UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ

ALEX FRANKLIN DA SILVA VALE

FRAMEWORK PARA CIÊNCIA DE DADOS NO CONTEXTO DE PEQUENAS E MÉDIAS EMPRESAS BRASILEIRAS

PATO BRANCO

2023

ALEX FRANKLIN DA SILVA VALE

FRAMEWORK PARA CIÊNCIA DE DADOS NO CONTEXTO DE PEQUENAS E MÉDIAS EMPRESAS BRASILEIRAS

FRAMEWORK FOR DATA SCIENCE IN THE CONTEXT OF SMALL AND MEDIUM-SIZED BRAZILIAN ENTERPRISES

Dissertação apresentada ao Programa de Pós-Graduação em Engenharia de Produção e Sistemas da Universidade Tecnológica Federal do Paraná, *Campus* Pato Branco, como requisito parcial à obtenção do título de Mestre em Engenharia de Produção - Área de Concentração: Gestão de Operações.

Orientador: Prof. Dr. Gilson Adamczuk Oliveira

Coorientador: Prof. Dr. Érick Oliveira Rodrigues

PATO BRANCO

2023

ALEX FRANKLIN DA SILVA VALE

FRAMEWORK PARA CIÊNCIA DE DADOS NO CONTEXTO DE PEQUENAS E MÉDIAS EMPRESAS BRASILEIRAS

Trabalho de pesquisa de mestrado apresentado como requisito para obtenção do título de Mestre Em Engenharia De Produção E Sistemas da Universidade Tecnológica Federal do Paraná (UTFPR). Área de concentração: Gestão Dos Sistemas Produtivos.

Data de aprovação: 31 de Julho de 2023

Dr. Gilson Adamczuk Oliveira, Doutorado - Universidade Tecnológica Federal do Paraná

Dr. Edson Pinheiro De Lima, Doutorado - Universidade Tecnológica Federal do Paraná

Dr. Ricardo Da Silva Braga, Doutorado - Petsupermarket Comercio de Produtos para Animais S/A

À Liandra Franklin, e a ela somente.

AGRADECIMENTOS

Agradeço primeiramente à minha esposa Liandra Franklin por me apoiar e estar ao meu lado nos momentos mais difíceis dessa jornada.

Agradecimentos ao meu orientador Prof. Dr. Gilson Adamczuk Oliveira por apontar o caminho, pela amizade, pela sinceridade e principalmente pelo incentivo à liberdade de criação.

Ao meu coorientador Prof. Dr. Érick Oliveira Rodrigues pelas regras de ouro no campo da pesquisa e pelas conversas esclarecedoras sobre *Machine Learning*.

Grato à Marisa Nilson por conceder de forma generosa o recurso mais escasso que temos: o tempo.

Agradeço à Eduarda Dutra por me ajudar a montar o projeto que me fez ser aceito no Programa de Pós-graduação em Engenharia de Produção da UTFPR.

Grato a meus colegas de curso por todo o compartilhamento durante esse período, em especial a Maiara Cristina Feliceti, Eloisa da Silva Garais, Arthur Facin de Bortoli e João Mazzochin pela torcida e por tornar esse passeio um lugar divertido.

Grato a UTFPR pelo ensino público e de altíssima qualidade, ao corpo docente por manter em altíssimo nível o padrão de formação acadêmica e um agradecimento especial também à Adriani Edith Michelon por ser um farol na vida de todos os alunos.

Meus sinceros agradecimentos a meus colegas e amigos pelas valiosas contribuições. Leandro Benitez, Ricardo Tanaka, Diego Peres, Kesley Miranda, Cesar Domingues, Marcos Bergonzi, Ray Lacerda, Bruno Curtarelli, Pedro Nishi, Márcio Marques, Valmir Costa, Leonardo Capellaro, William Calixto, Giuseppe Giovanelli, André Cotta, Rodolfo Real, Guilherme da Rosa, Diogo Couto e Marcello Tedardi, vocês fizeram total diferença.

Sinceros agradecimentos aos membros da banca, Prof. Dr. Edson Pinheiro e Dr. Ricardo da Silva Braga. Poder apresentar-lhes uma ideia é sem dúvida um privilégio.

“A educação é a arma mais poderosa que você pode usar para mudar o mundo” (Nelson Mandela).

Yibambe! (Dialeto Xhosa).

RESUMO

FRANKLIN, Alex. *Framework* para Ciência de Dados no contexto de pequenas e médias empresas brasileiras. Qualificação (Mestrado em Engenharia de Produção) – Programa de Pós-Graduação em Engenharia de Produção e Sistemas (Área de Concentração: Métodos Quantitativos), Universidade Tecnológica Federal do Paraná (UTFPR). Pato Branco, 2022.

Pequenas e Médias Empresas (PMEs) são responsáveis por uma fatia considerável de mercado, seja em economias emergentes ou desenvolvidas. Na China, engloba por volta de 75% da força de trabalho, enquanto no Canadá e Países da União Europeia, essa taxa chega a 64% e 67% respectivamente ao passo que no Brasil cerca de 78% dos empregos são gerados somente em Micro e Pequenas Empresas. No entanto, a representatividade econômica não reflete o nível de adoção de Tecnologias de Informação (TI) para PMEs onde se estima uma difusão entre 7% e 33% nessa modalidade de negócio, enquanto para empresas de grande porte por volta de 77%. Questões como nível de maturidade de TI, falta de investimento e capacidades técnicas muitas vezes limitam a utilização de tais tecnologias às grandes empresas ou *startups* que nascem em ambiente digital, enquanto PMEs ficam a margem de um mercado em amplo crescimento. Dada a relevância em incorporar tecnologias disruptivas e o seu impacto no sucesso das organizações, este estudo procurou reunir informação por meio de revisão de literatura e pesquisa de campo, elementos para proposição de um *Framework* para Ciência de Dados (FCD) no contexto das PMEs brasileiras. O processo metodológico se deu pela utilização de modelagem de tópicos para constituição do portfólio bibliográfico com aplicação do algoritmo Alocação Latente de Dirichlet (ALD) no âmbito de mineração de texto, somado às contribuições do mercado por meio de entrevista com profissionais atuantes no segmento de tecnologia e que parte de sua atuação tenha sido em Pequenas e Médias Empresas brasileiras. A percepção de valor e ajustes do FCD genérico, os entrevistados concordaram que o FCD poderia ser usado para orientar a adoção de processos de Ciência de Dados em empresas genéricas. A melhoria na governança da informação foi mencionada como o ponto de maior valor, seguida pela melhoria na eficiência dos processos e aumento no desempenho da equipe. Os entrevistados também destacaram a clareza no escopo do projeto/produto, tomada de decisões, melhoria na governança de TI e outros benefícios proporcionados pelo FCD. A falta de capital humano qualificado e a baixa percepção de valor foram apontadas como as principais barreiras em PMEs. Outros obstáculos mencionados incluem limitações financeiras, falta de organização da empresa, cultura organizacional e disponibilidade tecnológica insuficiente. Como informações complementares, os entrevistados propuseram desdobramentos sobre interoperabilidade de sistemas, a segmentação do FCD por porte de empresa e a entrega de valor em fases baseadas no nível de maturidade da empresa. Em resumo, a pesquisa mostrou que o FCD genérico pode ser aplicado em PMEs como um guia para a estruturação do fluxo de dados e melhoria na eficiência na tomada de decisões em PMEs.

Palavras-chave: Pequenas e Médias Empresas. Ciência de Dados. *Framework*. *Text Mining*.

ABSTRACT

FRANKLIN, Alex. Framework for Data Science in the context of Small and Medium-Sized Brazilian Enterprises. Dissertation (Masters in Production Engineer) – Pos Graduate Program in Production Engineer (Concentration Area: Operation Management), Federal University of Technology - Paraná (UTFPR). Pato Branco, 2023.

Small and Medium Enterprises (SMEs) are responsible for a considerable market share in emerging or developed economies. In China, it comprises around 75% of the workforce, while in Canada and European Union countries, this rate reaches 64% and 67%, respectively, while in Brazil, around 78% of jobs are generated only in Micro and Small businesses. However, economic representation does not reflect the adoption of Information Technologies (IT) for SMEs, where diffusion is estimated between 7% and 33% in this type of business. At the same time, it is around 77% for large companies. Issues such as IT maturity level, lack of investment, and technical capabilities often limit the use of such technologies to large companies or startups born in a digital environment. At the same time, SMEs still need to catch up on the sidelines of a rapidly growing market. Given the relevance of incorporating disruptive technologies and their impact on the success of organizations, this study sought to gather information through literature review and field research, elements for proposing a Data Science Framework (DCF) in the context of Brazilian SMEs. The methodological process was based on topic modeling to create the bibliographic portfolio with the application of the Latent Dirichlet Allocation (ALD) algorithm in the context of text mining, added to market contributions through interviews with professionals working in the technology segment. And that part of its work has been in Small and Medium-Sized Brazilian Companies. Perceiving the value and adjustments of the generic FCD, interviewees agreed that the FCD could be used to guide the adoption of Data Science processes in generic companies. Improving information governance was mentioned as the point of most significant value, followed by improving process efficiency and increasing team performance. Respondents also highlighted clarity in project/product scope, decision-making, improved IT governance, and other benefits provided by the FCD. The need for qualified human capital and the low perception of value were identified as the main barriers in SMEs. Other obstacles include financial limitations, lack of company organization, organizational culture, and insufficient technological availability. As additional information, the interviewees proposed developments on systems interoperability, the FCD segmentation by company size, and the delivery of value in phases based on the company's maturity level. In summary, research has shown that the generic FCD can be applied in SMEs as a guide for structuring the data flow and improving efficiency in decision-making in SMEs.

Keywords: Smal and Medium Sized Enterprise. Data Science. Framework. Text Mining.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE QUADROS

LISTA DE SIGLAS E ACRÔNIMOS

DOI	*Digital Object Identifier*
UE	União Europeia
PIB	Produto Interno Bruto
FCD	*Framework* de Ciência de Dados
IBGE	Instituto Brasileiro de Geografia e Estatística
PME	Pequenas e Médias Empresas
EPP	Empresas de Pequeno Porte
ME	Microempresa
BNDES	Banco Nacional de Desenvolvimento Econômico e Social
Sebrae	Serviço Brasileiro de Apoio às Micro e Pequenas Empresas
CSV	*Comma Separeted Value*
ECI	*Event-based Communication Interface*
API	*Application Programming Interface*
MAPE	*Mean Absolute Percentage Error*
MAE	*Mean Absolute Error*
RMSE	*Root Mean Square Error*
LDA	*Latent Dirichlet Allocation*
TF	*Term Frequency*
TF-IDF	*Term frequency–Inverse Document Frequency*
TI	Tecnologia da Informação

SUMÁRIO

1 INTRODUÇÃO

1.1 CONTEXTUALIZAÇÃO

Este estudo tem como foco Pequenas e Médias Empresas brasileiras, também conhecidas como PMEs. No entanto, embora seja um escopo de grande relevância como será demonstrado a seguir, não há, no Brasil, definição única sobre a classificação de empresas quanto ao porte. A legislação brasileira utiliza o faturamento bruto como critério de marcação de microempresas, empresas de pequeno e grande porte, sendo a marcação de companhias de médio porte uma medida indireta das duas últimas — acima do limite de faturamento bruto das de pequeno porte e abaixo do limite que define as de grande.

