões GQM:*** É descrever os objetivos que se deseja alcançar com a mineração de dados, indicando os novos dados a serem visualizados, com os seus padrões e/ou predições. Vale ressaltar que todas as etapas anteriores, idealmente, já foram definidas para um Data Mart de um projeto de BI, ou seja, neste caso, o que ocorrerá para a Mineração de Dados é o acréscimo de novas funcionalidades que poderão descrever os dados de novas formas, com descoberta de padrões, e/ou que poderão executar predições úteis à estratégia. User Stories do BI poderão ser incrementadas ou novas poderão ser criadas, considerando inclusive o uso de mídias dinâmicas (multimídia) associadas ao projeto, as quais podem conter filmagens de entrevistas e relatos dos clientes sobre as funcionalidades desejadas (Colaço Júnior et al., 2017) (Santos et al., 2020). Desta forma, nesta subatividade, **o objeto de entrega** é o objetivo geral da mineração de dados, bem como as predições e novas formas de visualização que estarão disponíveis, baseadas nas questões GQM, **com propósito de** descrever de forma geral os *insights* visuais e preditivos do incremento atual do projeto de mineração, **do ponto de vista** do Cliente Específico ou da Área Específica, **no contexto** da organização, influenciado pelo objetivo selecionado e pelo incremento atual do projeto de mineração.
5. ***Revisar e Ajustar:*** Depois de concluídas todas as subatividades anteriores, recomenda-se analisar e discutir o escopo preliminar e a grade

eleita em uma reunião grupal. Recomenda-se também que todas as pessoas de áreas da organização que são afetadas pelo objetivo definido façam parte dessa reunião. Nela, a área de inteligência explica todo o caminho que levou à definição do objetivo escolhido. Os participantes da reunião devem então verificar se os vínculos entre os objetivos e as estratégias são lógicos e pertinentes à realidade. Qualquer questão levantada em reunião é imediatamente discutida e a resolução de cada uma pode ser: Planejada, Abordada de imediato ou descartada em sessão.

4.4.2.1.3 RESULTADOS

Como saída esperada do subprocesso **Definir Objetivo da Mineração de Dados**, temos o "*Escopo preliminar e Grade eleita*", descritos nesta seção. Na tabela 7, disponibilizamos um modelo a ser utilizado.

Tabela 7 - Modelo de Saída do subprocesso: Definir Objetivo da Mineração de Dados

OBJETIVO				
<Descrição do objetivo estratégico que o projeto busca atender. Corresponde a um estado antecipado no futuro que uma organização deseja alcançar. Responda à pergunta: "O que deve ser alcançado?">				
DO PONTO DE VISTA DO NEGÓCIO				
OBJETO	**PROPÓSITO**	**FOCO DE QUALIDADE**	**PONTO DE VISTA**	**CONTEXTO**
< Objeto de Análise >	*< Finalidade do Projeto >*	*< Quais são as métricas possíveis para medir o objetivo de interesse de acordo com os membros do projeto? >*	*<Stakeholders, pessoas que influenciarão a saída da análise>*	*<Público-Alvo>*
DO PONTO DE VISTA DA MINERAÇÃO				
OBJETO	**PROPÓSITO**	**FOCO DE QUALIDADE**	**PONTO DE VISTA**	**CONTEXTO**
< Objeto de Análise >	*< Finalidade da mineração >*	*<Previsibilidade que se deseja alcançar, eficácia, eficiência, etc. >*	*<Stakeholders, pessoas que influenciarão a saída da análise>*	*<Contexto e suposições>*
FOCO DE QUALIDADE (QUESTÕES E MÉTRICAS)			**FATORES DE VARIAÇÃO**	

< Quais são as métricas possíveis para medir o objeto de interesse, de acordo com os membros do projeto?>	*< Quais fatores de contexto um membro do projeto espera influenciar as métricas? Isso fornece informações sobre quais outras informações são importantes para a compreensão das hipóteses da linha de base. >*
HIPÓTESES DE LINHA DE BASE	**IMPACTO NAS HIPÓTESES DE LINHA DE BASE**
< Qual é o conhecimento atual dos membros do projeto em relação a essas métricas? Isso pode estar disponível a partir de dados reais de projetos anteriores ou pode representar alguma forma de opinião de um especialista, ou seja, suposições do que pode ser verdade. >	*< Como esses fatores de variação podem influenciar as medições reais? Que tipo de dependência entre as métricas e os fatores de influência são assumidos? Isso fornece informações sobre quais outros dados são necessários para interpretar o modelo e as métricas. >*
INTERPRETAÇÃO DO MODELO	
< Interpretação dos objetivos descritos >	
OBJETIVO GERAL COM ESCOPO PRELIMINAR DE VISUALIZAÇÃO	
< Descrição do objetivo geral com a delimitação do escopo e resultado final esperado. Para formalização do objetivo, recomenda-se o uso do modelo GQM que segue abaixo. Recomenda-se também adicionar um esboço do resultado esperado, de forma a permitir um fácil entendimento do problema. Para isso, um protótipo inicial das saídas do processo de mineração e User Stories podem ser usados, conforme o exemplo mostrado no estudo de caso.> ***ANALISAR*** *<Objeto de Estudo>,* ***COM A FINALIDADE DE*** *<Objetivo>,* ***COM RESPEITO À(O)*** *<Enfoque>,* ***DO PONTO DE VISTA DE(A)*** *<Stakeholders>,* ***NO CONTEXTO DE(A)*** *<Contexto>.*	

4.4.2.2 PREPARAR DADOS

Esta atividade tem como foco preparar o conjunto de dados (*DataSet*) que será usado na mineração. A maioria dos dados usados em um processo de mineração é originalmente coletada e preservada para outros fins, necessitando de algum refinamento, integração, limpeza ou transformação, antes de estar pronta para o treinamento de um modelo de conhecimento. O objetivo, entradas, subatividades e resultados desta atividade são mostrados na Tabela 8, que segue abaixo.

Tabela 8 - Descritivo da atividade: Preparar Dados

Atividade: Preparar Dados

Objetivo	Construir o(s) DataSet(s)
Entradas	• Escopo preliminar e grade eleita.
Subatividades	1. Pré-Selecionar os Dados; 2. Supervisionar a Base; 3. Balancear a Base; 4. Normalizar a Base.
Resultados (Saídas)	• DataSet otimizado para o processo de mineração.

4.4.2.2.1 ENTRADAS

A entrada desse subprocesso é o documento descrito no item 4.4.2.1.3.

4.4.2.2.2 SUBATIVIDADES

É todo esforço necessário para execução da atividade macro denominada: *Preparar Dados*. Para essa atividade, foram definidas 4 subatividades:

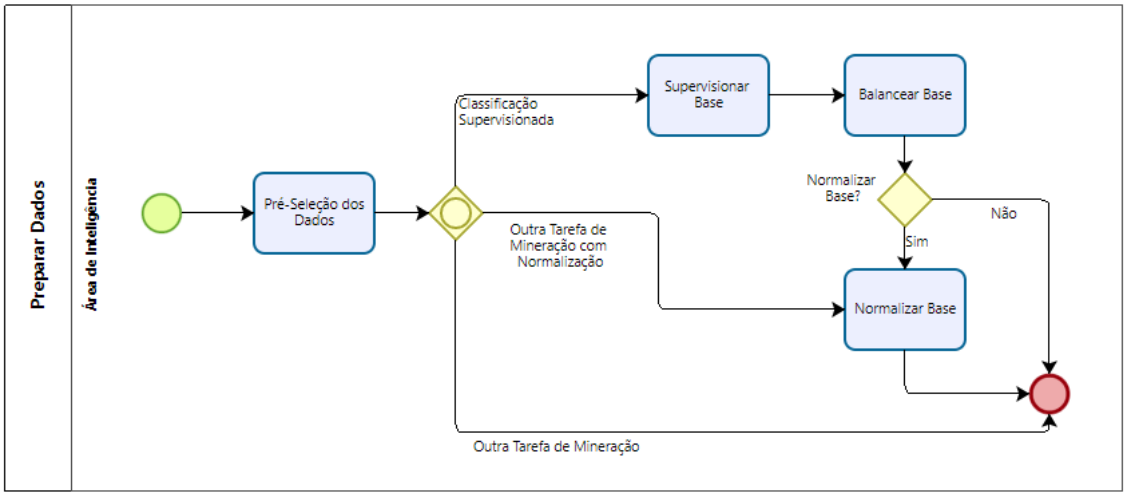
1. ***Pré-Selecionar os Dados:*** Deve-se decidir sobre os dados a serem usados para análise. Os critérios incluem relevância para os objetivos da mineração de dados, qualidade e restrições técnicas, tais como limites no volume de dados ou tipos de dados.
2. ***Supervisionar a Base:*** Em uma tarefa de classificação supervisionada, o supervisionamento é feito quando, a partir de um conjunto de dados rotulados previamente definido, deseja-se encontrar uma função que seja capaz de predizer rótulos desconhecidos.
3. ***Balancear a Base:*** Quando o conjunto de dados apresenta uma desigualdade entre as amostras de suas diferentes classes, faz-se necessária a aplicação de técnicas de balanceamento, de forma a se ter um conjunto de dados homogêneo. Se a métrica acurácia será utilizada, o balanceamento é condição *sine qua non*.
4. Dentre as principais técnicas de balanceamento, estão os métodos baseados em amostragem (*sampling*, ou *resampling*). Estes consistem na modificação da estrutura do conjunto de dados desbalanceado, de maneira a deixá-lo com quantidades equivalentes de amostras para as classes presentes, seja por meio da remoção (*undersampling*) ou adição (*oversampling*) de novas amostras (Maione, 2020).

5. ***Normalizar Base:*** A normalização de dados consiste em colocar atributos em uma mesma faixa de valores. É uma operação que ajusta a escala de valores de cada atributo de forma que os valores fiquem em pequenos intervalos, tais como de -1 a 1, ou de 0 a 1. Tal ajuste faz-se necessário para evitar que alguns atributos, por apresentarem uma escala de valores maior que outros, influenciem de forma tendenciosa em determinados métodos de padrão de dados.

   As variáveis podem ser normalizadas segundo a amplitude ou segundo a distribuição. Existem alguns métodos de normalização de dados: normalização linear (ou Max-min Equalizado), normalização por desvio padrão (ou Z-score), normalização por escala decimal, normalização pela soma dos elementos e normalização pelo valor máximo dos elementos (Mín/Max) (Goldschmidt & Passos, 2005).

Na imagem abaixo, podemos observar como essas subatividades estão dispostas no subprocesso ou na atividade macro *Preparar Dados*.

Figura 18 – Atividades do Processo Preparar Dados



4.4.2.2.3 RESULTADOS

Como saída esperada do subprocesso **Preparar Dados**, temos o *DataSet* otimizado que será usado na mineração. Na tabela 9, disponibilizamos um modelo a ser utilizado. Na primeira coluna da tabela, temos o atributo, na segunda coluna, a sua unidade de medida, na terceira coluna, o seu domínio, que representa o conjunto de valores possíveis do atributo, e, por fim, uma breve descrição do atributo.

Tabela 9 - Modelo de Saída do subprocesso: Preparar Dados

Atributo	Unidade	Domínio Final	Descrição
Atributo 1	Unidade do Atributo 1	Domínio Atributo 1	Descrição Atributo 1
Atributo 2	Unidade do Atributo 2	Domínio Atributo 2	Descrição Atributo 2
Atributo *n*	Unidade do Atributo *n*	Domínio Atributo *n*	Descrição Atributo *n*

4.4.2.3 PROJETAR MODELO

Aqui são selecionadas e aplicadas as técnicas de Data Mining mais apropriadas, com base nos objetivos identificados na primeira fase. O objetivo, entradas, subatividades e resultados desta atividade são mostrados na Tabela 10, que segue abaixo.

Tabela 10 - Descritivo da atividade: Projetar Modelo

Atividade: Projetar Modelo	
Objetivo	Construir modelo de Mineração de Dados
Entradas	• DataSet Otimizado.
Subatividades	1. Selecionar Atributos; 2. Selecionar Algoritmos; 3. Transformar Dados; 4. Definir Parâmetros.
Resultados (Saídas)	• Algoritmos e Atributos Selecionados, Parametrização e Transformação de Dados Necessária.

4.4.2.3.1 ENTRADAS

A entrada desse subprocesso é o documento descrito no item 4.4.2.2.3.

4.4.2.3.2 SUBATIVIDADES

É todo esforço necessário para execução da atividade macro denominada: *Projetar Modelo*. Para essa atividade, foram definidas 4 subatividades:

1. ***Selecionar Atributos:*** O objetivo da seleção de atributos é a eliminação de atributos redundantes e não informativos, bem como a criação de novos atributos. A eliminação desses atributos pode trazer benefícios, tais como facilitar o entendimento e a visualização dos dados, bem como reduzir o custo computacional do algoritmo aplicado.

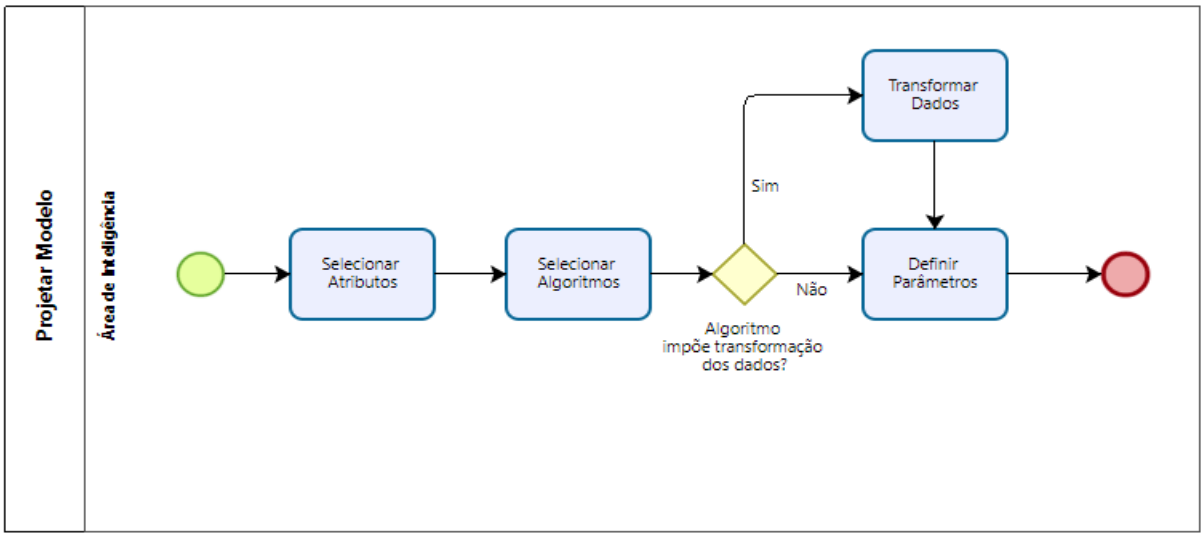
   Uma busca exaustiva pelo melhor subconjunto de atributos possíveis é normalmente inviável sob o ponto de vista computacional, devido à quantidade de subconjuntos possíveis. Assim, o problema de selecionar atributos pode ser abordado via métodos de busca heurísticas. Alguns dos

principais algoritmos utilizados são: *Random forest* (Ma & Fan, 2017), *Greedy hill Climbing*, *Best first* e Algoritmos Genéticos (Covões et al., 2010).

2. ***Selecionar Algoritmos:*** Consiste na seleção dos algoritmos que vão constituir o modelo de mineração de dados. Devem ser levados em consideração os tipos de variáveis envolvidas, as técnicas disponíveis nas ferramentas que serão utilizadas e os objetivos de negócio colhidos na primeira etapa deste processo.
3. ***Transformar Dados:*** O principal objetivo desta fase é transformar a representação dos dados a fim de superar quaisquer limitações existentes nos algoritmos que serão empregados para a extração de padrões. De forma geral, a decisão de quais transformações são necessárias depende do algoritmo que será utilizado na fase de mineração de dados. Determinadas ferramentas podem ser aplicadas apenas a conjuntos de dados com atributos nominais, enquanto outros algoritmos conseguem inferir e descobrir padrões apenas sobre variáveis numéricas, por exemplo.
4. ***Definir Parâmetros:*** Em qualquer algoritmo de mineração, geralmente há um grande número de parâmetros que podem ser ajustados. Deve-se listar os parâmetros e seus valores escolhidos, junto com a justificativa para a escolha das configurações dos parâmetros.

Na imagem abaixo, podemos observar o fluxo dessas atividades na atividade macro *Projetar Modelo*.

Figura 19 – Atividades do Processo Projetar Modelo



4.4.2.3.3 RESULTADOS

Como saída esperada do subprocesso **Projetar Modelo**, temos todos os algoritmos candidatos a compor o modelo de mineração, os quais serão avaliados e um ou apenas alguns serão eleitos para uso. Além disso, os atributos selecionados, a parametrização e a transformação de dados que se fez necessária. Na tabela 11, disponibilizamos um padrão a ser utilizado na documentação desse modelo. Na primeira coluna da tabela, temos os algoritmos selecionados para mineração, na segunda coluna, os atributos do *DataSet* usados pelo algoritmo, e, por fim, os valores dos parâmetros com os quais o algoritmo foi executado.

Tabela 11 - Modelo de Saída do subprocesso: Projetar Modelo

Algoritmo	**Atributos**	**Parâmetros**	**Transformação**
Algoritmo 1	Atributo 1, Atributo 2, Atributo n	Par 1: Valor, Par 2: Valor, Par n: Valor	Se houve, preencher com atributos transformados e fórmula aplicada.
Algoritmo 2	Atributo 1, Atributo 2, Atributo n	Par 1: Valor, Par 2: Valor, Par n: Valor	
Algoritmo n	Atributo 1, Atributo 2, Atributo n	Par 1: Valor, Par 2: Valor, Par n: Valor	

4.4.2.4 AVALIAR EXPERIMENTALMENTE

Esta etapa avalia o grau em que os algoritmos atendem aos objetivos do negócio e busca determinar se há algum motivo de negócio para o qual algum algoritmo é deficiente. Para isso, sugere-se o uso de uma abordagem experimental, com a qual é aplicado o método científico clássico, para validar se houve o cumprimento de todos os objetivos determinados na primeira etapa do processo. O objetivo, entradas, subatividades e resultados desta atividade são mostrados na Tabela 12, que segue abaixo.

Tabela 12 - Descritivo da atividade: Avaliar Experimentalmente

Atividade: Avaliar Experimentalmente	
Objetivo	Avaliar os algoritmos candidatos a compor o modelo de Mineração de Dados a ser construído
Entradas	• *DataSet* com atributos selecionados; • Algoritmos Selecionados e Parametrizados; • Transformações.
Subatividades	1. Definir Objetivo do Experimento; 2. Planejar Experimento; 3. Operar Experimento; 4. Analisar e Interpretar os Dados; 5. Descrever Ameaças à Validade.
Resultados (Saídas)	• Resultados e Detalhamento do Processo Experimental; • Modelo de Mineração de Dados Selecionado (com um ou mais algoritmos eleitos).

4.4.2.4.1 ENTRADAS

A entrada desse subprocesso é o documento descrito no item 4.4.2.2.3 e 4.4.2.3.3.

4.4.2.4.2 SUBATIVIDADES

É todo esforço necessário para execução da atividade macro denominada: *Avaliar Experimentalmente*. Para essa atividade, foram definidas 5 subatividades:

1. ***Definir Objetivo do Experimento:*** Nesta etapa, definimos o objetivo do experimento. Em outras palavras, sobre o que queremos aprender, qual o objeto de análise, quais os aspectos de interesse, qual a finalidade do estudo, sob qual ponto de vista e em que contexto o estudo será feito.
2. ***Planejar Experimento:*** Essa etapa corresponde ao planejamento do experimento, com as seguintes fases:

a. ***Selecionar o Contexto****:* Contexto geral, incluindo a localidade ou organização, juntamente com a totalidade do público alvo (a população e não a amostra), as pessoas às quais queremos que os resultados se apliquem.

b. ***Enunciar as Questões de Pesquisa:*** Uma questão de pesquisa é a declaração de uma indagação específica que o pesquisador deseja responder para abordar o problema de pesquisa. A questão ou questões de pesquisa orientam os tipos de dados a serem coletados e o tipo de estudo a ser desenvolvido.

c. ***Definir Variáveis Dependentes:*** São todas as variáveis que se deseja estudar em um processo de experimentação, ou seja, são as variáveis que se deseja observar o efeito das mudanças aplicadas nas variáveis independentes. São também chamadas variáveis de resposta e são derivadas diretamente das hipóteses estabelecidas.

d. ***Definir Variáveis Independentes:*** São todas as variáveis de um processo de experimentação que são manipuladas e controladas, como, por exemplo, um método de desenvolvimento, a experiência das pessoas e o ambiente. São também chamadas variáveis de entrada.

e. ***Formular Hipóteses Teóricas:*** O objeto de estudo deve ser traduzido em hipóteses formais. Uma hipótese é uma conjectura, uma resposta provisória que, de acordo com certos critérios, será Rejeitada ou Não-Rejeitada. Uma hipótese deve ser formulada de modo que se atribua ao **acaso** a ocorrência do fenômeno observado (hipótese nula). Formule outra, que sirva de alternativa à primeira, se ficar demonstrado que o **acaso** não pode ser responsabilizado pelo fenômeno observado (hipótese alternativa).

   Nesse momento, na formulação das hipóteses, deve-se fazer uso de variáveis teóricas, definidas por meio de conceitos abstratos. Na etapa de validação dos dados, será formalizada a operacionalização dessas variáveis, ou seja, para inferir previsões a partir das hipóteses, faz-se necessária a representação de uma

variável teórica por meio de uma variável operacional, buscando evitar uma possível confusão entre a hipótese operacionalizada e a teoria ou generalização que esta pretende testar (Kluger & Tikochinsky, 2001).

Por exemplo, em um estudo sobre depressão, que é um fenômeno relacionado a outras variáveis abstratas, tais como ruminação e ansiedade, medir essa variável diretamente é uma tarefa difícil, uma vez que a operacionalização das hipóteses exige objetividade. Ou seja, o pesquisador precisará recorrer a coisas concretas, tais como os níveis hormonais da pessoa ou questionários psicológicos que gerem scores objetivos.

Desta forma, depois da definição das hipóteses e das variáveis teóricas, deve-se escolher uma ou mais variáveis operacionais que representem bem o conceito a ser avaliado.

Na validação dos dados, depois do teste de normalidade, a hipótese operacional, definitiva matematicamente, poderá ser formalizada, uma vez que já estará definido, por exemplo, se será mesmo usada uma média (teste paramétrico) ou mediana (teste não paramétrico).

f. ***Selecionar Participantes e/ou Objetos:*** Diante da impossibilidade de avaliar todo o universo, na maioria dos casos, é necessário definir uma parcela do universo (amostra) que podemos avaliar. Essa amostra deve ser a mais representativa possível.

g. ***Projetar Experimento:*** O projeto experimental é um plano completo para avaliar as variáveis experimentais frente às variáveis de controle, acompanhando e mitigando a influência das variáveis de estado. Envolve os objetos, medidas, instruções, técnicas, formato experimental e tratamentos.

h. ***Definir Instrumentação:*** Aqui serão descritas as ferramentas e ambientes usados.

3. ***Operar Experimento:*** Após o planejamento do experimento, parte-se para a etapa em que o experimento é rodado. Essa etapa se divide em:

a. ***Preparar Experimento:*** Antes de conduzir o experimento real, a preparação pode incluir um estudo piloto para confirmar o cenário experimental, ajudar a organizar os fatores experimentais (por exemplo, experiência do sujeito) ou inocular os sujeitos.
b. **Executar Experimento:** Consiste das seguintes etapas:
    i. **Descrever Ambiente:** Descrição do ambiente no qual o experimento foi executado;
    ii. **Rodar o Experimento:** Execução do experimento propriamente dito;
    iii. **Coletar Dados:** Coleta dos dados do experimento.
c. **Definir e Executar Validação dos Dados:** Consiste na definição dos métodos estatísticos que serão usados para validar os dados coletados pelo experimento, bem como a sua execução. A análise dos dados pode incluir uma combinação de métodos quantitativos e qualitativos. A triagem preliminar dos dados, provavelmente usando gráficos e histogramas, geralmente precede a análise formal dos dados. O processo de análise dos dados requer a investigação de quaisquer suposições subjacentes (por exemplo, distributivas) antes da aplicação dos modelos estatísticos e testes. Ou seja, após a confirmação da normalidade e/ou homoscedasticidade dos dados, a hipótese operacional, em termos matemáticos, poderá ser formalizada e testada.

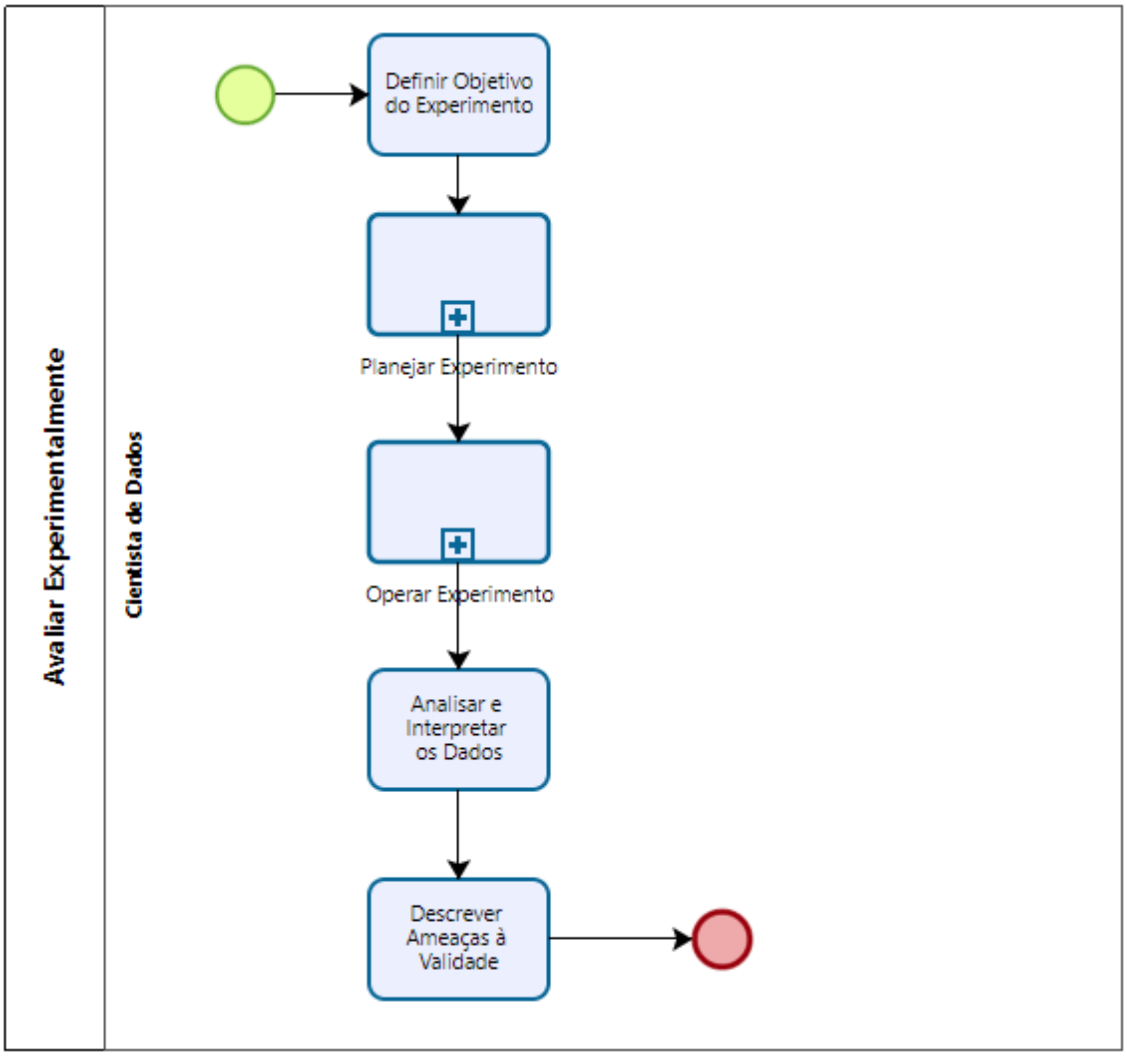
4. ***Analisar e Interpretar os Dados:*** Consiste na análise e interpretação dos resultados, descrevendo também os resultados estatísticos que serviram de base para as conclusões.
5. ***Descrever Ameaças à Validade:*** Descrição de tudo que possa ameaçar o resultado do experimento. As ameaças podem ser (Colaço Júnior, 2018):
    a. Conclusão: Relacionada à habilidade de chegar a uma conclusão correta a respeito dos relacionamentos entre o tratamento e o resultado. Exemplo: Escolha do teste estatístico;
    b. Construção: Aspectos relacionados ao projeto e fatores humanos. Exemplo: Pesquisador projeta baseado no que ele espera;
    c. Interna: Define se o relacionamento entre o tratamento e o resultado é causal, sem a influência de outro fator que pode não

ter sido medido. Exemplo: Um participante dá informação para outro;

d. Externa: Condições que limitam a capacidade de generalizar. Consequência das outras ameaças. Exemplo: Seleção de participantes não representativos, tempo e/ou lugar atípicos.