Segundo o Artigo 3º da Lei Nº 123 de 2006 (Brasil, 2006), empresas de pequeno porte auferem receita bruta anual superior a R$ 360.000 e igual ou inferior a R$ 4.800.000. Pela Lei 11.638 de 2007, Artigo 3º (Brasil, 2007), empresas de grande porte apresentam faturamento bruto superior a R$ 300.000.000. Não obstante, entende-se que empresas de médio porte devem apresentar receita bruta anual superior a R$ 4.800.000 e igual ou inferior a R$ 300.000.000. Entidades como o Banco Nacional do Desenvolvimento (BNDES) compartilham da mesma classificação legal, embora haja larga utilização do número de colaboradores para realizar tais definições. Para o Serviço Brasileiro de Apoio às Micro e Pequenas Empresas (Sebrae) e o Instituto Brasileiro de Geografia e Estatística (IBGE), por exemplo, esse critério é aplicado em conjunto com o setor econômico cujos limites são apresentados no Quadro 1.

Quadro 1: Classificação porte de empresas por número de empregados

Porte	**Comércio e Serviços**	**Indústria**
Pequeno	De 10 a 49 empregados	De 20 a 99 empregados
Médio	De 50 a 99 empregados	De 100 a 499 empregados

Fonte: Anuário do trabalho na micro e pequena empresa (2013)

Na Europa, a classificação Pequenas e Médias Empresas utiliza como critérios o número de colaboradores e o faturamento anual. Tal abordagem

apresenta-se como uma aplicação conjunta dos fatores que em geral, no Brasil, utiliza-se em separado. O Quadro 2 reúne os limites utilizados na União Europeia (UE) e nota-se que há discrepância se comparados com os limites brasileiros.

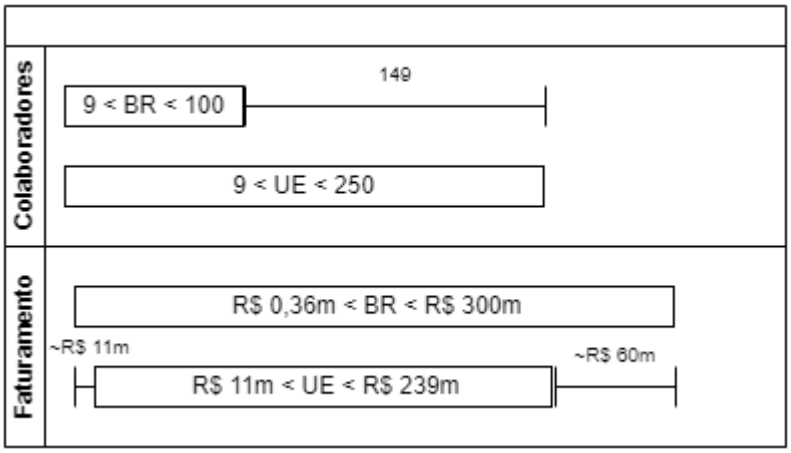
Quadro 2: Classificação Europeia para PMEs

Porte	Número de Empregados	Faturamento Anual
Pequeno	De 10 a 49 empregados	De € 2 milhões a € 10 milhões
Médio	De 50 a 249 empregados	De € 10 milhões a € 43 milhões

Fonte: European Commission (2022)

Para o fator "Número de Empregados", as PMEs brasileiras detêm entre 10 e 99 colaboradores contra o intervalo de 10 a 249 na UE. Segundo o Banco Central do Brasil (2023), a cotação do Euro no dia 31/12/2022 foi de R$ 5,5666, o que significa limites de faturamento para PME na UE entre R$ 11.133.200 e R$ 239.363.800. A Figura 1 ilustra uma sobreposição das fronteiras de classificação para ambos os cenários descritos acima — Brasil e União Europeia — evidenciando para este último uma tolerância maior para o número de colaboradores (149 a mais) e inferior para o faturamento anual (cerca de R$ 60 milhões a menos). Não obstante, resguardados demais fatores culturais e conjunturais dos dois cenários, tratar-se-á neste estudo como equivalentes PMEs brasileiras e PMEs estrangeiras.

Figura 1: Esquemático para classificação de empresas: Brasil x UE



Fonte: Autor (2023)

Hansen e Bøgh (2021) destacam que PMEs demonstram um papel preponderante na movimentação econômica de diversos países. Representam cerca de dois terços dos empregos gerados nos Estados Unidos e 99% dos empregos gerados na União Europeia. Apontam também sua vantagem em relação à

flexibilidade para adoção de processos de inovação por conta do menor número de camadas de decisão em relação a empresas de maior porte. Cerca de 600 milhões de empregos devem surgir até 2030 no mundo, o que reforça o papel das PMEs na absorção de 70% dessa mão de obra segundo World Bank SME Finance (2022).

O Brasil conta com mais de nove milhões de empresas conforme dados do Ministério da Economia (2022) das quais cerca de 10% são Empresas de Pequeno Porte (EPP) (Sebrae, 2020). Somadas a Micro Empresas, a participação no Produto Interno Bruto (PIB) é de 27% segundo dados do Sebrae em relação à pesquisa realizada junto à Fundação Getúlio Vargas em 2011 (Barretto, 2015). Segundo a mesma fonte, a riqueza gerada por Micro e Pequenas Empresas quadruplicou entre 2001 e 2011, passando de R$ 144 bilhões para R$ 599 bilhões. Quanto à mão de obra, o autor destaca que essa categoria de empresas é responsável por 52% dos contratos formais e 40% da massa salarial brasileira.

A participação econômica de PMEs no Brasil e no mundo não é acompanhada necessariamente de uma estrutura qualificada de informações e decisões orientada a dados. Mesmo para empresas de comércio eletrônico que atualmente fazem uso intensivo de ferramentas de gestão da informação objetivando monitorar o tráfego de atividades comerciais, o acesso ao conjunto de métricas de mais alto nível está disponível muitas vezes em pacotes comerciais distantes da realidade das PMEs (García et al.,2016). Dong e Yang (2020) destacam ainda que além das limitações de recursos financeiros, a limitação de capital humano pode ser um fator importante na adoção de novas tecnologias.

Palanivel Rajan D. et al. (2020) evidencia os principais desafios para o desenvolvimento e implementação de aplicações de Ciência de Dados em diferentes contextos. O autor aponta a interoperabilidade, acesso a informação e políticas de saúde como principais entraves no âmbito da indústria de saúde. Para comércio eletrônico, segurança da informação e validação em tempo real configuram-se nas principais barreiras. Da perspectiva agrícola, com semelhante dificuldade no tocante a interoperabilidade em diferentes padrões, a produção interna sob padrões controlados também agrega restrições às aplicações mencionadas. Na indústria financeira, as restrições estão fortemente ligadas a crime cibernéticos e regulações de mercado. Finalmente, para projetos de Ciência de Dados no contexto ambiental,

natureza e complexidade dos dados podem se apresentar como fatores limitantes em estudos e aplicações.

Azevedo e Almeida (2021) contrapõem os dois lados da transformação digital em PMEs, sendo que à primeira vista se tem organizações flexíveis e com processos ágeis, de outro empresas com restrições financeiras, mais conservadoras por conta da natureza familiar de gestão, inflexíveis e com dificuldade de inovar. Os autores demonstram que o segundo ponto pode se configurar em uma barreira mais forte a ser ultrapassada, com destaque para a falta de conscientização sobre as capacidades digitais, capital humano qualificado para a transição, recursos financeiros, padronização de processos e segurança da informação.

Ampliando o foco sobre o fator “capital humano”, Demchenko et al. (2017) trazem a discussão sobre a necessidade de formação de profissionais da área de análise de dados e que, o não atendimento desta demanda, pode acarretar em baixa performance econômica, industrial, acadêmica e perda de produtividade. O autor destaca ainda que tendo por base o caráter multidisciplinar da Ciência de Dados, uma economia moderna precisa passar por uma reformulação nos modelos tradicionais de ensino dos cursos existentes.

1.2 PROBLEMA DE PESQUISA

Embora os sistemas de informação estejam presentes em diversos segmentos de atividades, uma gama de empresas ainda se encontra à margem dos avanços e benefícios oriundos da digitalização. O desempenho de companhias nesta situação pode acarretar numa redução da competitividade frente a empresas de maior porte (Boonsiritomachai et al., 2016). No entanto, reduzir essa distância não é tarefa fácil na maioria das vezes, pois como apresentado por Gauzelin e Bentz (2017), dificuldades financeiras e a falta de experiência configuram barreiras importantes na implantação e gestão de sistemas orientados a dados. O World Bank SME Finance (2022) estima que cerca de 40% das PMEs de países em desenvolvimento possuem uma necessidade de financiamento não atendida o que representa aproximadamente U$ 5 trilhões por ano — Oriente Médio e Norte da África com 88% de necessidade não atendida, seguido de América Latina e Caribe

com 87% segundo o mesmo relatório.

Albaz et al. (2020) destacam a natureza jovem de PMEs e baixa escala como fator que dificulta o suporte por parte das políticas públicas. Em seu trabalho os autores citam que as agências governamentais devem focar seus esforços em aumentar a confiança nas PMEs, ações que permitam seu crescimento e aumentar a competitividade. O primeiro fator pode ser descrito por mudança na cultura e desburocratização fomentando o empreendedorismo. No segundo caso, criar canais que permitam o acesso a mercados globais, centros de inovação e facilitar a expansão do negócio. O último item identifica as agências governamentais e não governamentais como provedores de suporte para infraestrutura, produtividade e adoção de tecnologia.

Frente ao exposto, este estudo revela que até mesmo a produção de conteúdo formativo é impactada por estar limitada em grande parte a aplicações e estudos dirigidos a empresas de grande porte. Visando o contexto apresentado essa pesquisa busca responder a seguinte questão: “Qual estrutura tecnológica de Ciência de Dados se enquadra às necessidades e ao contexto de Pequenas e Médias Empresas brasileiras?”

1.3 OBJETIVOS

1.3.1 Objetivo Geral

Este estudo objetiva propor uma *Framework* de Ciência de Dados (FCD) no contexto de Pequenas e Médias Empresas Brasileiras.

1.3.2 Objetivos Específicos

Para o atingimento do objetivo geral, os seguintes objetivos específicos devem ser alcançados:

i. Selecionar elementos de Ciência de Dados por meio de revisão de literatura;

ii. Levantar requisitos para implantação de tecnologia em PMEs por meio de revisão de literatura;
iii. Elaborar *Framework* genérico de Ciência de Dados;
iv. Ajustar *Framework* genérico de Ciência de Dados ao contexto de PMEs por meio de pesquisa estruturada com especialistas;

1.4 JUSTIFICATIVA

Dados da FIESP (2004) apontam como principais causas de mortalidade de Pequenas e Médias Empresas o comportamento do empreendedor, a falta de planejamento, deficiências na gestão, falta de políticas públicas, conjuntura econômica e problemas pessoais. Estes dados mostram ainda que cerca de 40% sobrevivem após o quinto ano desde a fundação. Apesar da expressiva representatividade econômica citada e a dificuldade de sobrevivência, as PMEs são avaliadas do ponto de vista governamental, sob os mesmos quesitos de empresas de grande porte (Smits et al., 2018). Dessa forma, além de políticas públicas também se mostra relevante do ponto de vista acadêmico obter soluções enquadradas na conjuntura de Pequenas e Médias Empresas.

Mesmo para países desenvolvidos como França e Alemanha, a diferença de produtividade entre PMEs e empresas de grande porte é de 26% e 41% respectivamente segundo Albaz et al. (2020); no mundo a perda de valor adicionado está em torno de U$ 15 trilhões. Os autores destacam ainda que práticas de inovação no contexto de PMEs, similarmente à proposta deste estudo, podem impactar o desenvolvimento do país como um todo por dois fatores: primeiro por conta do potencial inexplorado na adoção de tecnologias e segundo por se apresentarem como fontes de inovação — dado que não estão prejudicadas por sistemas legados e estratégias desatualizadas como em empresas de grande porte.