Na imagem abaixo, podemos observar como essas subatividades estão dispostas na atividade macro *Avaliar Experimentalmente*.

Figura 20 – Atividades do Processo Avaliar Experimentalmente



Na Figura 21, temos o fluxo do subprocesso: Planejar Experimento e, na Figura 22, o fluxo do subprocesso: Operar Experimento.

Figura 21 – Atividades do Processo Planejar Experimento

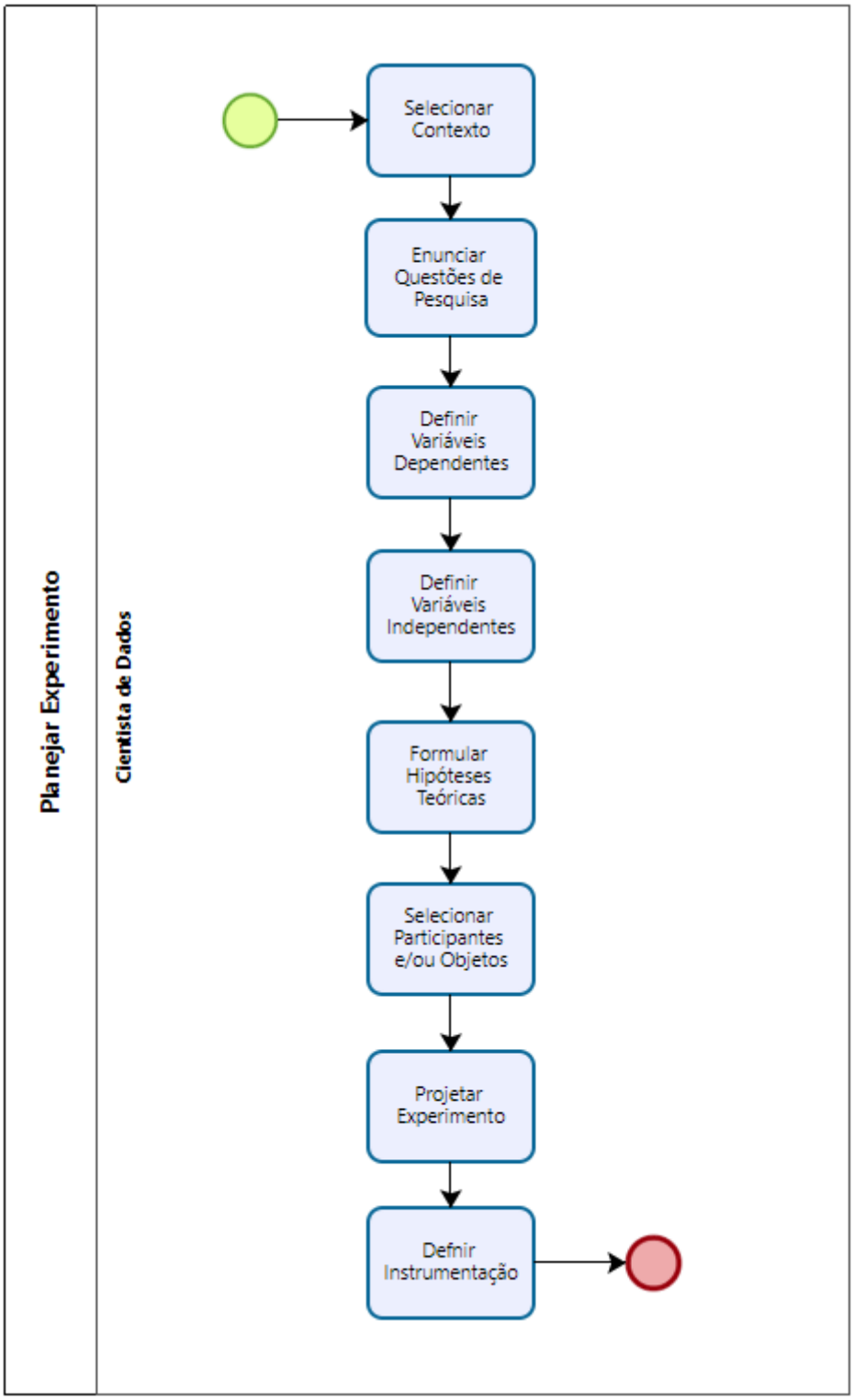
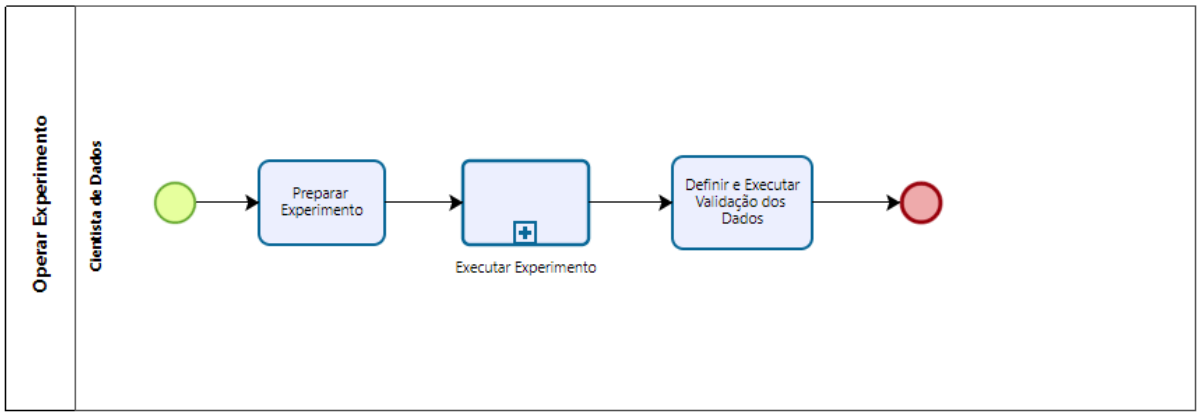


Figura 22 – Atividades do Processo Operar Experimento



4.4.2.4.3 RESULTADOS

Como saída esperada do subprocesso **Avaliar Experimentalmente**, temos a definição, a execução e o resultado de todo o projeto experimental, bem como o modelo de mineração eleito.

4.4.2.5 VALIDAR OBJETIVOS ESTRATÉGICOS

Nesta atividade, é validado o objetivo estratégico do DM. Como principal benefício, são colhidas validações e aceitação por parte do cliente, para a implantação do modelo de mineração criado. O objetivo, entradas, subatividades e resultados desta atividade são mostrados na Tabela 13, que segue abaixo.

Tabela 13 - Descritivo da atividade: Validar Objetivos Estratégicos

Atividade: Validar Objetivos Estratégicos	
Objetivo	Formalização e aceitação dos objetivos estratégicos selecionados para implantação do DM.
Entradas	• Escopo preliminar e grade eleita; • Modelo de Mineração de Dados Eleito; • Descrição e/ou filmagem, com User Stories novas ou incrementadas (opcional); • Protótipo de Visualização.
Subatividades	1. Validar Objetivos Estratégicos;
Resultados (Saídas)	• Documento de Aceitação ou Lista de não Conformidades

4.4.2.5.1 ENTRADAS

A entrada dessa atividade são os documentos descritos nos itens 4.4.2.1.3 e 4.4.2.4.3.

4.4.2.5.2 SUBATIVIDADES

É todo esforço necessário para execução da atividade macro denominada: *Validar Objetivos Estratégicos*. Para essa atividade, foram definidas a seguintes subatividades:

1. ***Validar Objetivos Estratégicos:*** A validação dos objetivos estratégicos é o momento de externar tudo aquilo que foi levantado e desenvolvido junto às equipes técnicas e a área de negócio, de tal forma que se possa validar, confirmar e assumir o compromisso da implantação do que foi definido para a DM. Inconformidades ainda podem ser encontradas e o processo recomeça para validação e ajustes. Nesta etapa, também é importante revalidar o protótipo da aplicação e visualizações a serem implementadas.
2. ***Revisar e Ajustar:*** Conforme subatividade de mesmo nome, da atividade macro *Desenvolver Objetivo do Processo de DM*.

4.4.2.5.3 RESULTADOS

Como saída esperada do subprocesso **Validar Objetivos Estratégicos**, há uma lista de inconformidades ou um documento de aceitação que autorizará o seguimento com o processo "*Implementar DM*". Na tabela 14, disponibilizamos um padrão a ser utilizado na documentação desse modelo.

Tabela 14 – Validação dos Objetivos Estratégicos

Checklist de Validação para Implantação do DM	
Equipe Avaliadora	**Data**
Área de Inteligência e Gerência	-
Objetivos Traçados e Protótipo	**Validação**
Objetivo Estratégico	<Aceito/Recusado>
Objetivo da Mineração	<Aceito/Recusado>
Protótipo	<Aceito/Recusado>
Conclusão	
Este documento formaliza o aceite da entrega, considerando-a em conformidade com os requisitos e os critérios de aceitação definidos, bem como considerando a validação de todos os documentos produzidos.	

Participante	Assinatura	Data
Área de Inteligência		-
Gerência (Área de Negócio)		-

4.4.2.6 IMPLEMENTAR DM

Por fim, uma vez que os objetivos estratégicos foram validados e aprovados, o modelo de mineração de dados deve ser implementado conforme o protótipo definido. O objetivo, entradas, subatividades e resultados desta atividade são mostrados na Tabela 15, que segue abaixo.

Tabela 15: Descritivo da atividade: Implementar DM

Atividade: Implementar DM	
Objetivo	Implementação do Modelo de Mineração Selecionado
Entradas	• Modelo de Mineração de Dados Eleito; • Protótipo de Visualização.
Subatividades	1. Selecionar Algoritmo(s) de Melhor Efetividade; 2. Treinar com toda a Base Disponível; 3. Implementar Aplicação com uso do(s) Algoritmo(s).
Resultados (Saídas)	• Modelo de Mineração de Dados Implementado

4.4.2.6.1 ENTRADAS

A entrada desse subprocesso é o documento produzido e descrito no item 4.4.2.1.3 e 4.4.2.4.3.

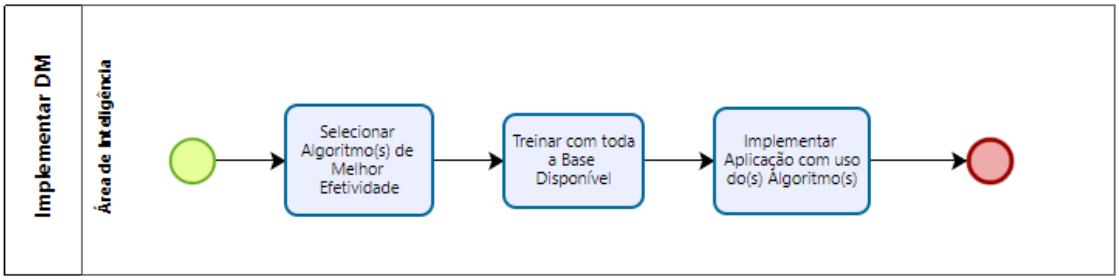
4.4.2.6.2 SUBATIVIDADES

É todo esforço necessário para execução da atividade macro denominada: *Implementar DM*. Para essa atividade, foram definidas a seguintes subatividades:

1. ***Selecionar Algoritmo(s) de Melhor Efetividade:*** Como resultado da avaliação experimental, tem-se o Modelo de Mineração de Dados Selecionado, com um ou mais algoritmos eleitos.
2. ***Treinar com toda a Base Disponível:*** Toda a base é usada para treinamento dos algoritmos selecionados.
3. ***Implementar Aplicação com uso do(s) Algoritmo(s)***: A aplicação é desenvolvida com base no protótipo definido e aprovado.

Na Figura 23, temos o fluxo do subprocesso *Implementar DM*.

Figura 23 – Atividades do Processo Implementar DM



4.4.2.6.3 RESULTADOS

Como saída esperada do subprocesso **Implementar DM**, temos a implementação do modelo de mineração de dados selecionado de acordo com o protótipo definido.

4.5 ESTUDO DE CASO

Para avaliar o processo proposto, foi planejado e realizado um estudo de caso dentro de uma instituição de ensino federal. Nas próximas seções, o estudo de caso é apresentado e detalhado.

4.5.1 ETAPAS E DIRETRIZES PARA O ESTUDO DE CASO

As principais etapas para a realização do estudo de caso foram as seguintes:

- Definição do objetivo: Definir os objetivos do estudo de caso;
- Planejamento: Planos, instrumentos e projeto do estudo de caso;
- Operação do Estudo de Caso: Definir a preparação e execução;
- Consolidação e divulgação de resultados.

O detalhamento de cada uma dessas etapas será realizado nas próximas seções.

4.5.2 DEFINIÇÃO DO OBJETIVO

O objetivo deste estudo foi avaliar o processo descrito na seção 4.4.2, o qual visa o alinhamento estratégico e a avaliação experimental no desenvolvimento de aplicações de *Data Mining* e *Data Science*.

Desta forma, as questões de pesquisa que precisam ser respondidas são as seguintes: 1ª) Um processo de BI dirigido à estratégia pode ser estendido para o desenvolvimento de aplicações de *Data Mining* e *Data Science* avaliadas experimentalmente?; 2ª) Um processo que encapsula experimentação disciplinará o exercício da Ciência de Dados?

Por fim, faz-se necessária a elaboração de uma suposição passível de investigação dentro da proposta desse trabalho. A suposição em questão é: Uma metodologia de BI dirigida

à estratégia pode ser estendida para a contemplação e o desenvolvimento de aplicações de *Data Mining* e *Data Science* avaliadas experimentalmente.

4.5.3 PLANEJAMENTO

Foi planejada a aplicação do processo proposto numa instituição de ensino federal. O projeto aconteceu nos meses de abril e maio de 2021, tendo a participação dos membros do "Programa 5A", projeto de inteligência acadêmica da referida instituição. Nessa aplicação, o processo guiou o desenvolvimento de um projeto de Data Mining, cumprindo as 6 atividades descritas anteriormente.

Foi planejada uma reunião de apresentação do processo para toda a equipe envolvida, bem como reuniões semanais de acompanhamento quanto ao uso do processo. Como plano de comunicação, foi acordado o uso dos mecanismos disponibilizados pela empresa, tais como: E-mail, Chat e Videoconferência.

4.5.3.1 SELEÇÃO DOS PARTICIPANTES

A escolha se deu por conveniência e por cota, pelo acesso dos pesquisadores à instituição, a qual possui uma rara equipe bem definida de inteligência, com características específicas da população que o projeto pretende atingir. As quantidades, funções e tempo de experiência profissional dos participantes, foram: um Gerente, com experiência de até 3 anos, especificamente nesta área de atuação; dois Analistas de Dados, um com até 3 anos de experiência e o outro entre 5 e 10 anos; e seis Bolsistas com menos de um ano de experiência na área.

4.5.3.2 INSTRUMENTAÇÃO

O processo de instrumentação se deu com o processo proposto na seção 4.4.2. A avaliação da execução do processo foi por meio de um questionário, disponível em https://forms.gle/CkRso2fVD7JfM7QM9.

4.5.4 OPERAÇÃO

Nesta subseção, estão descritas a preparação e execução do estudo em questão.

4.5.4.1 PREPARAÇÃO

Depois do treinamento no processo, a equipe entendeu a necessidade e levantou os objetivos estratégicos da organização, selecionando um prioritário para a aplicação do processo, o qual pode ser apoiado por aplicações inteligentes. Este levantamento ocorreu em conjunto

com a equipe estratégica, a qual concordou com um projeto de Mineração de Dados para apoio à mitigação da evasão escolar.

4.5.4.2 EXECUÇÃO

Para cada atividade, foram documentados os artefatos produzidos, que serviriam de requisitos para as atividades sucessoras. Por fim, foi validado o apoio ao respectivo objetivo estratégico selecionado. O detalhamento da execução será feito na próxima seção.

4.5.5 RESULTADO

Para que as questões da pesquisa pudessem ser respondidas, foram analisados os documentos produzidos, resultantes das atividades do processo proposto. Estas serão apresentadas nas próximas seções.

4.5.5.1 DESENVOLVER OBJETIVO DO PROCESSO DE DM

Com o propósito de desenvolver o projeto de Mineração para apoiar o atingimento do objetivo de negócio selecionado, foi desenvolvido o objetivo do processo de Mineração de Dados. O documento foi produzido a partir do *template* definido na seção 4.4.2.1.3, conforme tabela abaixo:

Tabela 16 – Escopo Preliminar do Objetivo do DM

OBJETIVO				
Diminuir a evasão escolar, identificando quais fatores podem contribuir para o insucesso acadêmico dos discentes.				
DO PONTO DE VISTA DO NEGÓCIO				
OBJETO	**PROPÓSITO**	**FOCO DE QUALIDADE**	**PONTO DE VISTA**	**CONTEXTO**
Discentes do Nível de Graduação de todos os *campi* do Instituto Federal	Avaliar quais as características dos discentes podem compor os fatores de insucesso acadêmico	Diminuir o insucesso acadêmico dos discentes, no quesito da evasão escolar, em 5%, até o final do ano	Reitor(a), Pró-reitor(a) de Ensino, Coordenadores(as) de Controle Docente e Discente, Gerentes e Diretores Ensino	Área de Ensino
DO PONTO DE VISTA DA MINERAÇÃO				
OBJETO	**PROPÓSITO**	**FOCO DE QUALIDADE**	**PONTO DE VISTA**	**CONTEXTO**
Algoritmos de	Avaliar e Predizer	Alcançar uma previsibilidade de	Gerentes, alunos,	Discentes do nível de

Mineração de Dados		evasão escolar com acurácia maior ou igual a 90%	profissionais de análise de dados e cientistas de dados	graduação de todos os *campi do IFS.*

FOCO DE QUALIDADE (QUESTÕES E MÉTRICAS)	**FATORES DE VARIAÇÃO**
PE-G-Q1 – Qual o percentual de evasão escolar em um ano letivo? • *Peva (A): Percentual de evasão escolar no ano A.* *ALG-G-Q1 – Quais as acurácias dos principais algoritmos de machine learning, para a tarefa de previsibilidade da evasão escolar?* • *Acur (n): Acurácia do algoritmo n*	-
HIPÓTESES DE LINHA DE BASE	**IMPACTO NAS HIPÓTESES DE LINHA DE BASE**
Peva(2020) = 8,22% *Acur(Árvore de Decisão) = 70%*	-

INTERPRETAÇÃO DO MODELO

PE-G-Q1 = Peva(2021) / Peva(2020) <= 0.95 <Diminuição de 5% na evasão escolar de 2021 em relação à evasão de 2020>

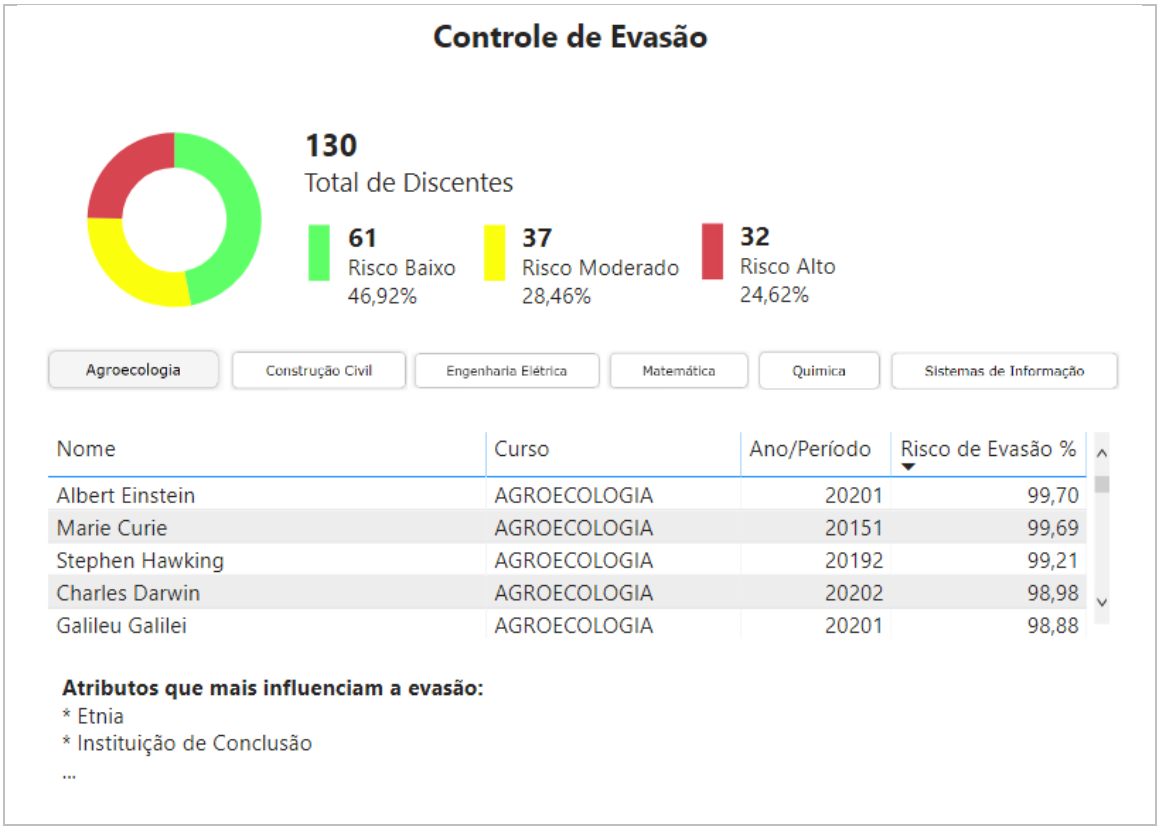
ALG-G-Q1 = Acur (n) >= 90%

OBJETIVO GERAL COM ESCOPO PRELIMINAR DE VISUALIZAÇÃO

ANALISAR algoritmos de mineração de dados, **COM A FINALIDADE DE** predizer e avaliar, **COM RESPEITO À** eficácia da previsão da evasão escolar, **DO PONTO DE VISTA** de Gerentes de Ensino, Alunos, Analistas e Cientistas de Dados, **NO CONTEXTO** de discentes ativos e matriculados no nível de graduação em um Instituto Federal.

Como resultado da mineração, será apresentada a probabilidade geral de evasão, bem como a probabilidade individual, identificando os discentes ativos e matriculados propensos a evadirem, possibilitando a adoção de medidas que venham a mudar esse quadro. Na imagem abaixo, temos um protótipo inicial das saídas do processo de mineração.

Figura 24 - Protótipo



4.5.5.2 PREPARAR DADOS

Na atividade, *Preparar Dados*, foi selecionado o seguinte conjunto de dados (vide parte dos dados do *DataSet*, na Tabela 17):

Tabela 17 – *Dataset*

Atributo	**Descrição**
sexo	Sexo do discente
idade	Idade do discente
tipo_instituicao_conclusao	Tipo de instituição de ensino que concluiu o segundo grau
raca	Etnia
est_civil	Estado civil
qtd_trac	Quantidade de disciplinas trancadas
reab_matricula	Indica se o discente efetuou reabertura de matrícula
qtd_ap_med_p	Quantidade média de disciplinas aprovadas por período
qtd_ap_1p	Quantidade de disciplinas aprovadas no primeiro período
qtd_rep_med_p	Quantidade média de disciplinas reprovadas por período
qtd_rep_1p	Quantidade de disciplinas reprovadas no primeiro período
qtd_per_cur	Quantidade de períodos cursados pelo discente
cra	Coeficiente de rendimento acadêmico
qtd_disciplinas_concluidas	Quantidade total de disciplinas que foi aprovado
qtd_disciplinas	Quantidade total de disciplinas que se matriculou
media_geral	Média geral de desempenho do discente no curso
media_faltas	Média geral de faltas do discente no curso
cotista	Indica se o discente ingressou por sistema de cotas

4.5.5.3 PROJETAR MODELO

Foram selecionados os algoritmos de classificação disponibilizados pela biblioteca Pycaret. O Pycare é uma biblioteca Python de machine learning, com código aberto, a qual foi selecionada devido a sua compatibilidade com o Google Colaboratory, ambiente colaborativo usado na organização (Gaian & Hotti, 2021). Os algoritmos são: Ada Boost Classifier, Decision Tree Classifier, Extra Trees Classifier, Gradient Boosting Classifier, K Neighbors Classifier, Light Gradient Boosting Machine, Linear Discriminant Analysis, Logistic Regression, Naive Bayes, Quadratic Discriminant Analysis, Random Forest Classifier, Ridge Classifier e SVM - Linear Kernel.

Para transformação dos dados, foram realizados os seguintes procedimentos: O atributo tipo_instituicao_conclusao teve os valores nulos preenchidos pelo valor ‘OUTRA’; O atributo est_civil teve os valores nulos preenchidos pelo valor ‘Outro’; Os atributos: 'sexo', 'tipo_instituicao_conclusao', 'raca' e 'est_civil' tiveram seus valores alterados para números ao invés de textos; Foram filtrados apenas os registros que possuem status em: 'CANCELADO', 'ATIVO', 'CONCLUÍDO', JUBILADO, 'ATIVO - GRADUANDO', 'ATIVO - FORMANDO'; Os status em textos foram substituídos para binários, sendo: Status: 'CANCELADO', 'JUBILADO', substituídos pelo número 1 (alvo - evasão); Status: 'ATIVO', 'CONCLUÍDO', 'ATIVO - GRADUANDO', 'ATIVO - FORMANDO', substituídos por 0 (controle - alunos que não evadiram).

Por questões de espaço, a tabela de parâmetros dos algoritmos não será apresentada.

4.5.5.4 AVALIAR EXPERIMENTALMENTE

Foi feito um experimento para avaliar o melhor algoritmo em termos de eficácia, com foco na evasão escolar. Os principais algoritmos poderão ser usados para formar um metamodelo de predição que ajudará a gestão da evasão. Desta forma, foi feito o experimento descrito na tabela abaixo.

Tabela 18 – Avaliação Experimental do Modelo de Mineração Proposto

DEFINIÇÃO DO OBJETIVO DO EXPERIMENTO
Definição do Objetivo
Já descrito na Tabela 16.
PLANEJAMENTO DO EXPERIMENTO
Seleção de Contexto

O experimento foi "in vitro", pois os dados foram retirados do ambiente real, para que pudessem ser transformados e então utilizados em um ambiente controlado. Foram considerados os dados de alunos de todos os cursos de Graduação, perfazendo ingressantes entre 2003, ano no qual iniciaram os primeiros cursos de graduação, e 2020. A seleção de dados levou em consideração atributos pessoais, acadêmicos e sociais.
Questões de Pesquisa
No contexto da evasão escolar, entre os algoritmos selecionados, qual deles apresenta os melhores indicadores em termos de eficácia?; A acurácia supera os 90% definidos como meta ?;
Variáveis Dependentes
Classificações, das quais podem ser derivadas: Acurácia, Sensibilidade, Precisão e Medida-F1.
Variáveis Independentes
Dataset descrito na Tabela 17 e os algoritmos anteriormente listados.
Formulação de Hipóteses Teóricas
• H_0: Os algoritmos($_{1,2..n}$) possuem a mesma eficácia. • H_1: Os algoritmos($_{1,2..n}$) possuem eficácias diferentes.
Seleção de Participantes e Objetos
Foram selecionados todos os alunos de graduação do Instituto, totalizando 10.949 alunos, dos quais 6.961 (63,57%) faziam parte da classe alvo (Evadidos) e 3.988 (36,42%) faziam parte da classe controle (Não Evadidos). Uma das métricas utilizadas foi a acurácia, a qual exige o balanceamento das classes. Para isso, foi necessário efetuar o processo de balanceamento, o qual levou em consideração a maior quantidade de registros presentes em cada classe, 3.988, sendo o total final um valor superior a uma amostra para população infinita, conforme embasamento na literatura (Pinto, 2015). Considerando a população do Instituto, a amostra final de 7.976 alunos tem margem de erro de 0,57%, para uma confiabilidade de 95%.
Projeto do Experimento
Para a avaliação do modelo, foi utilizada a abordagem 10-Fold Cross-validation, em que os dados são divididos em 10 partes, mantendo suas proporções. Assim, são realizados 10 testes, nos quais uma parte dos dados é separada para ser testada posteriormente. Além disso, será possível obter resultados anuais, semestrais, bimestrais, trimestrais, mensais ou quinzenais.
Instrumentação

Para o processo de mineração de dados, foi utilizada a biblioteca Python pycaret, a qual é uma biblioteca de aprendizado de máquina open source, de alto nível, cujo objetivo é maximizar o desempenho de comparação e usabilidade da biblioteca Scikit-learn. Para execução do código Python, foi utilizado o ambiente na nuvem do Google Colab, o qual é voltado à criação e execução de códigos em Python, diretamente em um navegador, sem a necessidade de nenhum tipo de instalação de software em máquinas locais. Os dados utilizados para a análise provêm do SIGAA, sistema acadêmico utilizado na instituição, o qual possui como SGBD o PostgreSQL.

OPERAÇÃO DO EXPERIMENTO

Preparação

Consistiu na preparação do dataset que foi usado na mineração. Essa preparação se deu seguindo o subprocesso Preparar Dados, do processo proposto.

Execução

Consistiu na realização do processo classificatório nos dados, planejado no projeto do experimento, para cada algoritmo de mineração selecionado, utilizando as demais variáveis independentes.

Validação dos Dados

Como auxílio para análise, interpretação e validação, foram utilizados três tipos de testes estatísticos: Teste Shapiro-Wilk, Teste de Levene e o Teste T Pareado.