Alta volumetria de dados permite melhor aderência de modelos e precisão das previsões em aplicações de Ciência de Dados (Han; Trimi, 2022; Philemotte, 2020), corroborando a necessidade de uma infraestrutura em nuvem para suportar uma demanda massiva de armazenamento e fornecer maior poder computacional às aplicações (Palanivel Rajan D. et al., 2020). A computação em

nuvem é fator decisivo no processo de adoção de tecnologias por permitir maior colaboração; sua utilização em larga escala fica atrás somente de serviços essenciais como água, eletricidade, gás e telefonia (Han; Trimi, 2022). Outro fator de grande relevância na utilização de provedores de serviços em nuvem é sua capacidade de fornecer uma gama de produtos de dados (armazenamento, processamento, análise, etc.) adequados a empresas de pequeno, médio e grande porte com custos de utilização variáveis (*pay-per-use*) Demchenko et al. (2017) evitando despesas imobilizadas com licenças ou estruturas subutilizadas.

Outro ponto a ser destacado, é que com o novo paradigma da Indústria 4.0 e seu caráter colaborativo, Pequenas e Médias Empresas podem se beneficiar da nova forma de abordar os processos e tecnologias de vanguarda. Mesmo sob as barreiras financeiras e capacitação técnica, a estrutura tecnológica defasada permite um espaço de crescimento superior ao de empresas de grande porte (Han; Trimi, 2022). Nesta conjuntura, Delgado et al. (2020) demostram que a Ciência de Dados vem se constituindo como uma disciplina própria, com aplicações em alto volume, variedade e velocidade de processamento tanto para dados estruturados como não estruturados o que possibilita uma abordagem mais versátil e adequada a cada contexto organizacional. Os autores destacam ainda a relevância de um *Framework* para Ciência de Dados no delineamento de processos para extração, análise, qualidade, integração e aplicações com dados.

Watson et al. (2017) aponta que a Ciência de Dados deriva da convergência de diferentes disciplinas nos campos da matemática, estatística e ciência da computação, atuando sobre dados gerados em atividades humanas como vendas *online*, pesquisas e exames médicos por exemplo, mas com parte da volumetria proveniente de máquinas (*Machine to Machine* ou M2M), tornando os avanços nos processos de coleta, armazenamento e análise de dados amplamente aceitos como o quarto paradigma da ciência. Yee et al. (2020) complementa a definição da área com a incorporação de disciplinas de comunicação, gestão e sociologia no estudo de dados de forma a prover um ambiente de geração de ideias e informação para o processo decisório.

A versatilidade de aplicações de Ciência de Dados pode ser percebida por meio dos diferentes contextos em que vem sendo empregada: “detecção de

fraude, algoritmos de recomendação, detecção de notícias falsas, *chatbot*, diagnóstico de câncer, reconhecimento facial e de emoções, precipitação de chuvas e carros autônomos" (Palanivel Rajan D. et al., 2020). Na esfera acadêmica, suas aplicações permitem ao pesquisador inferências que de outra forma seriam impraticáveis ou inviáveis Gupta et al. (2021). Para Han e Trimi (2022), sua utilização por PMEs pode prover melhor eficiência operacional, ganho na gestão da cadeia de suprimentos e prevenção de falhas.

Como fator limitante o planejamento e controle de processos fundamentados por projeções quantitativas, Han e Trimi (2022) mencionam o estudo no qual 78% das empresas já aplicavam tecnologias da Indústria 4.0, mas apenas 2% utilizavam *Big Data*, evidenciando a janela de oportunidade para o presente trabalho. Como evidência do fator de crescimento, Demchenko et al. (2017) nos mostra que o mercado de dados era da ordem de 60 bilhões de euros em 2016 com expectativa de crescimento para 106 bilhões de euros em 2020 e destaque para uma lacuna estimada em 9,8% de profissionais de dados para 2020. Neste mesmo contexto, Palanivel Rajan D. et al. (2020) avalia que o mercado de Ciência de Dados chegará a U$ 385 bilhões em 2025 — um crescimento anual em torno de 39%.

1.5 DELIMITAÇÃO DA PESQUISA

Do ponto de vista teórico, Berto e Nakano (1998) destacam que a pesquisa na Engenharia de Produção pode ser delineada segundo três critérios: modelo de pesquisa, conduta de pesquisa e propósito da pesquisa. Em relação ao modelo de pesquisa, este estudo é classificado como teórico-conceitual por valer-se de observações da literatura e compilação de ideias para modelagem teórica. No quesito conduta de pesquisa, utiliza-se uma combinação qualitativa-quantitativa uma vez que a seleção de materiais seguirá uma abordagem quantitativa — mensurável, genérica e replicável — enquanto a estruturação do modelo partirá de uma premissa qualitativa — enfoque desestruturado, hipóteses construídas ao longo do processo e com múltiplas fontes de dados (revisão de literatura). Referente ao propósito da pesquisa, este estudo se enquadra como uma pesquisa descritiva com intuito de examinar um fenômeno para defini-lo com maior precisão (*framework* mais

adequado às PMEs) e capturar a essência do fenômeno durante a coleta de dados (barreiras intrínsecas às PMEs).

Quanto ao porte das empresas, este trabalho versará apenas sobre Pequenas e Médias Empresas como já mencionado na contextualização. As micro empresas, apesar da forte representatividade em relação ao número de companhias existentes, Hairuddin et al. (2012) destacam uma série de barreiras à adoção de tecnologias que por si só já seriam objeto de estudo como: características do proprietário/gerente, características da empresa e custo e retorno sobre o investimento (ROI). Neste contexto, avaliar-se-á apenas empresas que já tenham avançado no sentido de usar tecnologia como vantagem competitiva.

Quanto à origem de coleta dos documentos, apenas documentos indexados na base de dados *Scopus* foram considerados por permitirem uma extração estruturada de metadados das suas publicações; o que facilita a etapa de pré-processamento e entrada de informações no modelo que está detalhado posteriormente na Seção 2. Cauchick Miguel et al. (2018) destacam que o processo de proposição de um modelo em Gestão de Operações parte de uma varredura horizontal (*literature search*) para identificar os pontos focais no assunto (autores e publicações) para só então realizar uma varredura vertical (*literature review*). No contexto abordado neste estudo, foi utilizado apenas a varredura vertical com eixos de pesquisa genéricos de forma a capturar o maior número de materiais e definir a agenda de pesquisa — isso por conta da quantidade limitada de materiais disponíveis. No tocante aos objetivos finais deste trabalho, sua participação encerra no modelo proposto, sem que haja desenvolvimento de tecnologias e aplicações práticas (sistema ou plataforma), possibilitando estudos futuros acerca do tema.

1.6 ESTRUTURA DA DISSERTAÇÃO

Inicialmente, este estudo traz em sua estrutura aspectos introdutórios como contexto, problema de pesquisa, objetivos, justificativa e delimitação da pesquisa. No Capítulo 2, a revisão de literatura é dividida em duas partes: análise completa da população de documentos referente a "*Framework* de *Ciência de Dados*" e seleção de portfólio referente a "*SMEs & Information Management*" por

meio de mineração de texto. No Capítulo 3 é apresentado o enquadramento metodológico e *framework* da pesquisa. O Capítulo 4 reúne-se os resultados em relação ao Framework Genérico de Ciência de Dados e pontos de ajuste para adequação a Pequenas e Médias Empresas e finalmente, no Capitulo 5, as considerações finais e propostas de estudos futuros.

2 REVISÃO DE LITERATURA

Um número cada vez maior de publicações está disponível a cada ano, o que possibilita uma gama de estudos aprofundados sobre uma variedade de assuntos. No entanto, tal disponibilidade traz consigo um aumento considerável na carga de trabalho por parte do pesquisador durante o processo de coleta e triagem de materiais.

O processo de seleção de materiais a ser utilizado neste estudo teve como entradas metadados ou atributos de cada documento: nome do autor, título da publicação, ano de publicação, periódico, DOI, link, resumo, palavras-chave utilizadas pelo autor, referências, idioma de publicação, tipo de publicação e filiação do autor. Estas informações podem ser obtidas manualmente de cada documento, no entanto, optou-se pela base de dados Scopus por permitir a extração de metadados em estrutura consolidada no formato CSV (*Comma-separated values*), o que facilita ainda a manipulação de dados nas etapas seguintes. Além da versatilidade, entende-se que não há prejuízo de volumetria e qualidade como assertado por Falagas et al. (2008) de que a Scopus detém um espectro mais amplo que *Web of Science* e *Pubmed*, por exemplo, e sua análise de citações inclui mais artigos que a *Web of Science* e inclui cerca de 20% a mais de documentos que essa base no mesmo período.

A segunda etapa foi definir quais eixos de pesquisa utilizar (Tabela 1). Esta prática é normalmente utilizada para realizar uma primeira triagem, visando convergir os resultados de pesquisa para o tema de estudo e reduzir a base bruta de materiais. Com somente um resultado, a primeira combinação mostrou-se pouco promissora com o cruzamento direto dos eixos principais do estudo — *Data Science Framework & SMEs*. Desta forma, optou-se por segmentar a coleta de materiais em duas etapas genéricas de modo a ampliar o portfólio bibliográfico. A primeira, visou reunir documentos que mapeassem o conhecimento em torno do tema *Framework* de Ciência de Dados sem utilização de restrição referente às características organizacionais (objetivo específico i). No segundo grupo, a busca se deu pela seleção de estudos que identificassem o processo e a estrutura de gestão da informação em Pequenas e Médias Empresas (objetivo específico ii).

Tabela 1: Eixos de pesquisa

Combinações	**Query**	**#Resultados**
Data Science Framework & SMEs	*("data science platform" OR "data science framework" OR "data science model") AND ("small and medium sized enterprise" OR "small firm" OR "small business" OR sme OR "small and medium enterprise")*	1
Data Science Framework	*("data science platform" OR "data science framework" OR "data science model")*	173
SMEs & Information Management	*("small and medium sized enterprise" OR "small firm" OR "small business" OR sme OR "small and medium enterprise") AND ("information technology" OR "information management" OR "information systems")*	4393

A coleta de informações do grupo *Data Science Framework* (Tabela 1) retornou uma quantidade de materiais factível de avaliação populacional (173 documentos) — os critérios de inclusão e exclusão serão apresentados em detalhes na seção 2.1. Não obstante, no grupo referente a "*SMEs & Information Management*" a quantidade sobressalente de materiais (4.393 documentos) permitiu ao pesquisador lançar mão de métodos semiautônomos de agrupamento como a aplicação de algoritmos de modelagem de tópicos e selecionar o grupo de documentos cujo contexto está pertence a pesquisa. O Método aplicado será detalhado na seção 2.2.