Foi utilizado o Teste Shapiro-Wilk para o teste de Normalidade e o Teste de Levene para o teste de homocedasticidade. Uma vez que o pressuposto da normalidade foi atendido e o da homocedasticidade não, foi utilizado o Teste T Pareado para testar as hipóteses.

RESULTADOS

Análise e Interpretação dos Dados

Após a execução dos algoritmos, utilizando a abordagem 10-Cross-validation, foram obtidos os resultados das classificações, os quais serão apresentados na Figura 25.

Figura 25 - Comparativo das Métricas dos Algoritmos

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
lightgbm	Light Gradient Boosting Machine	0.9070	0.9693	0.8912	0.9207	0.9056	0.8141	0.8147	0.163
gbc	Gradient Boosting Classifier	0.9045	0.9670	0.8891	0.9179	0.9031	0.8091	0.8098	0.919
rf	Random Forest Classifier	0.9022	0.9654	0.8933	0.9101	0.9014	0.8044	0.8050	0.866
ada	Ada Boost Classifier	0.8841	0.9564	0.8790	0.8883	0.8835	0.7682	0.7685	0.325
et	Extra Trees Classifier	0.8807	0.9522	0.8797	0.8820	0.8807	0.7614	0.7617	0.823
dt	Decision Tree Classifier	0.8714	0.8718	0.8733	0.8706	0.8717	0.7428	0.7431	0.057
knn	K Neighbors Classifier	0.8494	0.9154	0.8665	0.8383	0.8520	0.6987	0.6994	0.157
lr	Logistic Regression	0.8110	0.9005	0.8175	0.8076	0.8123	0.6221	0.6225	0.911
ridge	Ridge Classifier	0.8105	0.0000	0.8264	0.8016	0.8135	0.6210	0.6219	0.034
lda	Linear Discriminant Analysis	0.8105	0.8992	0.8257	0.8021	0.8134	0.6210	0.6218	0.068
nb	Naive Bayes	0.7625	0.8535	0.6804	0.8186	0.7397	0.5251	0.5360	0.032
svm	SVM - Linear Kernel	0.6896	0.0000	0.7741	0.7410	0.7003	0.3791	0.4400	0.110
qda	Quadratic Discriminant Analysis	0.5157	0.5160	0.1682	0.5710	0.2444	0.0319	0.0496	0.044

Os resultados foram utilizados para responder à questão de pesquisa. Os algoritmos Light Gradient Boosting Machine e Gradient Boosting Classifier (GBC) possuem médias para as métricas bem semelhantes e próximas, uma vez que o Light é uma customização da Microsoft, contudo, este possui tempo de execução inferior ao GBC, que é seguido, em termos de eficácia, pelo Random Forest Classifier, com médias também semelhantes aos primeiros colocados. Os demais algoritmos apresentados não alcançaram a meta mínima definida de 90%.

Apesar do posicionamento do algoritmo Light, para definir qual o melhor algoritmo entre os que atingiram a meta, são necessárias evidências estatísticas suficientemente conclusivas. Desta forma, foi definido um nível de significância de 0,05. Ao aplicar o Teste de Shapiro-Wilk, para análise da normalidade da distribuição dos dados, foram obtidos os p-values apresentados na tabela abaixo, na qual, observa-se valores acima do nível de significância adotado, concluindo-se que as distribuições dos dados para as métricas Acurácia e Medida-F1 são normais.

Tabela 19 - Resultado do Teste de Shapiro-Wilk, para análise da normalidade dos dados

Algoritmo	**Acurácia**	**Medida-F1**
Light Gradient Boosting Classifier	0.0856	0.6809
Gradient Boosting Classifier	0.2881	0.1274
Random Forest Classifier	0.1525	0.7854

Como ficaram apenas 3 algoritmos e os dados são normais, optou-se por comparar os algoritmos dois a dois, por meio do Teste T Pareado. Além disso, também devido à normalidade, a operacionalização da hipótese já pode ser realizada com o uso da média. Desta forma, foram formalizadas as seguintes hipóteses (para cada métrica avaliada e para cada dupla de algoritmos):

- H_0: Os algoritmos $(_{1,2})$ possuem médias iguais para a métrica.
  μ_1(métrica) = μ_2(métrica);
- H_1: Os algoritmos $(_{1,2})$ possuem médias diferentes para a métrica.
  μ_1(metrica) ≠ μ_2(métrica).

Aplicando o teste T Pareado para cada dupla de algoritmos, foram obtidos os seguintes p-values:

Tabela 20 - Resultado do Teste T Pareado

Algoritmo 1 contra Algoritmo 2	Acurácia	Medida-F1

Light Gradient Boosting Classifier/Gradient Boosting Classifier	0.4951	0.5045
Light Gradient Boosting Classifier/Random Forest Classifier	0.08812	0.1038
Gradient Boosting Classifier/Random Forest Classifier	0.142	0.1603

Uma vez que os p-values apresentaram valores maiores do que o nível de significância definido, as Hipótese nulas (H_0) não podem ser rejeitadas. Em outras palavras, as diferenças entre as médias dos algoritmos não são grandes o suficiente para serem estatisticamente significativas, indicando que os 3 algoritmos se equivalem, quanto às medidas comparadas.

Considerando as métricas mais importantes para o problema em questão, pode ser construído um sistema com um modelo que usa os três algoritmos vencedores, decidindo a classificação por meio do resultado que mais ocorre. A criação desse metamodelo gera uma nova hipótese a ser testada.

Ameaças à Validade

Ameaças à validade interna: O sistema acadêmico atual está presente na Instituição desde 2017, o qual herdou a base do sistema acadêmico legado, com várias informações inconsistentes, principalmente até meados de 2007. Esta ameaça foi mitigada com a realização do processo de limpeza dos dados, diminuindo a probabilidade do uso de informações mais antigas incorretas.

4.5.5.5 VALIDAR OBJETIVOS ESTRATÉGICOS

Na atividade, *Validar Objetivos Estratégicos*, foi aplicado um *checklist* de validação para implementação do DM.

Tabela 21 – Validação dos Objetivos Estratégicos

Checklist de Validação para Implantação do DM	
Equipe Avaliadora	**Data**
Área de Inteligência e Gerência	18/05/2021
Objetivos Traçados e Protótipo	**Validação**
Objetivo Estratégico	Aceito
Objetivo da Mineração	Aceito
Protótipo	Aceito
Conclusão	
Este documento formaliza o aceite da entrega, considerando-a em conformidade com os requisitos e os critérios de aceitação definidos, bem como considerando a validação de todos os documentos produzidos.	

Participante	Assinatura	Data
Área de Inteligência	<Suprimido>	18/05/2021
Gerência (Área de Negócio)	<Suprimido>	18/05/2021

4.5.5.6 IMPLEMENTAR DM

A Atividade de implementação não foi acompanhada por esse estudo.

4.5.6 AVALIAÇÃO DO PROCESSO

Para que possamos interpretar os resultados apresentados, uma avaliação qualitativa foi proposta. Segundo Demo (2012), a avaliação qualitativa de uma pesquisa busca preservar e procurar informações na realidade. A informação qualitativa pode obter confiabilidade sobre a correta execução dos procedimentos.

4.5.6.1 MÉTODO DE AVALIAÇÃO

A avaliação foi aplicada à equipe de especialistas da instituição, com experiência na área, utilizando o questionário descrito na seção 4.5.3.2, o qual contém um conjunto de perguntas qualitativas sobre o processo aqui proposto. Todo o processo descrito neste trabalho, bem como os artefatos produzidos, foi disponibilizado e vivenciado por cada avaliador. Ao final do projeto, o questionário foi aplicado com supervisão.

4.5.6.2 ANÁLISE DAS AVALIAÇÕES

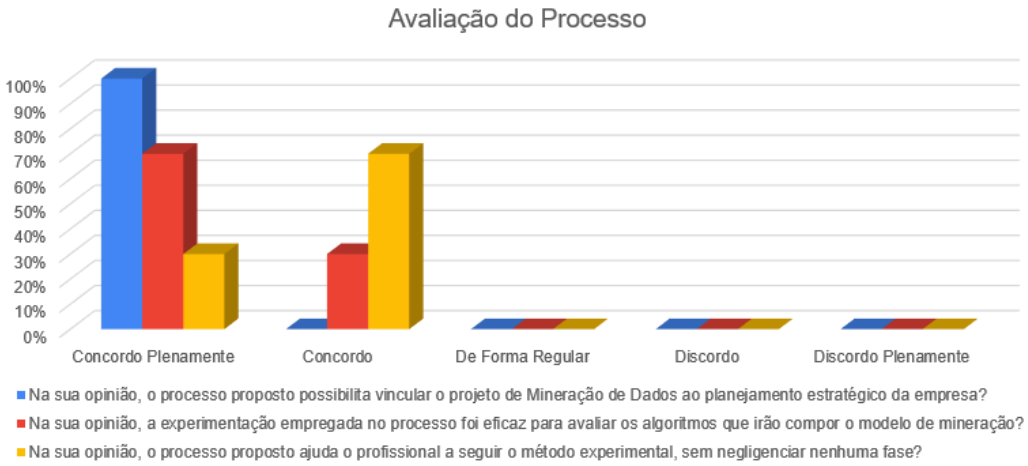
Confirmando os dados evidenciados por Lima et al. (2017) e a revisão *quasi*-sistemática da literatura executada na metodologia deste trabalho (Cruz, Colaço Júnior & Gois, 2021), a instituição estudada não utilizava nenhuma metodologia formal para desenvolvimento de seus projetos de Mineração de Dados.

Todos os entrevistados colocaram como importante o Alinhamento Estratégico dos projetos de Mineração de Dados. Além disso, três informaram que a experimentação também pode ser usada para orientar o desenvolvimento de diversos produtos e permitir que a organização avalie o ROI dos projetos de software.

Sobre a possibilidade de vincular o projeto de Mineração de Dados ao planejamento estratégico da empresa, a concordância foi plena em relação ao apoio dado pelo processo proposto. Quanto à eficácia da experimentação na avaliação dos algoritmos que irão compor os modelos de mineração, 70% dos entrevistados responderam que concordam plenamente e 30% que concordam.

Perguntados se o processo proposto ajuda o profissional a seguir o método experimental, sem negligenciar nenhuma fase, 30% responderam que concordam plenamente e 70% responderam que concordam. A síntese da avaliação pode ser vista na Figura 26.

Figura 26 – Avaliação do Processo Proposto



Desta forma, os resultados iniciais trouxeram evidências de que é possível integrar alinhamento estratégico, método científico e uma metodologia de desenvolvimento de aplicações de DM, fomentando a experimentação como elemento de retroalimentação para o Planejamento Estratégico.

4.5.6.3 AMEAÇAS À VALIDADE

O estudo de caso pode ser influenciado positivamente pelos proponentes do processo, os quais tendem a defender os seus produtos e ocultar problemas. Este fato pode ocasionar um fenômeno estudado pela psicologia denominado *Demand Characterization*, o qual considera que um artefato experimental pode ter uma interpretação, pelos participantes, do propósito do experimento, levando à mudança de comportamento inconsciente, para se adaptar a esta interpretação (Orne, 1962). Para mitigar este fator, pode-se dizer que foram utilizadas pelo menos duas abordagens diferentes: *The More The Merrier e Unobtrusive Manipulations and Measures* (Orne, 1962). Respectivamente, na primeira, para mitigar o viés, o processo foi aplicado pelos funcionários da empresa, não envolvidos com a pesquisa. A segunda nos norteou a não informar, durante a condução, quais fatores e métricas seriam avaliados, de modo que os participantes não tivessem pistas sobre a hipótese de pesquisa. Por fim, uma entrevista foi feita com os participantes, com o intuito de avaliar qualitativamente os resultados iniciais.

4.6 CONCLUSÃO E TRABALHOS FUTUROS

A principal contribuição deste trabalho foi proposta e avaliação de uma abordagem para disciplinar o alinhamento estratégico e a construção experimental no desenvolvimento de aplicações de Data Mining e Data Science. O trabalho foi consolidado pela realização de um estudo de caso em uma instituição de ensino federal, a fim de ajudar a consolidar o processo proposto.

A combinação da abordagem GQM+Strategies e da experimentação foi bem-sucedida, uma vez que foi possível disciplinar e alinhar o desenvolvimento da aplicação de DM ao planejamento estratégico da organização, baseando-se em método científico e em métricas que apoiam a implementação deste planejamento. No que diz respeito ao alinhamento estratégico em si e à aplicação do método científico, o fomento, enfoque e comunicação são beneficiados com o mapeamento de processos que transformem o alcance de metas estratégicas em requisitos explícitos, bem como a abordagem sistemática do método científico é facilitada.

É importante ressaltar que os processos de descoberta de conhecimento podem ter suas metas alteradas, uma vez que os processos de negócio apoiados por este processo de tecnologia podem não estar funcionando bem, de acordo com os seus indicadores de performance. Em outras palavras, as metas a serem atingidas pelas aplicações de DM podem ser diretamente técnicas, tais como, por exemplo, atingir um tempo mínimo para execução de uma consulta e exigência de acuracidade das informações, no entanto, estas metas precisam refletir também o alcance de metas do negócio, tais como, por exemplo, a conquista de novos clientes e diminuição da evasão. Se a aplicação inteligente atinge suas metas e o processo de negócio apoiado não, deve ser identificado onde está o problema (nível de software e/ou de negócio) e um replanejamento, com pacote de melhorias e novas metas, deve ser colocado em prática. Se o sistema indica que a Etnia do aluno é fator preponderante para evasão, quais as ações tomadas sob esse ponto de vista?.

Por fim, como trabalhos futuros, faz-se necessária a aplicação do processo em outras empresas, com tamanhos e complexidades diferentes, avaliando a adesão do processo por qualquer organização.

Uma vez apresentadas a proposta e a avaliação do processo, serão explanadas, no próximo capítulo, uma sintése narrativa da Revisão *Quasi*-Sistemática realizada e sugestões feitas para o processo.

5.0 DISCUSSÃO

Neste capítulo, será apresentada uma discussão dos resultados obtidos após a realização da Revisão *Quasi*-Sistemática, bem como comentários e sugestões feitos pelos avaliadores do processo. Desta forma, além da discussão sobre experimentação, estratégia e resultados brutos, realizada na Revisão da Literatura, uma discussão dos principais aspectos e lições aprendidas sobre possíveis melhorias no processo de desenvolvimento Experimental de aplicações de *Business Intelligence* e *Data Mining* é apresentada a seguir.