2.1 *FRAMEWORK* PARA CIÊNCIA DE DADOS

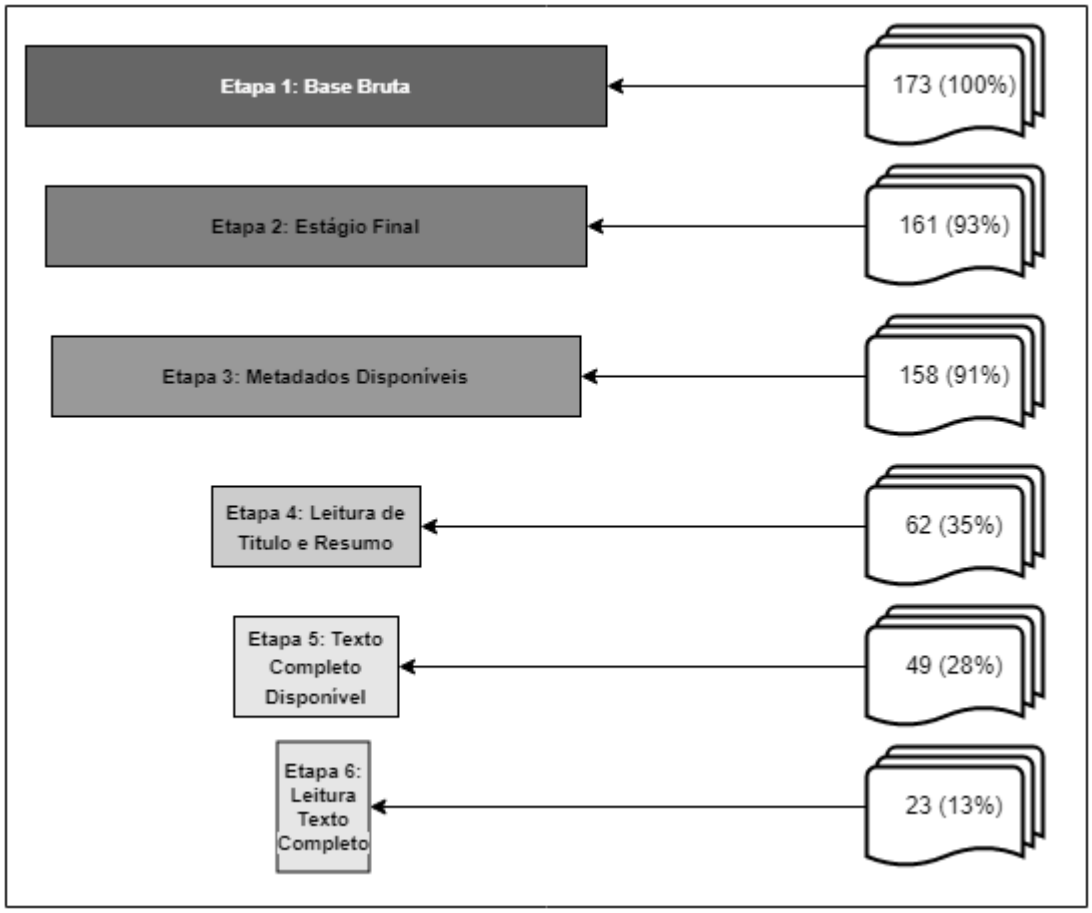
2.1.1 Estratégia para Seleção de Portfólio

Como supracitado, esta seção discorre sobre o processo de seleção de portfólio referente ao grupo *Data Science Framework* da Tabela 1 e as 173 publicações

resultantes da aplicação da query de busca "*"data science platform" OR "data science framework" OR "data science model"*" na base de dados *Scopus*; esta etapa constitui a base bruta de materiais.

A Figura 2 esboça o funil de publicações segundo os filtros aplicados. O primeiro corte foi realizado sobre registros que estejam em estágio final de publicação, restando 161 documentos nesta etapa. Na terceira etapa, selecionou-se apenas registros cujos atributos utilizados para avaliação do documento estavam disponíveis — DOI, Nome do Autor, Filiação e Resumo — e sem os quais não seria possível analisar a aderência com este estudo e o desempenho acadêmico do material, do autor e impacto do periódico no qual a publicação foi submetida; 158 documentos seguiram à próxima etapa.

Figura 2: Funil de seleção do portfólio bibliográfico – *Data Science Framework*



Na quarta fase do processo de seleção, adotou-se a avaliação de aderência por meio da leitura do título e resumo das publicações remanescentes da

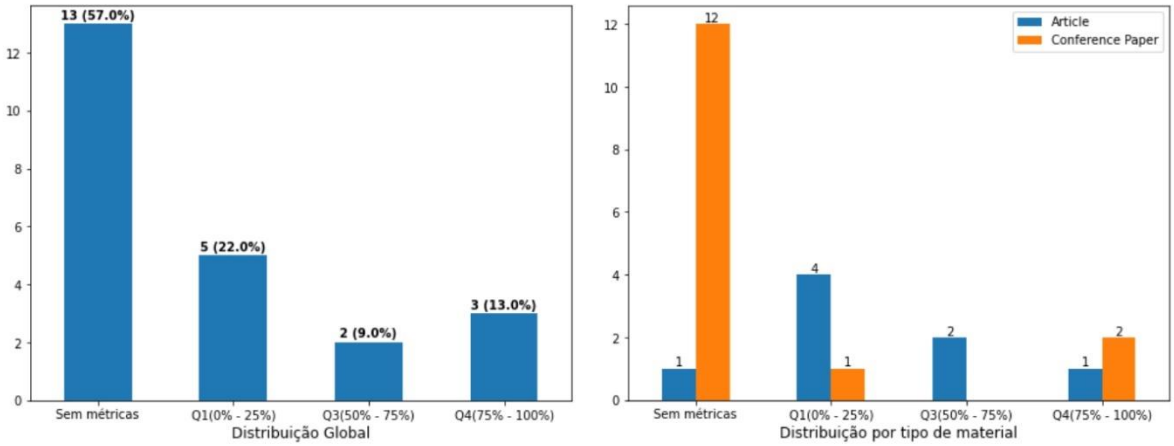
Etapa 3. Esta abordagem resulta em 35% da quantidade bruta inicial — 62 documentos. Na última fase, Etapa 5, considerou-se apenas para publicações disponíveis em sua forma completa, o que leva a remoção de 13 materiais e constituição do portfólio de 49 documentos para leitura completa; 28% da quantidade original. Nesta última etapa, 26 documentos foram desconsiderados por baixa aderência a agenda de pesquisa e 23 publicações formaram o portfólio final que dá subsídio teórico a este estudo.

2.1.2 Análise de Desempenho

Esta etapa tem por objetivo realizar o enquadramento do portfólio constituído na etapa anterior em relação à relevância acadêmica dos materiais em relação a seus pares e extrair informações que permitam inferências em relação à agenda de pesquisa deste estudo.

Relativo ao primeiro objetivo, a Figura 3 apresenta a distribuição de publicações do portfólio em relação a seu posicionamento na distribuição quartílica do *CiteScore* 2020. Agência de Bibliotecas e Coleções Digitais (2016) definem *CiteScore* como “o número médio de citações recebidas em um ano por todos os itens publicados nesse periódico nos três anos anteriores”.

Figura 3: Número de documentos por quartil do CiteScore 2020 Ciência de Dados



A ilustração mostra que 57% (13 artigos) portfólio deste estudo referente ao contexto de Ciência de Dados & PMEs, não possuem métricas disponíveis; sendo que destes, 12 são Artigos de Conferência e 1 Artigo de

Periódico. Associado à escassez de materiais a respeito do tema, a volumetria de artigos de alto impacto (13%) corrobora a lacuna de pesquisa existente referente ao tema deste estudo.

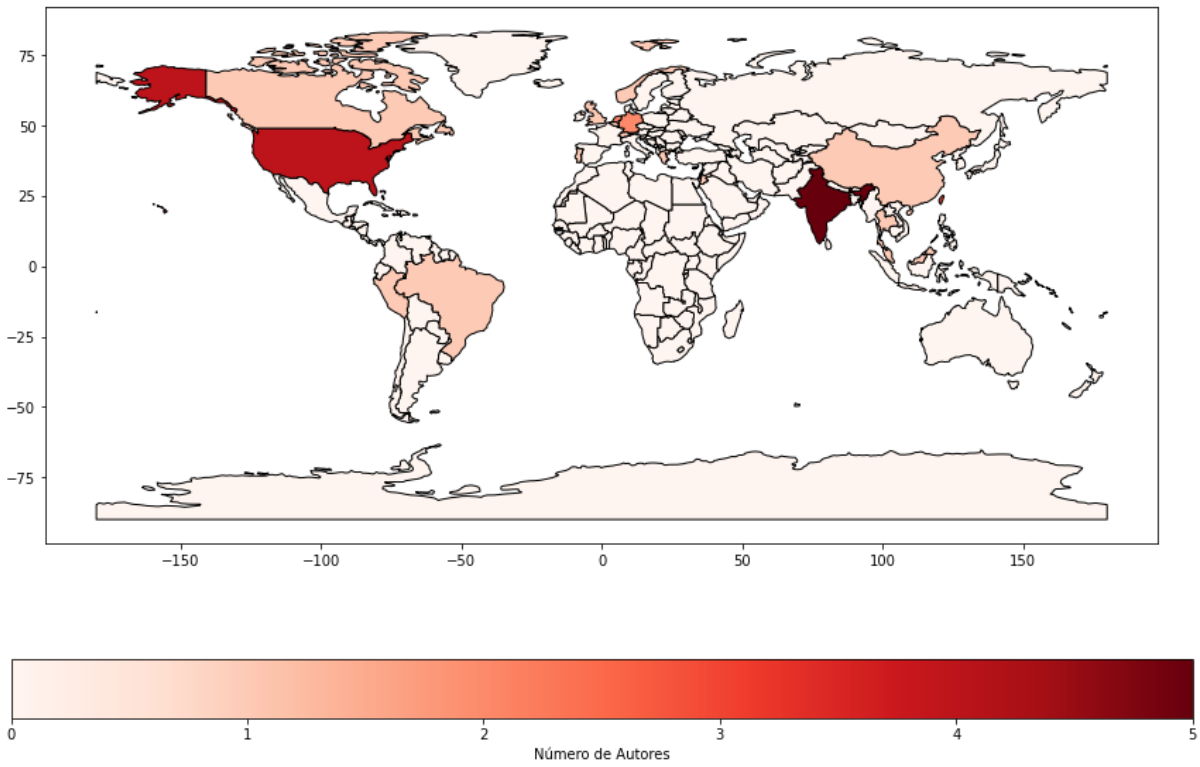
O agrupamento apresentado na Tabela 2 mostra evidências de uma distribuição uniforme entre os periódicos do portfólio com *CiteScore* 2020 disponível — com mais de uma ocorrência apenas *ACM International Conference Proceeding Series* e *Advanced Engineering Informatics* com dois artigos cada. Em relação à área de conhecimento, percebe-se ainda uma concentração de publicações no âmbito técnico — *software, artificial intelligence, computer networks and communications, computer science applications, computer vision and pattern recognition, human-computer interaction, information systems* — enquanto publicações envolvendo um direcionamento estratégico e gerencial (*strategy and management* e *business and international management*) apresenta-se como área de pouca expressão dentro do portfólio.

Tabela 2: Número de Publicações do portfólio por Periódico/Área de Conhecimento

Journal	Area	#documentos
ACM International Conference Proceeding Series	*computer networks and communications;computer vision and pattern recognition;human-computer interaction;software*	2
Advanced Engineering Informatics	*artificial intelligence;information systems*	2
IEEE Potentials	*education;electrical and electronic engineering;strategy and management*	1
Informatica (Slovenia)	*artificial intelligence;computer science applications;software;theoretical computer science*	1
Journal of Library Administration	*library and information sciences;public administration*	1
Journal of Nursing Scholarship	*nursing*	1
Robotics and Computer-Integrated Manufacturing	*computer science applications;control and systems engineering;industrial and manufacturing engineering;mathematics ;software*	1
Technological Forecasting and Social Change	*applied psychology;business and international management;management of technology and innovation*	1

Geograficamente, a Figura 4 mostra Estados Unidos (5) e Índia (4) como mais expressivos no quesito filiação dos autores por país. Mas do ponto de vista continental, Asia e Europa ainda se configuram como centros geradores de conteúdo a respeito do tema com doze e dez autores respectivamente; América do Sul e América do Norte totalizam 7 autores conjuntamente. Dessa lista, o Brasil é um dos países de menor expressão com apenas uma publicação (Gaedke Nomura et al., 2021).

Figura 4: Distribuição geográfica de filiações do portfólio bibliográfico Ciência de Dados

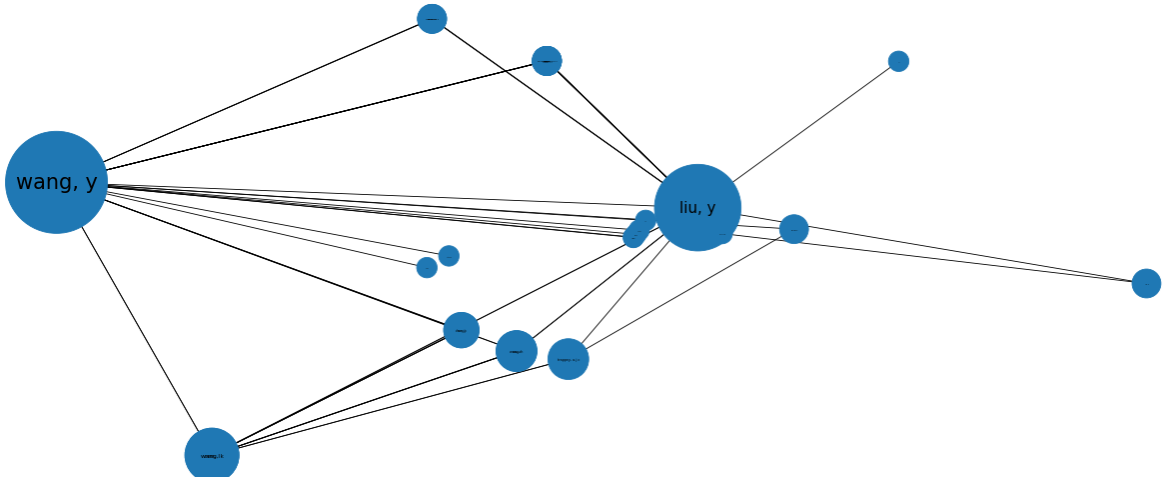


Com objetivo de identificar os autores de maior expressão dentro do portfólio, utilizou-se uma análise de rede por meio de uma matriz de coocorrência. O processo prevê a marcação dos autores de cada referência os conectando com os demais autores do mesmo documento, em seguida avaliando suas conexões com os demais artigos do portfólio. As métricas utilizadas nesta análise foram o número de conexões de cada autor com a população de autores e a coocorrência par a par entre dois autores, indicando um centro de competência em relação ao tema.

A Figura 5 ilustra os resultados obtidos quando se seleciona apenas autores que estão em coocorrência com outros autores mais de uma vez e que possuem número de conexões acima do percentil 99%. Esta última régua foi

atribuída com intuito de encontrar autores com relevância considerável em relação aos demais. De posse da análise visual supracitada, procedeu-se a identificação dos materiais com presença dos autores mais relevantes, direcionando o processo investigativo num primeiro momento aos materiais de maior relevância acadêmica (Quadro 3).

Figura 5: Análise de rede em relação as referências do portfólio bibliográfico Ciência de Dados



Quadro 3: Ocorrência de autores representativos no portfólio bibliográfico Ciência de Dados

Referência	Wang, Y	Liu, Y
Qin et al. (2020)	x	-
Chen et al. (2021)	x	x
Han; Trimi (2022)	x	x

2.1.3 Análise de Conteúdo: Elementos de Ciência de Dados

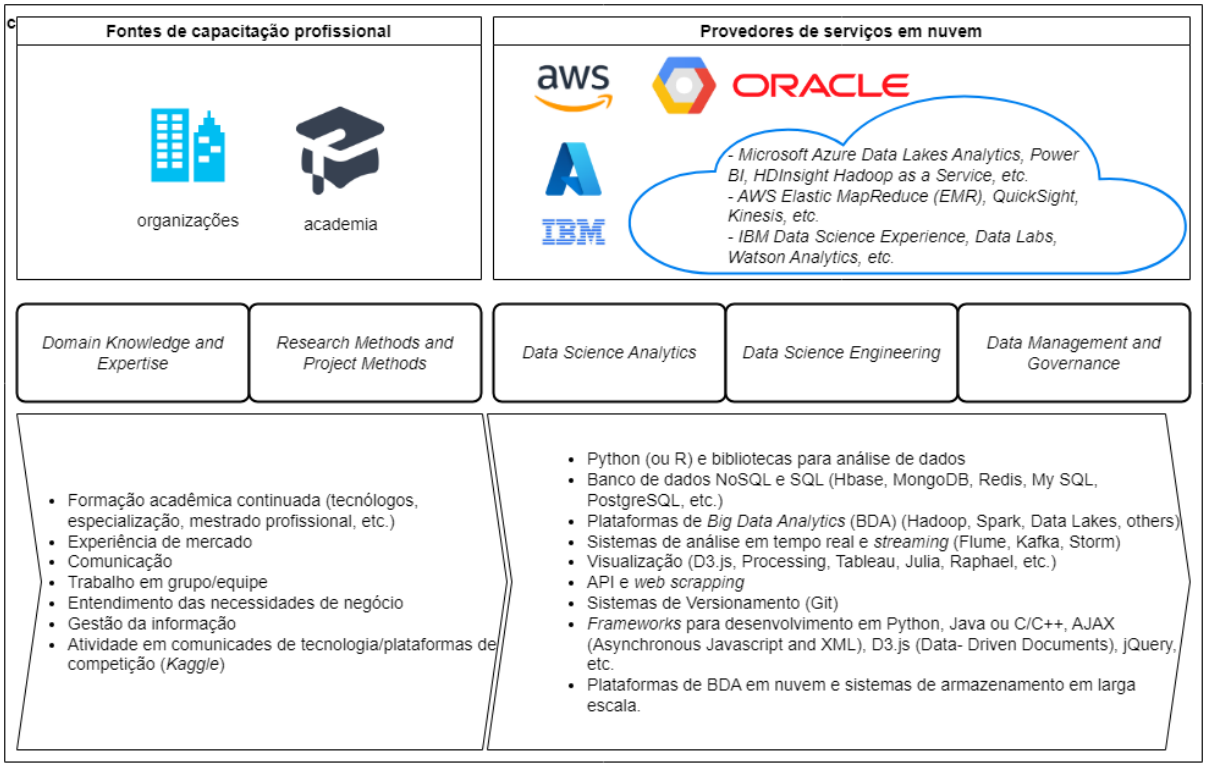
É esperado desta seção atender ao objetivo específico i "*Selecionar elementos de Ciência de Dados por meio de revisão de literatura*". O ponto de partida deste estudo busca entender, *a priori*, quais elementos fazem parte do escopo de um profissional que atua nesta frente. EDISON (2017) reúne uma coleção de documentos para definição da Ciência de Dados como profissão — com acesso gratuito, essa coleção visa direcionar orientadores educacionais, instrutores, empregadores, gerentes e até mesmo Cientistas de Dados. Com base nestes documentos, Demchenko et al. (2017) abordam quais habilidades pessoais e técnicas são esperadas de Cientistas de Dados e os grupos funcionais de atuação.

Os autores supracitados dividem as habilidades requeridas em dois tipos — Tipo A e Tipo B. O primeiro caso refere-se à experiência e competências adquiridas ao longo da carreira (*soft skills*), enquanto o segundo caso está relacionado a habilidades computacionais (*hard skills*) como linguagens de programação, ambientes de desenvolvimento e plataformas em nuvem. No que se refere aos grupos funcionais, a Ciência de Dados foi dividida segundo o mesmo estudo em *Data Science Analytics* (Análise Estatística, Aprendizagem de máquina, Mineração de dados, Análise de negócio, outros), *Data Science Engineering* (Engenharia de software, *Data Warehousing*, Infraestrutura de e ferramentas de *Big Data*), *Data Management and Governance* (administração, curadoria e preservação de dados), *Research Methods and Project Methods*, *Domain Knowledge and Expertise* (contexto científico).

Estruturas em nuvem como *Amazon Web Services* (AWS), *Google Cloud Platform* (GCP) e *Microsoft Azure* permitem que cientistas de dados tenham acesso a grande poder computacional e com alta elasticidade. Neste cenário a alocação de recursos é realizada conforme demanda que, somado ao maior controle de custos, também permite acesso simultâneo por diversos colaboradores. O autor cita ainda o *Jupyter Notebook* como ferramenta que pode ser utilizada no contexto colaborativo por fornecer um ambiente em código aberto e que suporta uma gama de *frameworks*, linguagens de programação e bibliotecas de *Deep Learning* como Apache Spark, Python, R e Keras, para o desenvolvimento de modelos preditivos, por exemplo. Marungo et al. (2015) sugere ainda que um ambiente de teste desconectado da estrutura principal (*sandbox*) seja utilizado para que não haja prejuízo ou danos aos dados originais (Lei et al., 2020).

A Figura 6 ilustra o relacionamento das competências e habilidades mencionadas acima com os grupos funcionais, bem como provedores de conhecimento e serviços de tecnologia em nuvem. Apesar da imagem se apresentar como um mapa e deter caráter estático, é preciso ressaltar que há fluxo de informação ao longo dos grupos, uma vez que tanto as competências pessoais e interpessoais são requeridas em camadas mais técnicas, quanto o conhecimento relativo as camadas técnicas podem suportar decisões em camadas de cunho estratégico.

Figura 6: Visão geral da atuação em Ciência de Dados



Fonte: Adaptado de Demchenko et al. (2017)

Berinato (2019) desta que uma operação de ciência de dados engloba uma ampla combinação de possibilidades para projetos desde uma abordagem simples para análise até a aplicação de algoritmos de aprendizagem de máquina mais sofisticados. Dessa forma, o autor aponta que uma equipe deve ser composta com base nas competências — Gestão de Projetos, Limpeza de dados, Análise de Dados, Especialista de Negócio, *Design* e *Storytelling* — e não pelo número de colaboradores, disponibilizando um conjunto de habilidades requeridas em cada etapa do processo de ciência de dados.

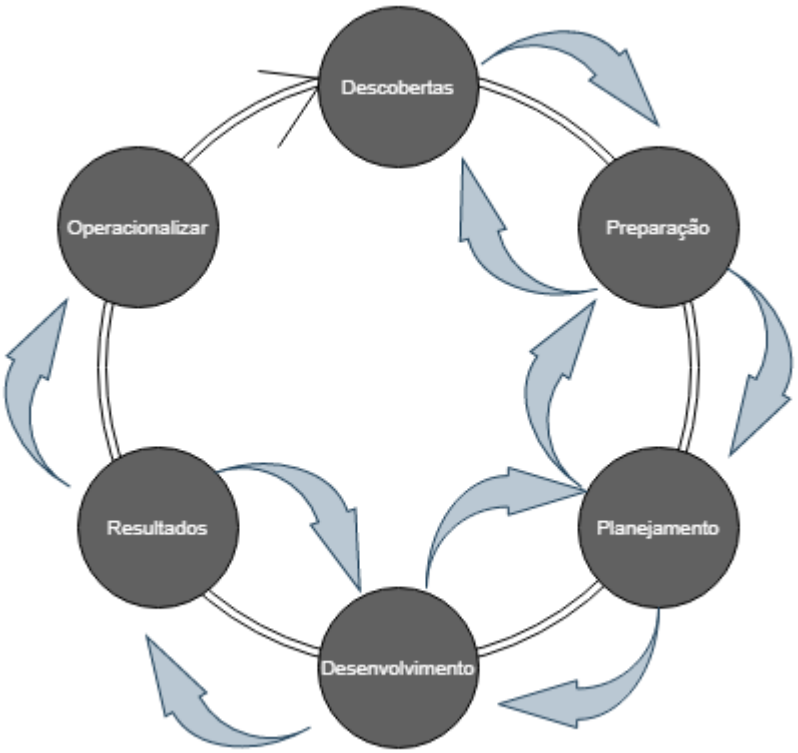
A ciência de dados opera sobre ambientes de alta complexidade em virtude principalmente do caráter distribuído de processamento que, apesar de beneficiar-se da alta disponibilidade de recursos, também gera o ônus de evitar inconsistências ao longo da jornada de dados Philemotte (2020). Saraee; Silva (2018) aponta que o ciclo de vida proposto por Dietrich et al. (2015) e apresentado na Figura 7 não reflete todos os processos necessários em projetos de ciência de cados como no caso de projetos geoespaciais, cuja atenção principal deve ser dada na fase de aquisição dos dados de múltiplas fontes não consideradas no modelo.