Antes do início da discussão, é importante revisitar o objetivo da revisão: "identificar e caracterizar as metodologias de desenvolvimento de aplicações de *Business Intelligence* e *Data Mining* dirigidas à estratégia e/ou que preveem avaliação experimental". Neste contexto, é importante ressaltar que não foi possível encontrar pelo menos uma abordagem que abrangesse o escopo requisitado, impedindo uma listagem destas metodologias, ou seja, a exposição de evidências sobre metodologias está limitada à escassez das publicações, as quais não existem ou estão limitadas pelo empirismo não publicado. Demonstrar essa lacuna também é um resultado considerável deste trabalho.

Como exceção limitada, podemos listar apenas o processo publicado por Colaço Júnior et al. (2019), o qual mescla a abordagem GQM+*Strategies* (vide Tabela 22) com uma metodologia de desenvolvimento ágil de aplicações de *Business Intelligence* proposta pelo autor, visando garantir o alinhamento estratégico. Apesar do avanço de disciplinar o alinhamento estratégico, o processo ainda não abrange soluções de Data Mining e não prevê uma validação experimental dos modelos de conhecimento, partes integrantes principais do objetivo deste trabalho.

Melhorias no processo de desenvolvimento de aplicações inteligentes são propostas desde a década de 50, quando o termo *Business Intelligence* foi utilizado pela primeira vez por Hans Peter Luhn, um pesquisador da IBM, no artigo intitulado "*A Business Intelligence System*" (LUHN, 1958). Nesse período, vários modelos de processo de descoberta de conhecimento foram propostos por pesquisadores e profissionais. Exemplos incluem: Fayyad, et al. (1996), Berry & Linoff (1997), Cabena et al. (1998), Cios et al. (2000), CRISP-DM (2003), IBM (2005), SAS (2005), Sharma, Osei-Bryson & Kasper (2012) e Ławrynowicz & Potoniec (2014).

Especificamente sobre o acoplamento de *Data Mining* às aplicações de BI, além da seminal KDD (*Knowledge Discovery in Databases*) (Fayaad et al., 1996), apesar de não terem sido desenvolvidas pela comunidade científica, CRISP-DM (Wirth & Hipp, 2000; Kurgan & Musilek, 2006) e SEMMA (Sample, Explore, Modify, Model, Assess) (Mariscal, Marban & Fernandez, 2020; Matignon, 2007) são amplamente usadas, todavia, também não apresentam

uma boa integração dos conceitos de BI e *Data Mining*, bem como não dão cobertura aos aspectos relacionados ao planejamento estratégico da organização. Uma alternativa, não encontrada nesta revisão para *Data Mining*, encontrada apenas para BI (Colaço et al., 2019), seria o uso de métodos auxiliares para o alinhamento estratégico, todavia, normalmente, o uso destes métodos é focado apenas em aspectos específicos ou estão relacionados a outros domínios da organização. Os métodos mais conhecidos da atualidade para alinhamentos de projetos de TI, tais como os de *Data Mining*, são apresentados na Tabela 22.

Tabela 22: Métodos Auxiliares para o Alinhamento Estratégico

Método	Descrição	Fonte
COBIT	O *Control Objectives for Information and Related* Technology (COBIT) é definido como um conjunto de diretrizes baseadas em auditoria para processos, práticas e controles de TIC voltadas à redução de riscos, busca pela integridade, confiabilidade e segurança da informação.	(Cobit, 2016)
BSC	*Balanced Scorecard* (BSC) é um método oriundo da governança corporativa, o qual exerce bem o papel de medição, mas não contempla boas práticas. Seus conceitos vêm sendo incorporados ao processo de plano estratégico de Tecnologia da Informação.	(Tonelli, et al., 2014)
GQM+Strategies	Abordagem sistemática que integra os objetivos de negócio, adaptando-os aos modelos de processos de software, produtos e perspectivas de interesse de qualidade, com base nas necessidades específicas do projeto.	(Basili, et al., 2014)

Tais abordagens fornecem métodos específicos que estão associados ao planejamento estratégico ou à tecnologia da informação, mas a publicação de pesquisas fazendo a mescla destas abordagens com aplicações inteligentes praticamente inexiste. Isto reforça os dados obtidos por Lima et al., (2017), em um *Survey* realizado no Brasil, o qual verificou que 72% das empresas não utilizavam um método específico para o desenvolvimento de aplicações de BI alinhadas ao planejamento estratégico da organização. Fato que pesquisadores têm apontado como uma das causas para o insucesso de projetos nessa área (Shi et al., 2010; Olszak, 2012).

Essa ausência de publicações pode indicar três coisas: (1) que a área já produziu resultados definitivos; (2) que a formalização de metodologias está falhando em alcançar níveis estratégicos; (3) ou que a área está se desenvolvendo na indústria, com informações não publicadas e tratadas como vantagem competitiva. O segundo ou terceiro caso são os mais prováveis, pois BI, *Data Mining*, IA e *Data Science* são tópicos muito importantes na

administração moderna, bem como esta é uma área de pesquisa ampla, com muito a oferecer às comunidades de pesquisa em Gestão Estratégica e em Ciência de Dados.

Sob esse ponto de vista da Ciência de Dados e da Experimentação, o atual contexto Big Data, em que os dados são gerados em velocidade, volume e variedade cada vez maiores, tem vetado o uso da maioria dos métodos estatísticos convencionais. Para gerenciar esses conjuntos de dados novos e potencialmente inestimáveis, novos métodos e novas aplicações na forma de análise preditiva estão sendo desenvolvidos sob a égide da Ciência de Dados.

Uma alternativa para atender os pressupostos da Ciência de Dados é padronizar os projetos de inteligência para o uso de uma abordagem experimental (Bock et al., 2018; Costa et al., 2015; Costa et al., 2016; Ławrynowicz & Potoniec, 2014; Santos et al., 2017; Sharma, Osei-Bryson & Kasper, 2012), uma vez que a aplicação de um método científico rigoroso coaduna com a tentativa de tornar a análise de dados uma ciência, com princípios que diminuem as ameaças à validade do conhecimento gerado.

No entanto, este trabalho evidenciou que ainda é tímido o uso do método científico no desenvolvimento de aplicações inteligentes. Apenas 28,57% dos trabalhos encontrados validaram suas soluções por meio de Experimentos Controlados, sendo a Aplicação Prática o método de pesquisa mais utilizado (47,61%). Esses números mostram a necessidade de aumentar o uso do método científico nessa área, com repetições de estudos que permitirão avaliar se outros pesquisadores chegarão, independentemente, aos mesmos resultados.

Além disso, mesmo os que validaram, não seguiram ou propuseram uma metodologia de BI ou de *Data Mining* dirigida à experimentação, ou seja, que prevê uma fase experimental na validação dos resultados.

Em resumo, qualquer abordagem, metodologia ou novo processo proposto deve se concentrar em atender os objetivos estratégicos organizacionais e validar sua utilidade em experimentos ou estudos de caso bem executados. Isso está longe de propor um novo modelo de conhecimento e de executar alguns estudos de viabilidade sobre este. Pesquisadores e profissionais devem se concentrar em responder a perguntas tais como: Minha abordagem pode ser usada no mundo real?; Como meus resultados se generalizam para outras organizações?. Caso contrário, sempre haverá um problema com a validade externa da abordagem proposta e será difícil passar do estado da arte para o estado da prática. Ainda nesse contexto, a colaboração entre a indústria e a academia é baixa. Atividades cooperativas levariam a uma melhoria mais rápida das abordagens existentes e a um entendimento mais profundo da área, com a produção de metodologias ou processos mais completos.

No que diz ao processo de extração dos dados, destaca-se a dificuldade para classificação das abordagens encontradas como alinhadas estrategicamente ou como experimentais. Mesmo com a possibilidade do uso da mineração inteligente de textos e da utilização de robôs baseados nesta tecnologia, muita intervenção humana faz-se necessária para definir um estudo como experimental e averiguar alinhamento estratégico. Foi preciso ler os textos inteiros e interpretar minunciosamente os dados, com algumas associações não automáticas, as quais teriam que ser implementadas em uma inovadora ferramenta de extração. Por exemplo, a presença de um objetivo estratégico não implica que a metodologia usada previu o uso deste objetivo ou atendeu a este.

Em relação ao estudo de caso, vale destacar um comentário feito, *ipsis litteris*, por um dos avaliadores:

> "A proposta se mostra bem inovadora, contudo, alguns desafios a sua implementação giram em torno de exigir uma cultura organizacional aprimorada, relacionando o Planejamento Estratégico da Organização e as aplicações de *Data Mining* e *Data Science*, sendo bastante desafiador alcançar esse nível de maturidade, principalmente no serviço público onde os processos costumam ser bem estáticos, não obstante, os artefatos propostos são coerentes principalmente em cenários onde pretende-se desenvolver essa nova cultura, por conseguinte, observam-se poucas abordagens desta temática na literatura".

Por fim, como sugestão, um dos entrevistados indicou a publicação de uma página na Web, onde seja possível a manutenção *online* dos artefatos gerados, bem como estabelecimento de uma comunidade para compartilhamento de experiências relacionadas ao uso do processo proposto.

6.0 CONCLUSÃO

Esta dissertação teve como objetivo propor e avaliar um processo para o desenvolvimento experimental de aplicações de *Data Mining* e *Data Science*, alinhadas ao planejamento estratégico da organização.

Como parte integrante desta dissertação, foi realizada uma Revisão *Quasi*-Sistemática da Literatura, com o objetivo de identificar e caracterizar as metodologias de desenvolvimento de aplicações de *Business Intelligence* e *Data Mining* dirigidas à estratégia e/ou que preveem avaliação Experimental. Por meio da revisão, não foi possível identificar trabalhos que apresentassem alguma abordagem para disciplinar o alinhamento estratégico no desenvolvimento desse tipo de aplicação.

Os estudos realizados serviram de base para a definição de um processo que pudesse disciplinar o alinhamento estratégico no desenvolvimento de aplicações de *Data Mining e Data Science*, por meio da combinação do GQM+*Strategies* e da aplicação do método científico.

Para avaliar o processo proposto, foi realizado um estudo de caso em uma instituição de ensino federal. Nesta fase, foram utilizadas técnicas para a avaliação qualitativa, por meio da aplicação de um questionário e de uma análise geral dos resultados obtidos.

Houve evidências de que um processo de BI dirigido à estratégia pode ser estendido para o desenvolvimento de aplicações de *Data Mining* e *Data Science* avaliadas experimentalmente, indagação central desta pesquisa, a qual foi respondida com base na aceitação inicial do processo por uma equipe de inteligência.

6.1 RESULTADOS E CONTRIBUIÇÕES

Dando prosseguimento ao tópico anterior, a principal contribuição deste estudo consiste na proposta e avaliação de um processo para o desenvolvimento experimental de aplicações de Data Mining (DM) e Data Science, alinhadas ao planejamento estratégico da organização. Além disso, destacam-se as seguintes contribuições:

- Processo para o Desenvolvimento de Aplicações de Business Intelligence Dirigido à Estratégia. Seus resultados foram submetidos e publicados no CONTECSI USP - *International Conference on Information Systems and Technology Management*;
- Revisão Quasi-Sistemática de metodologias de desenvolvimento de aplicações de *Business Intelligence* e *Data Mining* dirigidas à estratégia e/ou que preveem avaliação Experimental. Seus resultados foram submetidos à Revista Ibero-Americana de Estratégia – RIAE;

- Processo para o Desenvolvimento Experimental de Aplicações de *Data Mining e Data Science* Alinhadas ao Planejamento Estratégico da Organização. Seus resultados foram submetidos ao periódico JISTEM - *Journal of Information Systems and Technology Management*.

Na próxima seção, serão apresentados possíveis desdobramentos relacionados a este trabalho.

6.2 TRABALHOS FUTUROS

Para consolidar o processo proposto, é necessário aplicá-lo em outras empresas, com tamanhos e complexidades diferentes, avaliando a adesão do processo por qualquer organização.

Outro trabalho futuro é o desenvolvimento de uma ferramenta que gere os artefatos/documentos de cada atividade do processo, facilitando o seu preenchimento. Além disso, uma ferramenta desse porte pode ser incorporada à organização, agilizando as atividades preliminares.

REFERÊNCIAS

Alexander, A. (2014). Case studies: business intelligence. *Accounting Today,(June)*, 32.

Aradau, C., & Van Munster, R. (2011). *Politics of catastrophe: genealogies of the unknown*. Routledge.

Araújo, M. V. M., & Dornelas, J. S. (2017). Mapeamento perceptual da associação entre sucesso de projetos de TI e fatores promotores do alinhamento estratégico. *EnANPAD*.

Astley, W. G., Axelsson, R., Butler, R. J., Hickson, D. J., & Wilson, D. C. (2017). Complexity and cleavage: dual explanations of strategic decision-making. *In The Bradford studies of strategic decision making* (pp. 47-65). *Ashgate*.

Barbieri, C. (2011). *BI2--Business intelligence: Modelagem & Qualidade*. Elsevier Editora.

Basili, V. R. (1996, March). The role of experimentation in software engineering: past, current, and future. In *Proceedings of IEEE 18th International Conference on Software Engineering* (pp. 442-449). IEEE.

Basili, V., Heidrich, J., Lindvall, M., Munch, J., Regardie, M., & Trendowicz, A. (2007, September). GQM+Strategies – Aligning Business Strategies with Software

Measurement. In *First international symposium on empirical software engineering and measurement (ESEM 2007)* (pp. 488-490). IEEE.

Basili, V. R., Lindvall, M., Regardie, M., Seaman, C., Heidrich, J., Münch, J., ... & Trendowicz, A. (2010). Linking software development and business strategy through measurement. *Computer*, *43*(4), 57-65.

Basili, V., Trendowicz, M. Kowalczyk, J. Heidrich, C. Seaman, J. Münch, D. Rombach. (2014). *Aligning Organizations Through Measurement: The GQM+Strategies Approach*. *Springer Publishing Company, Incorporated.*

Batista, C. F., Souza, E. P. R., Correia Neto, J. D. S., & Dornelas, J. S. (2012). Proposta de Data Mart para Análise de Faturamento de Empresa de Varejo Utilizando Software Livre. *Revista Brasileira de Administração Científica*, 3(2).

Bautista, R. M. (2018). Clustering failed courses of engineering students using association rule mining. *Journal of Theoretical & Applied Information Technology*, v. 96(4).

Bergin, S., & Wraight, P. (2006). Silver based wound dressings and topical agents for treating diabetic foot ulcers. *Cochrane Database of Systematic Reviews*, (1).

Berry, M. J., & Linoff, G. S. (2004). *Data mining techniques: for marketing, sales, and customer relationship management*. John Wiley & Sons.

Brannon, N. (2010). Business intelligence and E-discovery. *Intellectual Property & Technology Law Journal*, 22(7), 1.

Bock, C., Gumbsch, T., Moor, M., Rieck, B., Roqueiro, D., & Borgwardt, K. (2018). Association mapping in biomedical time series via statistically significant shapelet mining. *Bioinformatics*, *34*(13), i438-i446.