No entanto, este ciclo de vida lança um enfoque importante sobre cada

uma das etapas que é o processo de *feedback*. Dietrich et al. (2015) destaca que as equipes de ciência de dados ou de análise de dados, durante o desenvolvimento, na grande maioria das vezes, geram conhecimentos em fases intermediárias que as permitiria retornar as fases predecessoras e realizar ajustes ou refino do trabalho com base nos novos insights e novas descobertas.

Lei et al. (2020) aponta que em grande parte dos casos as aplicações de Ciência de Dados são compartilhadas ao cliente final somente após o atendimento de requisitos internos de modelagem. Fator este que impede o processo de *feedback* e por conseguinte ajuste do modelo já durante a etapa de testes, o que acarreta perda de valor ou baixa aderência no modelo ao objeto estudado. Gaedke Nomura et al. (2021) apresenta um processo de feedback nas etapas de análise exploratória podendo haver reformulação do processo de entendimento da questão objetivo, na preparação de dados com reavaliação da etapa de exploração de dados e por fim no processo de acurácia podendo acarretar remodelagem da solução.

Figura 7: Ciclo de Vida em análise de dados



Fonte: Adaptado de Dietrich et al. (2015)

O sucesso de uma atividade vinculada a Ciência de Dados está diretamente ligado a integridade do ambiente digital e por conseguinte garantia de qualidade antes mesmo de qualquer desenvolvimento ou aplicação, corroborando a necessidade de um processo de governança bem estruturado e validado, o que permite um cenário de baixo custo e risco para projetos deste contexto e garante maximização na entrega de valor (Brous et al., 2020).

À luz dos argumentos fornecidos para garantia de boas práticas de governança, foi reunido na Quadro 4 elementos que compõem um processo de ciência de dados e quais atividades devem ser realizadas em cada etapa. O quadro sinaliza para qual material do portfólio bibliográfico foi mencionado cada elemento.

Quadro 4: Elementos de Ciência de Dados

Elemento	***Sinônimos em Inglês***	**Atividades**	**Referências**
Definição do Problema	*Problem Definition*	Definição de Objetivos, Mapeamento de Condições de contorno, *Stakeholders*	Chen et al. (2021), Gaedke Nomura et al. (2021), Saraee; Silva (2018)
Coleta de Dados	*Data Acquisition, Data Collection, Data Capture, Data Extraction, Data Entry, Signal Reception, Datasets, Data Integration*	Conexão bases de dados internas e/ou externas (Banco de Dados relacional e não relacional, ECI (*Event-based Communication Interface*), API (*Application Programming Interface*), arquivos locais, sensores etc.), criação de banco de dados, carga de dados, estruturação/organização de dados (*Data Warehouse, Online Analytical Processing*)	Gupta et al. (2021), Han; Trimi (2022) Watson et al. (2017), Qin et al. (2020), Lee et al. (2022), Chen et al. (2021), Lee; Tsai (2019), Gonzales et al. (2019), Chaniotakis et al. (2021), Gaedke Nomura et al. (2021), Marrapu et al. (2022), Yee et al. (2020), Prakash et al. (2020), Lei et al. (2020), Joshi et al. (2018) , Saraee; Silva (2018)

Elemento	***Sinônimos em Inglês***	**Atividades**	**Referências**
Limpeza de Dados	*Data Cleaning,*	Análise de qualidade dos dados, Remoção de registros incompletos e/ou inconsistentes (ruídos), Remoção de duplicados, Repopulação artificial de dados faltantes, Análise de Colinearidade, *Outliers*	Gupta et al. (2021), Han; Trimi (2022), Watson et al. (2017), Lee et al. (2022), Lee; Tsai (2019), Chaniotakis et al. (2021), Gaedke Nomura et al. (2021), Marrapu et al. (2022), Prakash et al. (2020), Joshi et al. (2018) , Saraee; Silva (2018)
Transformação de Dados	*Data Munging, Data Wrangling, Data Transformation, Rearranging Data, Data Segmentation, Data Reduction*	Ordenação, Filtragem, Agrupamento, Normalização, Segmentação, Decomposição, Definir Constructos, Criptografia de dados sensíveis, Suavizações de variáveis	Han; Trimi (2022), Watson et al. (2017), Lee et al. (2022), Chen et al. (2021), Gonzales et al. (2019), Chaniotakis et al. (2021), Marrapu et al. (2022) , Saraee; Silva (2018)
Análise Exploratória de Dados	*Data Explore, Statistical Analysis Feature Selection*	Identificação de dependência entre variáveis, Análise Descritiva (tendências e/ou dispersão), Análise Inferencial (representação populacional, teste de hipótese), Análise de causa e efeito, Correlações, Identificação de *outliers,*	Gupta et al. (2021), Philemotte (2020), Watson et al. (2017), Qin et al. (2020), Gonzales et al. (2019), Gaedke Nomura et al. (2021), Marrapu et al. (2022), Yee et al. (2020), Prakash et al. (2020), Correia et al. (2018) , Saraee; Silva (2018)

Elemento	***Sinônimos em Inglês***	**Atividades**	**Referências**
		Extração de constructos, Análise de qualidade, Redução de dimensionalidade	
Parametrização de Modelos	*Parameter Selection*	Particionamento de dados, Treinamento de modelos supervisionados, condições de contorno e definição de hiper parâmetros em métodos não supervisionados	Gupta et al. (2021), Philemotte (2020), Han; Trimi (2022), Watson et al. (2017), Qin et al. (2020), Lee et al. (2022), Chen et al. (2021), Lee; Tsai (2019), Gonzales et al. (2019), Chaniotakis et al. (2021), Gaedke Nomura et al. (2021), Marrapu et al. (2022), Yee et al. (2020), Lei et al. (2020), Joshi et al. (2018) , Saraee; Silva (2018)
Seleção de Modelo	*Model Selection*	Estatística, Modelos de *Machine Learning*, *Deep Learning* ou Heurísticas — Análise Descritiva, Análise Prescritiva, Análise Preditiva, Clusterização, Regressão	Gupta et al. (2021), Philemotte (2020), Han; Trimi (2022), Watson et al. (2017), Lee et al. (2022), Chen et al. (2021), Lee; Tsai (2019), Chaniotakis et al. (2021), Gaedke Nomura et al. (2021), Yee et al. (2020), Prakash et al. (2020), Lei et al. (2020),

Elemento	*Sinônimos em Inglês*	Atividades	Referências
			Joshi et al. (2018)
Análise de Precisão do Modelo	*Accuracy*	Métricas de desempenho dos modelos selecionados (MAPE, MAE, RMSE etc.), avaliação de falsos positivos, custo de processamento, tempo de processamento, eficiência do modelo, impacto dos resultados	Gupta et al. (2021), Philemotte (2020), Han; Trimi (2022), Qin et al. (2020), Lee et al. (2022), Chen et al. (2021), Lee; Tsai (2019), Chaniotakis et al. (2021), Gaedke Nomura et al. (2021), Marrapu et al. (2022), Prakash et al. (2020), Lei et al. (2020), Joshi et al. (2018) , Saraee; Silva (2018)
Lançamento	*Deployment*	Disponibilização e/ou automação da solução	Philemotte (2020), Watson et al. (2017), Qin et al. (2020), Lee et al. (2022), Gonzales et al. (2019), Gaedke Nomura et al. (2021)
Retreinamento	*Retraining, Feedback*	Retroalimentação de saídas do modelo para ajustes de parâmetros e adaptação da solução	Philemotte (2020), Han; Trimi (2022), Lee; Tsai (2019), Saraee; Silva (2018)
Visualização e Monitoramento	*Monitoring, Quality Control*		Philemotte (2020), Han; Trimi (2022), Yee et al. (2020), Joshi et al. (2018), Saraee; Silva (2018)

Segundo Yee et al. (2020), o fluxo de processamento de dados parte de cinco elementos essenciais: local de armazenamento, ferramentas de exploração e rápida prototipação, ferramentas de desenvolvimento, ferramentas de *machine learning* e de visualização. Nota-se aqui e corroborado pela baixa concentração de autores na seção de definição do problema do Quadro 4, a escassez de conteúdo e direcionamento no que concerne a primeira etapa desta modalidade de projeto.

Para Gaedke Nomura et al. (2021), o processo de definição do problema em questão normalmente é suprimido em grande parte dos materiais publicados, no entanto é destacado pelo autor como um dos pontos de maior alocação de tempo em detrimento de alinhamento com *stakeholders*, análise de protocolos e políticas de utilização de dados. Saraee e Silva (2018) apontam como erro comum em projetos de *Big Data* — em seu caso particular projetos geoespaciais — reside no fato de o pesquisador alocar esforços iniciais para coleta e análise de dados sem garantir a priori um plano adequado para o projeto de entendimento dos requisitos de negócio. O prejuízo da negligência neste caso é não poder validar os resultados por ausência de critérios de referência, falha na estimativa de custos e prazo de entrega, baixa satisfação do cliente e acurácia do projeto; os autores sugerem como ferramenta o método 5W1H (*What?, Who?, Where?, When?, Why? and How?*) por conta da facilidade de entendimento e aplicação.

Correia et al. (2018) lança um enfoque sobre dois contextos de aplicação no universo de dados e com objetivos complementares. Na primeira fase de seu estudo, o autor aloca os processos e recursos para análise de dados, processamento de dados e visualização de indicadores (*dashboards*) em *Business Intelligence* (BI); para grandes volumes ferramentas e processos de *Big Data*. Neste último caso, Saraee e Silva (2018) nos mostram que existem duas abordagens gerais: processamento em lote (*batch*) para dados coletados periodicamente ou processamento em tempo real (*stream*). Os autores sugerem como plataforma de computação paralela as ferramentas de código aberto Hadoop e Spark.

Operando sob escala de *Big Data*, ferramentas tradicionais tornam-se ineficientes ou até mesmo inviáveis em detrimento da alta volumetria, velocidade e variedade requerida (3Vs). Somado a qualidade, outro ponto de atenção reside na

estrutura e formato em que os dados estão sendo gerados e gravados — Estruturados, Semiestruturados e Não Estruturados. Primeiramente, a persistência de informações pode ser realizada em linhas e colunas em banco de dados relacional. Não obstante, dados semiestruturados apesar de apresentarem estrutura definida, não se encaixam no formato de tabela com linhas e colunas e tampouco são passíveis de armazenamento em banco de dados relacional. Dados não estruturados adicionam maior complexidade a rotina de projeto por não estar armazenados em sistemas de gerenciamento relacional e como tal permitir um consumo direto e com conexões unívocas. A exemplo de arquivos de imagens, textos livres e áudio que geralmente são persistidos em arquivos ou em formato serializado como BLOB (*binary large object*) ou CLOB (*character large object*) (Marungo et al. (2015) e Saraee; Silva (2018)).

As definições supracitadas mostram o nível de complexidade requerida no processo de armazenamento e por conseguinte maior capital humano e custos envolvidos. Brous et al. (2020) indica que nesta conjuntura *Data Lakes* permitem o rápido acesso a dados armazenados em sua forma bruta, mas em contrapartida lança um enfoque sobre importância no processo de gestão de dados no que tange qualidade, segurança e controle de acessos.

A atuação no contexto de Ciência de Dados “requer habilidades em diferentes linguagens de programação (Python, R, SQL, etc.), gestão de banco de dados, modelos preditivos, Algoritmos de *Machine Learning*, *Big Data* e comunicação” (Gaedke Nomura et al., 2021). O profissional com a designação de Cientista de Dados, tem o desafio constante de suportar todas as etapas de pré-processamento de dados seja por ausência de recursos para o setor de Engenharia de Dados ou por decisão de verticalizar o processo de geração de informação, o que pode gerar um custo excedente de até 80% no tempo de projeto Philemotte (2020) Watson et al. (2017). Corroborado por Marrapu et al. (2022), esta etapa é fundamental para o aumento de desempenho do conjunto de dados quando submetidos aos modelos de Ciência de Dados (principalmente em grandes volumes), o que gera a necessidade de ser realizado por um profissional com larga experiência e domínio do processo (Keller et al., 2020).

A etapa de limpeza dos dados consiste em basicamente da remoção

de atributos não relacionados ao objeto de estudo, substituição de valores nulos (médias, por exemplo) e eliminação de ruídos cuja permanência pode gerar resultados inconsistentes ou adicionar complexidade ao modelo. Na fase de exploração é possível identificar a relação entre variáveis que permitirão ao pesquisador na etapa de seleção de atributos definir a estratégia de menor custo em tempo e processamento (Prakash et al., 2020; Saraee; Silva, 2018).

O processo de normalização, comumente abordado em aplicações de algoritmos de *Machine Learning*, é utilizado como meio para eliminação de distorções dissidentes de dados em múltiplas escalas, o que gera maior convergência de melhora na precisão do modelo Chen et al. (2021).

Aprendizagem de máquina ou *Machine Learning* se apresenta como um segmento da ciência da computação cujo foco encontra-se na aplicação de algoritmos com capacidade de aprender e adaptar-se para tomada de decisão. Os algoritmos de *Machine Learning* estão subdivididos geralmente em dois grupos — Supervisionados e Não Supervisionados. No Primeiro caso os resultados de teste são conhecidos fazendo parte da modelagem de dados, enquanto no segundo o objetivo é identificar padrões e grupos no conjunto de dados sem que haja uma variável alvo. De forma mais ampla, no aprendizado supervisionado um conjunto de dados rotulados para treinamento são utilizados para estimar dados de saída em função do dados de entrada, enquanto para o não supervisionado os rótulos não são conhecidos e portanto apenas grupos podem ser formados com base e comportamentos semelhantes (Albayati; Altamimi, 2019; Chaniotakis et al. 2021).

Albayati e Altamimi (2019) descrevem *Data Mining* ou Mineração de Dados como um processo para extração de conhecimento sobre uma grande volumetria de dados com objetivo de buscar padrões, realizar validações e previsões para obtenção de informações úteis. Para os autores tal processo requer uma sequência de procedimentos e técnicas aplicadas sob a ótica de um conjunto de disciplinas — estatística, aprendizado de máquina, banco de dados, visualização e reconhecimento de padrões. Corroborado por Joshi et al. (2018), a Mineração de Dados surge como parte primordial no processo de inferir conhecimento sobre dados brutos provenientes de fontes diversas. De forma complementar, Marungo et al. (2015) utiliza-se do *framework* proposto por Fayyad et al. (1996) para o processo

de mineração de dados com intuito de descoberta de conhecimento (*knowledge discovery in databases* ou KDD) como base para elaboração de uma plataforma de Ciência de Dados para desenvolvimento de modelos de risco em tratamentos oncológicos:

> (1) compreensão do domínio do problema e dos trabalhos anteriores na área; (2) selecionar um conjunto de dados alvo; (3) limpeza e pré-processamento de dados; (4) redução e projeção de dados; (5) combinando o conhecimento objetivos de descoberta com uma abordagem de mineração de dados; (6) exploratório análise com hipótese e modelo testes; (7) mineração de dados; (8) interpretação dos resultados; e (9) agir sobre o conhecimento descoberto (Fayyad et al., 1996, p. 30).

Gaedke Nomura et al. (2021) aponta que a visualização de dados permite ao pesquisador identificar com mais clareza cenários de dados de alta complexidade, o que otimiza a utilização de recurso e eficiência no planejamento da pesquisa. Para Joshi et al. (2018, p. 2), “o processo de visualização de dados gera vários modelos descritivos e preditivos que ajudam a definir não apenas a natureza dos dados, mas também fornecem o modelo de funcionamento geral dos conjuntos de dados” cujos padrões que em outros formatos, texto por exemplo, não seriam facilmente identificados (Saraee; Silva, 2018).

Para Marrapu et al. (2022), somente o uso de visualização não garante o reconhecimento de padrões em dados como realizado pela aplicação de algoritmos. No entanto, permite um caminho mais acessível e eficiente para humanos no processo de análise e exploração de dados. Os autores destacam ainda a importância de desenvolvimento conjunto entre equipe de Ciência de Dados e equipe atuante no contexto estudado, uma vez que os dados podem ser armazenados em diferentes formatos e padrões, cuja complexidade pode ser mitigada em um esforço coordenado, uma vez que “as pesquisas em Ciência de Dados devem ser gerenciadas com eficiência para obter melhores resultados. Para resultados ideais, os desenvolvedores de pesquisa precisam ser competentes e ter amplo conhecimento sobre pesquisa” Marrapu et al. (2022, p.5).

Brous et al. (2020) apresentam um conjunto de proposições que relaciona governança e o sucesso em projetos de Ciência de Dados: 1. Propõe um “preço” para conjunto de dados de forma a dar maior transparência e entendimento dos dados como ativos da organização; 2. Monitoramento e controle da qualidade de

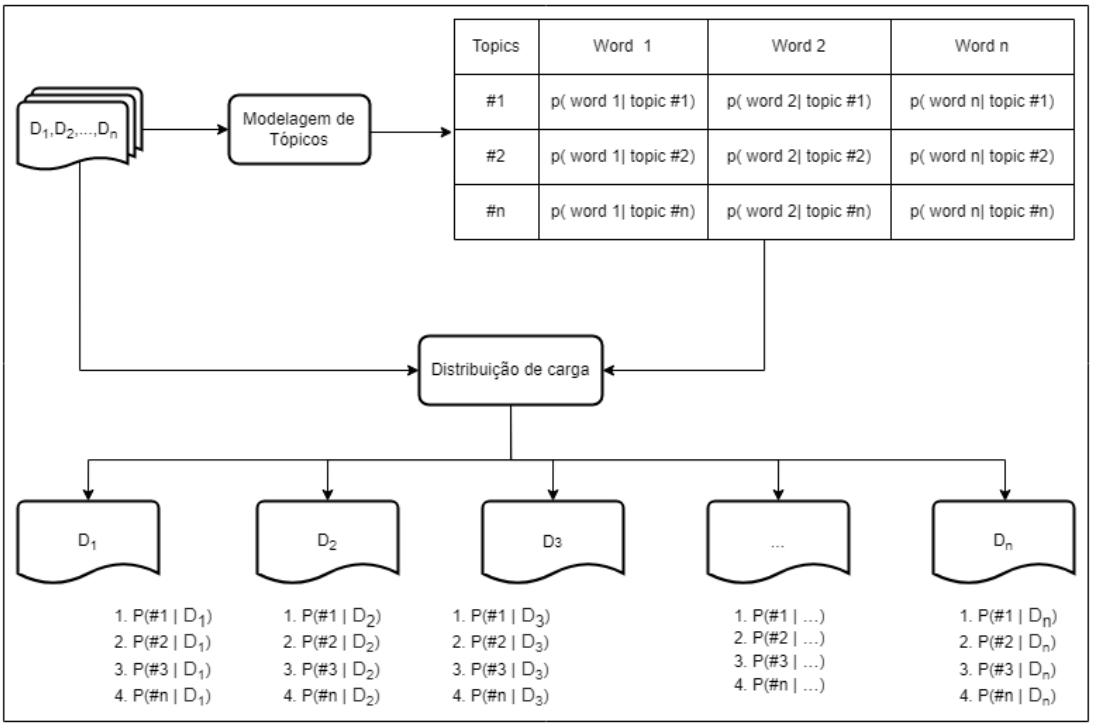
dados, o que permite maior conscientização e economia de custos; 3. Cumprimento dos aspectos legais de utilização de dados. Em detrimento da alta volumetria de dados gerados atualmente, os desafios das práticas de governança crescem sobremaneira para que as organizações consigam garantir adequação do consumo de dados com regulamentações legais, diretivas, políticas de uso e procedimentos internos dentro das fronteiras de cada um dos elementos mencionados.

2.2 GESTÃO DA INFORMAÇÃO EM PEQUENAS E MÉDIAS EMPRESAS

A revisão de literatura tradicional busca, por meio da aplicação de filtros sucessivos, reduzir a grande quantidade de estudos obtidos em bases de dados acadêmicas a números viáveis de serem avaliados em sua totalidade para extração de conteúdo. A produção científica cresce a uma taxa de 9% ao ano chegando ao dobro da quantidade em cerca de oito anos, o que gera o desafio no processo de organização e estruturação do conhecimento em relação a determinado tema Patterson et al. (2017), sem mencionar o risco de trabalhos de relevância ao estudo não integrarem o portfólio final conforme citam OMara-Eves, J. et al. (2016). No âmbito organizacional, estima-se que cerca de 80% da fatia de informações relevantes seja de dados não estruturados ou semiestruturados, texto por exemplo (Palanivel Rajan D. et al., 2020).