Bologa, A., & Bologa, R. (2011). Business intelligence using software agents. *Database Systems Journal*, 2(4), 31-42.

Bosch-Sijtsema, P., & Bosch, J. (2015). User involvement throughout the innovation process in high-tech industries. *Journal of Product Innovation Management*, *32*(5), 793-807.

Botelho, F. R., & Filho, E. R. (2014). Conceituando o termo Business Intelligence: Origem e Principais Objetivos. *Sistemas, Cibernética e Informática*, vol. 11, n.º 11, pp. 55–60.

Brackett, M. H. (1996). *The data warehouse challenge: taming data chaos*. John Wiley & Sons, Inc..

Cabena, P., Hadjinian, P., Stadler, R., Verhees, J., & Zanasi, A. (1998). *Discovering data mining: from concept to implementation*. Prentice-Hall, Inc.

Campbell, B. R. (2005). Alignment: Resolving ambiguity within bounded choices. In *Pacific Asia Conference on Information Systems*. University of Hong King.

Castellion, G. (2008). Do it wrong quickly: how the web changes the old marketing rules by Mike Moran. *Journal of Product Innovation Management*, v. 25, n. 6, p. 633-635.

Chan, Y. E., Huff, S. L., Barclay, D. W., & Copeland, D. G. (1997). Business strategic orientation, information systems strategic orientation, and strategic alignment. *Information systems research*, 8(2), 125-150.

Chaudhuri, S., Dayal, U., & Narasayya, V. (2011). An overview of business intelligence technology. *Communications of the ACM*, 54(8), 88-98.

Chen, M. S., Han, J., & Yu, P. S. (1996). Data mining: an overview from a database perspective. *IEEE Transactions on Knowledge and data Engineering*, 8(6), 866-883.

Cheng, H., Lu, Y. C., & Sheu, C. (2009). An ontology-based business intelligence application in a financial knowledge management system. *Expert Systems with Applications*, *36*(2), 3614-3622.

Cielen, D., & Meysman, A. (2016). *Introducing data science: big data, machine learning, and more, using Python tools*. Simon and Schuster.

Cios, K. J., Teresinska, A., Konieczna, S., Potocka, J., & Sharma, S. (2000). Diagnosing myocardial perfusion from PECT bull's-eye maps-A knowledge discovery approach. *IEEE Engineering in Medicine and Biology Magazine*, *19*(4), 17-25.

Cobit. (2016). *What is Cobit 5? It's the leading framework for the governance and management of enterprise IT. Information Systems Audit and Control Foundation* (ISACA). [Online] 20 de Junho de 2019. http://www.isaca.org/COBIT/Pages/default.aspx.

Clancy, T. (1995). The standish group report. *Chaos report*.

Colaço Júnior, M., de Fátima Menezes, M., Corumba, D., Mendonça, M., & Santos, B. S. (2015). Do software engineers have preferred representational systems?. *Journal of Research and Practice in Information Technology*, 47(1), 23-46.

Colaço Júnior, M. (2018). *Vocabulário e Definição de Estudos Experimentais* [Material da Disciplina de Engenharia de Software Experimental]. Mestrado em Ciência da

Computação, Universidade Federal de Sergipe, São Cristóvão, Sergipe.

Colaço Júnior, M., Cruz, R. F. & Lima, A. S. (2019). Proposta e Avaliação de um Processo para o Desenvolvimento de Aplicações de Business Intelligence Dirigido à Estratégia. In: *International Conference on Information Systems and Technology Management*, 2019, São Paulo. ContecSI.

Côrte-Real, N., Oliveira, T., & Ruivo, P. (2017). Assessing business value of Big Data Analytics in European firms. *Journal of Business Research*, *70*, 379-390.

Cortez, P., & Santos, M. F. (2013). *Knowledge discovery and business intelligence*, v. 30, n. 4, p. 283-284.

Costa, A. D. S., Nascimento, A. V. D., Cruz, E. B., Terra, L. L., & Ramalho, M. (2013). *O uso do método estudo de caso na Ciência da Informação no Brasil* (Vol. 4, 1).

Costa, J. K. G., Santos, I. P. O., Nascimento, A. V. R., & Júnior, M. C. (2015, May). Experimentation at Industrial Setting to Improve the Effectiveness of the ETL Procedures Implementation in a Business Intelligence Environment. In: *Proceedings of the annual conference on Brazilian Symposium on Information Systems: Information Systems: A Computer Socio-Technical Perspective-Volume 1. Brazilian Computer Society* (pp. 459-466).

Costa, J. K., Santos, I. P., junior, M. C., & Nascimento, A. V. (2016, May). An Experiment in an Industrial Business Intelligence environment to improve data loads maintenance. In *Proceedings of the XII Brazilian Symposium on Information Systems on Brazilian Symposium on Information Systems: Information Systems in the Cloud Computing Era-Volume 1* (pp. 534-541).

Costa, S. C. M., de Mattos Pimenta, C. A., & Nobre, M. R. C. (2007). A estratégia PICO para a construção da pergunta de pesquisa e busca de evidências. *Revista Latino-Americana de Enfermagem*, 15(3).

Covões, T. F. (2010). *Seleção de atributos via agrupamento* (Doctoral dissertation, Universidade de São Paulo).

CRISP-DM. (2003). *Cross Industry Standard Process for Data Mining 1.0: Step by Step Data Mining Guide.* [Online] 20 de Junho de 2019. http://www.crisp-dm.org/.

Cruz, R. F., Colaço Júnior, Methanias & GOIS, V. M. (2021). Quão experimentais e estratégicas são as aplicações de Business Intelligence (BI), Data Mining e Inteligência

Artificial (IA)? Em fase de revisão pela *Revista Ibero-Americana de Estratégia*.

Davenport, T. H., & Harris, J. G. (2007). Competing on analytics: Harvard Business School Press. *Metrics and analytics in SEM. Whitepaper*. Retrieved on June, 19, 2009.

Dedić, N., & Stanier, C. (2016). *An evaluation of the challenges of multilingualism in data warehouse development*.

Dedić, N., & Stanier, C. (2017). Measuring the success of changes to Business Intelligence solutions to improve Business Intelligence reporting. *Journal of Management Analytics*, 4(2), 130-144.

Deloitte. (2017). *Trajetória entre perfis: No rumo da geração de valor ao negócio*. [Online] 10 de Julho de 2019. https://www2.deloitte.com/br/pt/pages/technology/articles/cio-survey.html.

Demo, P. A. e Silva, R. (2012). *Pesquisa e Informação Qualitativa* 5ª edição. . São Paulo :s.n..

Disner, D. D. S. (2015). *Mineração de dados para obtenção de conhecimento em Big data*.

Dittrich, Y., Nørbjerg, J., Tell, P., & Bendix, L. (2018, May). Researching cooperation and communication in continuous software engineering. In *2018 IEEE/ACM 11th International Workshop on Cooperative and Human Aspects of Software Engineering (CHASE)* (pp. 87-90). IEEE.

Duan, L., & Da Xu, L. (2012). Business intelligence for enterprise systems: A survey. *IEEE Transactions on Industrial Informatics*, 8(3), 679-687.

Endres, A., & Rombach, H. D. (2003). *A handbook of software and systems engineering: Empirical observations, laws, and theories*. Pearson Education.

Esfandiari, N., Babavalian, M. R., Moghadam, A. M. E., & Tabar, V. K. (2014). Knowledge discovery in medicine: Current issue and future trend. *Expert Systems with Applications*, 41(9), 4434-4463.

Fagerholm, F., Guinea, A. S., Mäenpää, H., & Münch, J. (2017). The RIGHT model for continuous experimentation. *Journal of Systems and Software*, *123*, 292-305.

Falessi, D., Juristo, N., Wohlin, C., Turhan, B., Münch, J., Jedlitschka, A., & Oivo, M. (2018). Empirical software engineering experts on the use of students and professionals in experiments. *Empirical Software Engineering*, 23(1), 452-489.

Farias, M. A., Xisto, R., Santos, M. S., Fontes, R. S., Colaço, M., Spínola, R., & Mendonça, M.

(2019, May). Identifying technical debt through a code comment mining tool. In *Proceedings of the XV Brazilian Symposium on Information Systems* (pp. 1-8).

Fayyad, U. M., Piatetsky-Shapiro, G., & Smyth, P. (1996, August). Knowledge Discovery and Data Mining: Towards a Unifying Framework. In *KDD* (Vol. 96, pp. 82-88).

Gain, U., & Hotti, V. (2021, February). Low-code AutoML-augmented Data Pipeline–A Review and Experiments. In *Journal of Physics: Conference Series* (Vol. 1828, No. 1, p. 012015). IOP Publishing.

Galvão, N. D., & Marin, H. D. F. (2009). Técnica de mineração de dados: uma revisão da literatura. *Acta Paulista de Enfermagem*, 22(5), 686-690.

Goldratt, E. M., & Cox, J. (2016). *The goal: a process of ongoing improvement*. Routledge.

Goldschmidt, R., & Passos, E. (2005). *Data mining: um guia prático*. Gulf Professional Publishing.

Grover, V., Chiang, R. H., Liang, T. P., & Zhang, D. (2018). Creating strategic business value from big data analytics: A research framework. *Journal of Management Information Systems*, 35(2), 388-423.

Hall, A. L., & Rist, R. C. (1999). Integrating multiple qualitative research methods (or avoiding the precariousness of a one-legged stool). *Psychology & Marketing*, 16(4), 291-304.

Han, R., Nie, L., Ghanem, M. M., & Guo, Y. (2013, October). Elastic algorithms for guaranteeing quality monotonicity in big data mining. In *2013 IEEE International Conference on Big Data* (pp. 45-50). IEEE.

Hans, R. T., & Mnkandla, E. (2013, September). Modeling software engineering projects as a business: A business intelligence perspective. In *2013 Africon* (pp. 1-5). IEEE.

Henry, R., & Venkatraman, S. (2015). Big Data Analytics the Next Big Learning Opportunity. *Journal of Management Information & Decision Sciences*, 18(2).

Hohnhold, H., O'Brien, D., & Tang, D. (2015, August). Focusing on the long-term: It's good for users and business. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1849-1858).

IBM. (2005). *Analytics solutions unified method*. ftp://ftp.software.ibm.com/software/data/sw-library/services/ASUM.pdf.

Inmon, W. H. (1997). *Como construir o Data Warehouse. Rio de Janeiro: Campus*.

Isaca. (2018). *COBIT® 2019 Framework: Introduction & Methodology. Information Systems Audit and Control Foundation* (ISACA).

Ju, J., Liu, L., & Feng, Y. (2018). Citizen-centered big data analysis-driven governance intelligence framework for smart cities. *Telecommunications Policy*, 42(10), 881-896.

Juristo, N., & Moreno, A. M. (2013). *Basics of software engineering experimentation*. Springer Science & Business Media.

Kanavos, A., Nodarakis, N., Sioutas, S., Tsakalidis, A., Tsolis, D., & Tzimas, G. (2017). Large scale implementations for twitter sentiment classification. *Algorithms*, 10(1), 33.

King, W. R. (1988). How effective is your information systems planning?. *Long range planning*, 21(5), 103-112.

Kitchenham, B. (2004). Procedures for performing systematic reviews. *Keele, UK, Keele University*, *33*(2004), 1-26.

Kitchenham, B., Brereton, O. P., Budgen, D., Turner, M., Bailey, J., & Linkman, S. (2009). Systematic literature reviews in software engineering–a systematic literature review. *Information and software technology*, 51(1), 7-15.

Kluger, A. N., & Tikochinsky, J. (2001). The error of accepting the" theoretical" null hypothesis: the rise, fall, and resurrection of commonsense hypotheses in psychology. *Psychological bulletin*, *127*(3), 408.

Kohavi, R., Longbotham, R., Sommerfield, D., & Henne, R. M. (2009). Controlled experiments on the web: survey and practical guide. *Data mining and knowledge discovery*, *18*(1), 140-181.

Kohavi, R., Deng, A., Frasca, B., Walker, T., Xu, Y., & Pohlmann, N. (2013, August). Online controlled experiments at large scale. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 1168-1176).

Kohavi, R., Deng, A., Longbotham, R., & Xu, Y. (2014, August). Seven rules of thumb for web site experimenters. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 1857-1866).

Kohavi, R., & Longbotham, R. (2017). Online Controlled Experiments and A/B Testing. *Encyclopedia of machine learning and data mining*, *7*(8), 922-929.

Kohtamäki, M., & Farmer, D. (2017). Strategic Agility—Integrating Business Intelligence with

Strategy. In *Real-time Strategy and Business Intelligence* (pp. 11-36). Palgrave Macmillan, Cham.

Koua, E. L., & Kraak, M. J. (2004). Geovisualization to support the exploration of large health and demographic survey data. *International journal of health geographics*, 3(1), 1-13.

Kubina, M., Varmus, M., & Kubinova, I. (2015). Use of big data for competitive advantage of company. *Procedia Economics and Finance*, *26*, 561-565.

Kurgan, L. A., & Musilek, P. (2006). A survey of knowledge discovery and data mining process models. *Knowledge Engineering Review*, *21*(1), 1-24.

Ławrynowicz, A., & Potoniec, J. (2014). Pattern based feature construction in semantic data mining. *International Journal on Semantic Web and Information Systems (IJSWIS)*, *10*(1), 27-65.

Lima, Adriano, Colaço Júnior, Methanias & Nascimento, Andre Vinicius RP. (2017). Um Survey com Empresas Brasileiras acerca da Utilização de Business Intelligence (BI) e um diagnóstico sobre a infraestrutura e metodologias associadas. *Conferência Ibero-Americana de Engenharia de Software* – Trilha de Engenharia de Software Experimental.

Lin, Y. F., Huang, C. F., & Tseng, V. S. (2017). A novel methodology for stock investment using high utility episode mining and genetic algorithm. *Applied Soft Computing*, 59, 303-315.

Ma, L., & Fan, S. (2017). CURE-SMOTE algorithm and hybrid algorithm for feature selection and parameter optimization based on random forests. *BMC bioinformatics*, 18(1), 1-18.

Maione, C. (2020). *Balanceamento de dados com base em oversampling em dados transformados*. 2020. 135 f. Tese (Doutorado em Ciência da Computação em Rede) - Universidade Federal de Goiás, Goiânia.

Mandić, V., Basili, V., Harjumaa, L., Oivo, M., & Markkula, J. (2010, September). Utilizing GQM+ Strategies for business value analysis: An approach for evaluating business goals. In *Proceedings of the 2010 ACM-IEEE International Symposium on Empirical Software Engineering and Measurement* (pp. 1-10).

Manigandan, E., Shanthi, V., & Kasthuri, M. (2019). Parallel clustering for data mining in CRM. In *Data Management, Analytics and Innovation* (pp. 117-127). Springer, Singapore.

Manzi, J. (2019). *Uncontrolled: The surprising payoff of trial-and-error for business, politics,*

and society. Basic Books (AZ).

Mariscal, G., Marban, O., & Fernandez, C. (2010). A survey of data mining and knowledge discovery process models and methodologies. *The Knowledge Engineering Review*, 25(2), 137.

Martin, R. C. (2002). *Agile software development: principles, patterns, and practices*. Prentice Hall.

Martínez-Plumed, F., Contreras-Ochando, L., Ferri, C., Orallo, J. H., Kull, M., Lachiche, N., ... & Flach, P. A. (2019). CRISP-DM twenty years later: From data mining processes to data science trajectories. *IEEE Transactions on Knowledge and Data Engineering*.

Matignon, R. (2007). *Data mining using SAS enterprise miner* (Vol. 638). John Wiley & Sons.

Medeiros Júnior, J. V., de Sousa Neto, M. V., Añez, M. E. M., & de Moraes, E. A. (2017). Identifying mechanisms to develop information technology capabilities. *Revista Ibero-Americana de Estratégia*, 16(4), 37-49.

Mola, L., Rossignoli, C., Carugati, A., & Giangreco, A. (2015). Business intelligence system design and its consequences for knowledge sharing, collaboration, and decision-making: an exploratory study. *International Journal of Technology and Human Interaction (IJTHI)*, 11(4), 1-25.

Monino, J. L., & Sedkaoui, S. (2016). *Big Data, Open Data and Data Development*, vol. 3.

More, S. (2014, May). Modified path traversal for an efficient web navigation mining. In *2014 IEEE International Conference on Advanced Communications, Control and Computing Technologies* (pp. 940-945). IEEE.

Münch, J., Fagerholm, F., Kettunen, P., Pagels, M., & Partanen, J. (2013). The effects of GQM+ Strategies on organizational alignment. *arXiv preprint arXiv:1311.6221*.

Nunes, F. M., Júnior, M. C., da Silva Junior, J. B., Costa, L. B. B., & Recchi, E. C. S. (2019, May). Galactus-Um ambiente inteligente para apoio à tomada de decisão no âmbito do Ministério Público de Sergipe. In *Anais Estendidos do XV Simpósio Brasileiro de Sistemas de Informação* (pp. 153-156). SBC.

Obeidat, M., North, M., Richardson, R., & Rattanak, V. (2015). *Business intelligence technology, applications, and trends*.

Olsson, H. H., Alahyari, H., & Bosch, J. (2012, September). Climbing the" Stairway to

Heaven"--A Mulitiple-Case Study Exploring Barriers in the Transition from Agile Development towards Continuous Deployment of Software. In *2012 38th euromicro conference on software engineering and advanced applications* (pp. 392-399). IEEE.

Olsson, H. H., & Bosch, J. (2014). The HYPEX model: from opinions to data-driven software development. In *Continuous software engineering* (pp. 155-164). Springer, Cham.

Olszak, C. M., & Ziemba, E. (2012). Critical success factors for implementing business intelligence systems in small and medium enterprises on the example of upper Silesia, Poland. *Interdisciplinary Journal of Information, Knowledge, and Management*, 7(2), 129-150.

Orne, M. T. (1962). *Sobre a psicologia social da experiência psicológica: Com referência particular para exigir características e suas implicações*.

Petersen, K., Feldt, R., Mujtaba, S., & Mattsson, M. (2008, June). Systematic mapping studies in software engineering. In *12th International Conference on Evaluation and Assessment in Software Engineering (EASE)* 12 (pp. 1-10).

Pinto, P. (2015). *Introdução à Análise Estatística-Vol 2* (Vol. 2). Sílabas & Desafios.

Phillips-Wren, G., & Hoskisson, A. (2015). An analytical journey towards big data. *Journal of Decision Systems*, 24(1), 87-102.

Primak, F. V. (2008). *Decisões com BI (Business Intelligence).* Fabio Vinicius Primak.

Puppala, M., He, T., Chen, S., Ogunti, R., Yu, X., Li, F., ... & Wong, S. T. (2015). METEOR: an enterprise health informatics environment to support evidence-based medicine. *IEEE Transactions on Biomedical Engineering*, 62(12), 2776-2786.

Rainer, R. K., & Cegielski, C. G. (2011). *Introdução a sistemas de informação* (3rd ed.). Rio de Janeiro: Elsevier. 2011.

Reich, B. H., & Benbasat, I. (1996). Measuring the linkage between business and information technology objectives. *MIS quarterly*, 55-81.

Rodríguez, P., Haghighatkhah, A., Lwakatare, L. E., Teppola, S., Suomalainen, T., Eskeli, J., ... & Oivo, M. (2017). Continuous deployment of software intensive products and services: A systematic mapping study. *Journal of Systems and Software*, *123*, 263-291.

Rogalewicz, M., & Sika, R. (2016). Methodologies of knowledge discovery from data and data mining methods in mechanical engineering. *Management and Production Engineering*

Review.

Rosemary Williams DBA, C. P. A. (2015). Using data analytics for oversight and efficiency. *The journal of government financial management*, 64(2), 18.

Roy, R. K. (2001). *Design of experiments using the Taguchi approach: 16 steps to product and process improvement*. John Wiley & Sons.

Ruggieri, S., Pedreschi, D., & Turini, F. (2010). Integrating induction and deduction for finding evidence of discrimination. *Artificial Intelligence and Law*, 18(1), 1-43.

Santos, I. P. O., Costa, J. K. G., Júnior, M. C., & Nascimento, A. V. R. (2017, April). Experimental Evaluation of Automatic Tests Cases in Data Analytics Applications Loading Procedures. In *ICEIS (1)* (pp. 304-311).

Santos, B. S., Junior, M. C., & de Souza, J. G. (2018, June). An Experimental Evaluation of the NeuroMessenger: A Collaborative Tool to Improve the Empathy of Text Interactions. In *2018 IEEE Symposium on Computers and Communications (ISCC)* (pp. 00573-00579). IEEE.

Santos, A. C. M., Colaço Junior, Methanias, & de Carvalho Andrade, E. (2020). Multimedia resources as a support for requirements engineering and software maintenance. In *Journal of Software: Evolution and Process*.

SAS. (2005). *Semma data mining methodology*. http://www.sas.com/technologies/analytics/datamining/miner/semma.html.

Saunders, M., Lewis, P., & Thornhill, A. (2009). *Research methods for business students*. Pearson education.

Schäfer, F., Zeiselmair, C., Becker, J., & Otten, H. (2018, November). Synthesizing CRISP-DM and quality management: A data mining approach for production processes. In *2018 IEEE International Conference on Technology Management, Operations and Decisions (ICTMOD)* (pp. 190-195). IEEE.

Sedkaoui, S. (2018). Statistical and Computational Needs for Big Data Challenges. In *Big Data Analytics in HIV/AIDS Research* (pp. 21-53). IGI Global.

Sell, D. & Pacheco, R. C. S. (2001). Uma arquitetura para distribuição de componentes tecnológicos de sistemas de informações baseados em Data Warehouse. In: *XXI Encontro Nacional De Engenharia De Produção - ENEGEP*, 2001, Salvador. Rio de Janeiro:

ABEPRO.

Sjøberg, D. I., Hannay, J. E., Hansen, O., Kampenes, V. B., Karahasanovic, A., Liborg, N. K., & Rekdal, A. C. (2005). A survey of controlled experiments in software engineering. *IEEE transactions on software engineering*, *31*(9), 733-753.

Sharma, S., Osei-Bryson, K. M., & Kasper, G. M. (2012). Evaluation of an integrated Knowledge Discovery and Data Mining process model. In *Expert Systems with Applications*, *39*(13), 11335-11348.

Shi, Y., & Lu, X. (2010, November). The role of business intelligence in business performance management. In *2010 3rd International Conference on Information Management, Innovation Management and Industrial Engineering* (Vol. 4, pp. 184-186). IEEE.

Shmueli, G., Bruce, P. C., Yahav, I., Patel, N. R., & Lichtendahl Jr, K. C. (2017). *Data mining for business analytics: concepts, techniques, and applications in R. John Wiley & Sons*.

Sholom, M. W., & Indurkhya, N. (1999). *Predict Data Mining*. Morgan Kaufmann Publishes, Inc.

Singh, B. (2016). *The Lean Startup: How Today's Entrepreneurs Use Continuous Innovation to Create Radically Successful Businesses*. *Bangalore* Vol. 11, Ed. 2.

Sowmya, R., & Suneetha, K. R. (2017, January). Data mining with big data. In *2017 11th International Conference on Intelligent Systems and Control (ISCO)* (pp. 246-250). IEEE.

Sun, Y., Bauer, B., & Weidlich, M. (2017, November). Compound trace clustering to generate accurate and simple sub-process models. In *International Conference on Service-Oriented Computing* (pp. 175-190). Springer, Cham.

Thamir, A., & Poulis, E. (2015). Business intelligence capabilities and implementation strategies. *International Journal of Global Business*, 8(1), 34.

Tonelli, A. O., Bermejo, P. H. D. S., & Zambalde, A. L. (2014). Using the bsc for strategic planning of it (information technology) in brazilian organizations. *JISTEM-Journal of Information Systems and Technology Management*, 11(2), 361-378.

Trendowicz, A., Heidrich, J., & Shintani, K. (2011, November). Aligning software projects with business objectives. In *2011 Joint Conference of the 21st International Workshop on Software Measurement and the 6th International Conference on Software Process and Product Measurement* (pp. 142-150). IEEE.

Trninic, J., Durkovic, J., & Rakovic, L. (2011). Business intelligence as support to knowledge management. *Perspectives of Innovations, Economics, and Business*, 8(1231-2016-100753), 35-40.

Tuler, E., Prates, R. O., Almir, F., Rocha, L., & Meira Jr, W. (2006, November). Caracterizando desafios de interação com sistemas de mineração de regras de associação. In *Proceedings of VII Brazilian symposium on Human factors in computing systems* (pp. 40-49).

Turban, E., Sharda, R., Aronson, J. E., & King, D. (2009). *Business intelligence: um enfoque gerencial para a inteligência do negócio*. Bookman Editora.

Turban, E., & Volonino, L. (2013). *Tecnologia da Informação para Gestão-: Em Busca de um Melhor Desempenho Estratégico e Operacional*. Bookman Editora.

Vasconcelos, N., Júnior, M. C., Almeida, T., & da Silva, V. M. (2019). Comparative Analysis of Data Mining Algorithms Applied to the Context of School Dropout. In *FedCSIS (Communication Papers)* (pp. 3-10).

Vercellis, C. (2009). *Business intelligence: data mining and optimization for decision making* (pp. 1-18). New York: Wiley.

Vergara, S. C. (2006). *Projetos e relatórios de pesquisa*. São Paulo: Atlas.

Vitt, C. A., & Xiong, H. (2015, November). The impact of patent activities on stock dynamics in the high-tech sector. In *2015 IEEE International Conference on Data Mining* (pp. 399-408). IEEE.

Yin, R. (2015). *Estudo de Caso - 5.Ed.: Planejamento e Métodos*. s.l. : BOOKMAN.

Yu, L., Zheng, J., Shen, W. C., Wu, B., Wang, B., Qian, L., & Zhang, B. R. (2012, August). BC-PDM: data mining, social network analysis and text mining system based on cloud computing. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 1496-1499).

Wang, X., & Sun, Z. (2013, November). The design of water resources and hydropower cloud GIS platform based on big data. In *International Conference on Geo-Informatics in Resource Management and Sustainable Ecosystem* (pp. 313-322). Springer, Berlin, Heidelberg.

Weber, M., & Klein, A. Z. (2013). Gestão estratégica em empresas de tecnologia da informação: um estudo de caso. *Revista Ibero Americana de Estratégia*, 12(3), 37-65.

Wirth, R., & Hipp, J. (2000, April). CRISP-DM: Towards a standard process model for data mining. In *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining* (Vol. 1). London, UK: Springer-Verlag.

Wohlin, C., Runeson, P., Höst, M., Ohlsson, M. C., Regnell, B., & Wesslén, A. (2012). *Experimentation in software engineering*. Springer Science & Business Media.

APÊNDICE

Formulário de Avaliação do Processo Proposto

*Obrigatório

Você tem quantos anos de experiência na sua área de atuação? *
( ) Até 3 anos
( ) Acima de 3 a 5 anos
( ) Acima de 5 a 10 anos
( ) Acima de 10 anos

Qual a sua posição?
( ) DBA
( ) Administrador de Dados
( ) Analista
( ) Programador
( ) Gerente
Outro:

Na sua empresa é utilizada alguma metodologia formal no desenvolvimento de projetos de Mineração de Dados? (Mais de uma opção pode ser marcada) *
[ ] CRISP
[ ] SEMMA
[ ] ASUM-DM
[ ] KDD
[ ] Nenhuma
Outro:

Na sua opinião, o Alinhamento Estratégico é importante para o desenvolvimento de Projetos de Mineração de Dados? *
( ) Concordo Plenamente
( ) Concordo
( ) De Forma Regular
( ) Discordo
( ) Discordo Plenamente

Na sua empresa, é utilizada alguma metodologia para o alinhamento estratégico no desenvolvimento de projetos de Mineração de Dados? (Mais de uma opção pode ser marcada) *
[ ] GQM
[ ] GQM+Strategies
[ ] Balanced Scorecard (BSC)
[ ] COBIT
[ ] Nenhuma
Outro:

Na sua opinião, o processo proposto possibilita vincular o projeto de Mineração de Dados ao planejamento estratégico da empresa? *
( ) Concordo Plenamente
( ) Concordo
( ) De Forma Regular
( ) Discordo
( ) Discordo Plenamente

Você já participou de algum projeto experimental? *
( ) Sim
( ) Não

Na sua opinião, a experimentação pode ser usada para orientar o desenvolvimento de produtos e permitir que a organização avalie o ROI dos projetos no desenvolvimento de software? *
( ) Concordo Plenamente
( ) Concordo
( ) De Forma Regular
( ) Discordo
( ) Discordo Plenamente

Na sua opinião, a experimentação empregada no processo foi eficaz para avaliar os algoritmos que irão compor o modelo de mineração? *
( ) Concordo Plenamente
( ) Concordo
( ) De Forma Regular
( ) Discordo
( ) Discordo Plenamente

Na sua opinião, o processo proposto ajuda o profissional a seguir o método experimental, sem negligenciar nenhuma fase?
( ) Concordo Plenamente
( ) Concordo
( ) De Forma Regular
( ) Discordo
( ) Discordo Plenamente

No geral, como você avalia o processo proposto para o Desenvolvimento Experimental de Aplicações de Data Mining e Data Science Alinhadas ao Planejamento Estratégico da Organização? (Opcional)
_____
_____
_____
_____

Reserve esse espaço para dar alguma sugestão (Opcional)
_____
_____

_____
_____