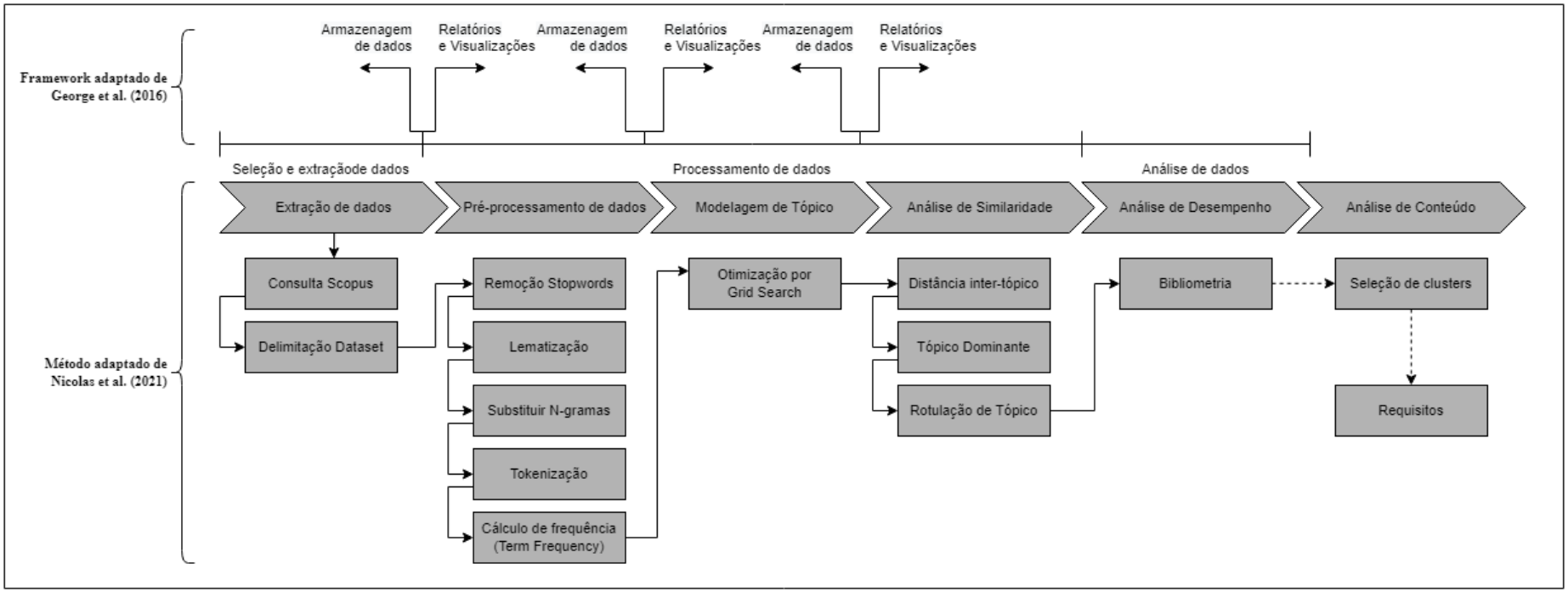
De forma a minimizar os efeitos supracitados e analisar uma quantidade maior de publicações, este estudo incorpora técnica de mineração de texto sobre os 4.393 documentos do terceiro eixo (Tabela 1), tanto para seleção de portfólio bibliográfico quanto para extração de conteúdo. O algoritmo utilizado, Alocação Latente de Dirichlet (DLA), trata-se uma abordagem não supervisionada para identificação de padrões e estruturas ocultas ou latentes (Modelagem de tópicos) em dados não estruturados como documentos de texto (BLEI et al, 2002), rotulando e interpretando o conteúdo baseado nas palavras contidas no corpus de documentos (Tavana et al. , 2020). Em linhas gerais a aplicação do algoritmo busca identificar os tópicos latentes analisando as palavras mais frequentes presentes em cada tópico e em seguida, agrupar os documentos em relação à proporção que apresentam de cada tópico como ilustrado na Figura 8.

Figura 8: Esquema de aplicação do algoritmo LDA



Com objetivo de garantir a reprodutibilidade do estudo em outros contextos, George et al. (2016) propuseram um *framework* de trabalho com as principais etapas de um projeto de ciência de dados e *Big Data* dividido em cinco etapas: 1) Coleta de dados, 2) Armazenamento de dados, 3) Processamento de dados, 4) Análise de dados e 5) Relatórios e Visualizações. No entanto, o contexto de utilização apresentado para este estudo não prevê a sustentação e manutenção da solução com entradas e saídas sendo atualizadas por eventos e, portanto, não apresenta a necessidade de uma camada de armazenamento (3) e uma de relatoria (5) dedicadas. Ambas as camadas estarão difundidas nas demais como no modelo apresentado por Nicolas et al. (2021) e esquematizado na Figura 9.

Figura 9: Método de revisão de literatura utilizando LDA



2.2.1 Extração de Dados

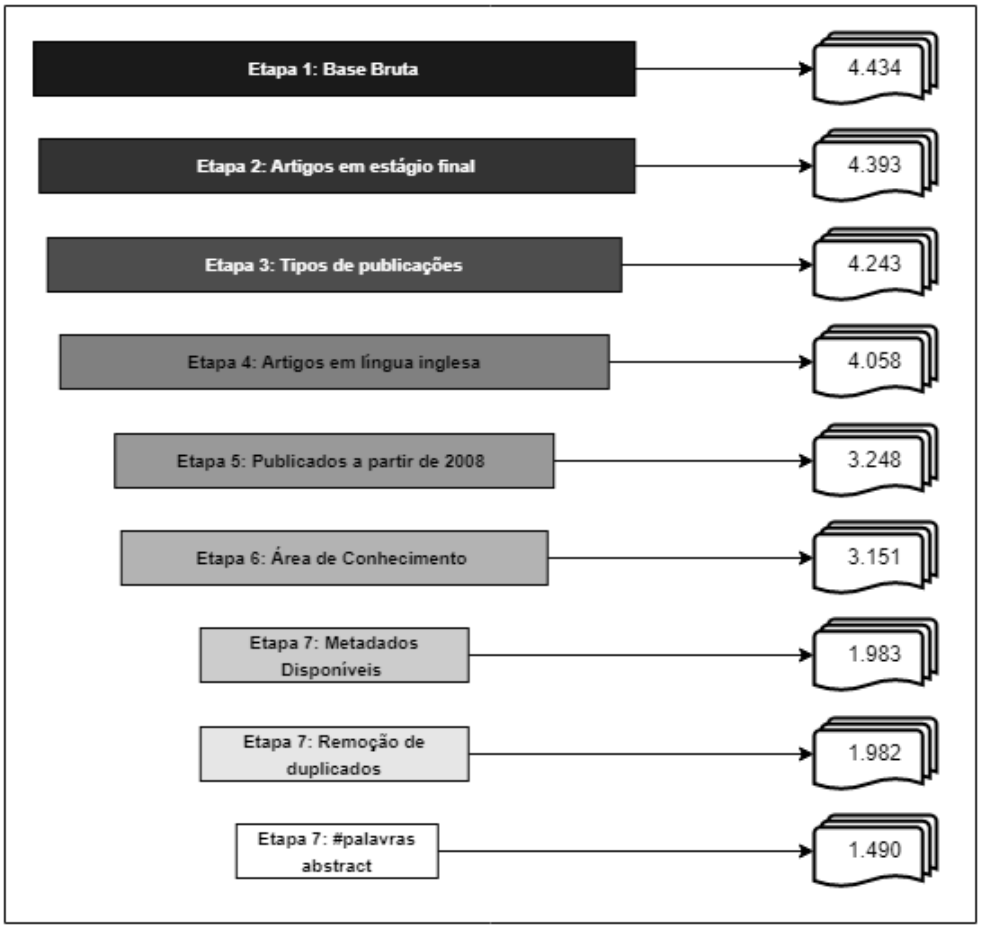
A primeira etapa do processo de mineração de texto prevê a seleção da base de dados e coleta de dados brutos. Como já estabelecido na seção de delimitação do estudo, a base de dados Scopus foi utilizada e a base bruta de materiais obtidas por meio da *query* de busca relatada no eixo *SMEs & Information Management* da Tabela 1.

Dos 4.434 materiais resultantes, 4393 estavam em estágio final de publicação e seguiram à Etapa 3, na qual apenas documentos cujo tipo está listado como *Conference Paper*, *Article*, *Review* ou *Conference Review* foram mantidos — 4.243 documentos. A quarta etapa optou-se por materiais em língua inglesa dado à disponibilidade de pacotes computacionais para esse idioma, enquanto para outras línguas proceder-se-ia à tradução a priori para a língua inglesa e só então utilizar um dos *frameworks* disponíveis; 4.058 artigos restantes.

Como filtro temporal, o ano de 2008 foi selecionado como ponto de partida por já apresentar volumetria média de publicações estável referentes ao tema, sem grandes flutuações. A quantidade residual cai para 3.248 documentos. Em relação às áreas de conhecimento, foram selecionadas *Computer Science*, *Business, Management and Accounting*, *Decision Sciences*, *Engineering*, *Social Sciences*, *Mathematics*, *Economics*, *Econometrics and Finance*, *Environmental Science*, *Materials Science*, *Multidisciplinary*, *Chemical Engineering* por estarem em consonância com o objeto de estudo envolvendo a temática de gestão e Ciência de Dados (3.151 artigos).

Relativo ao processo de limpeza dos dados, 1.982 documentos restaram após retirada de materiais cujos metadados principais não estavam disponíveis (Nome do Autor, *Abstract*) e remoção de duplicados. Como mencionado anteriormente, modelos de aprendizagem de máquina operam sobre uma volumetria considerável e interfere na precisão dos resultados. Por esta assertiva, optou-se por remover artigos cujo número de palavras estivesse abaixo do primeiro quartil (25%), chegando ao número final da etapa de coleta em 1490 documentos. A Figura 10 reúne cada etapa do processo descrita nesta seção e o fluxo de materiais ao longo do processo.

Figura 10: Funil coleta de documentos *SME & Information Management*



2.2.2 Pré-Processamento de Dados

Esta seção do método visou transformar o conjunto de dados de forma a adequar às condições de aplicação do modelo LDA. Em primeira instância, procede-se a remoção de palavras sem valor explicativo como preposições, artigos (definidos e indefinidos), números e pontuações — a estas palavras dá-se a definição de *stopwords*. Na sequência, com objetivo de remover variações da mesma palavra por flexão de gênero e número, reduziu-se cada a palavra a seu lema e a este processo dá-se o nome de lematização.

Outro processo largamente utilizado é stemmatização, na qual cada palavra é reduzida a sua raiz. Apontado por Divya et al. (2020), o *stemm* (raiz) não se configura em um texto de linguagem real como no caso do lema. Os autores destacam ainda que em virtude do “Conhecimento linguístico profundo necessário

para formar o glossário que permite ao algoritmo buscar a parte significativa da palavra no processo de lematização, o resultado será mais preciso" Divya et al. (2020, p. 356). Frente a argumentação, este trabalho utilizou somente o processo de lematização.

> Lematização e Stemming desempenham um papel importante no processamento de texto e linguagem natural. Ambos criam a parte base dos textos-palavra flexionados. A diferença reside no fato de que a raiz não é um texto de palavra real, enquanto o lema é um formato de texto de linguagem real Divya et al. (2020, p. 357).

Um processo comumente utilizado é a justaposição de palavras que ocorrem em sequência e só passam a fazer sentido se avaliadas em conjunto, a exemplo das palavras *Data* e *Science*, que mesmo havendo possibilidade de aplicações separadas, quando utilizadas justapostas passam a representar um terceiro contexto, Ciência de Dados. A combinação pode dar-se por duas ou mais palavras e por este motivo dá-se a definição de N-gramação, com N > 1. Um bigrama representa a junção de duas palavras, um trigrama de três e assim sucessivamente.

Este estudo realiza a mineração de texto sobre os resumos dos materiais. Ao passo que os insumos desta abordagem não chegam a integrar um escopo considerável a ponto de uma locução como "*Data Science*" ocorrer mais de uma vez, dessa forma, o treinamento para esta etapa foi realizado sobre um documento único resultante da concatenação de todos os 1.490 resumos. Em seguida, após a identificação dos N-gramas (a partir de cinco ocorrências) realizou-se a substituição dentro de cada um dos resumos.

Na próxima fase, cada resumo representado por uma sequência única de texto foi dividido em fragmentos (*tokens*) cuja quebra foi realizada na ocorrência de espaço e apenas tokens com três ou mais letras foram considerados com entrada para a próxima etapa.

Para aplicação do algoritmo LDA, é preciso que cada token seja incorporado a entrada seguido da sua frequência dentro da sentença (resumo, neste caso). Pode-se optar por incorporar frequência e ordem de ocorrência (*Term Frequency Inverse Document Frequency* ou TF-IDF) (Roque et al., 2019) ou somente a frequência (*Term Frequency* ou TF) (Nicolas et al., 2021). Com a

aplicação do método de modelagem de tópicos busca-se neste trabalho como principal saída o agrupamento dos documentos (clusterização) e a análise de conteúdo será realizada a posteriori. Dessa forma, a ordem de ocorrência de cada *token* dentro da sentença passa a não apresentar contribuição significativa e, portanto, apenas o TF foi utilizado.

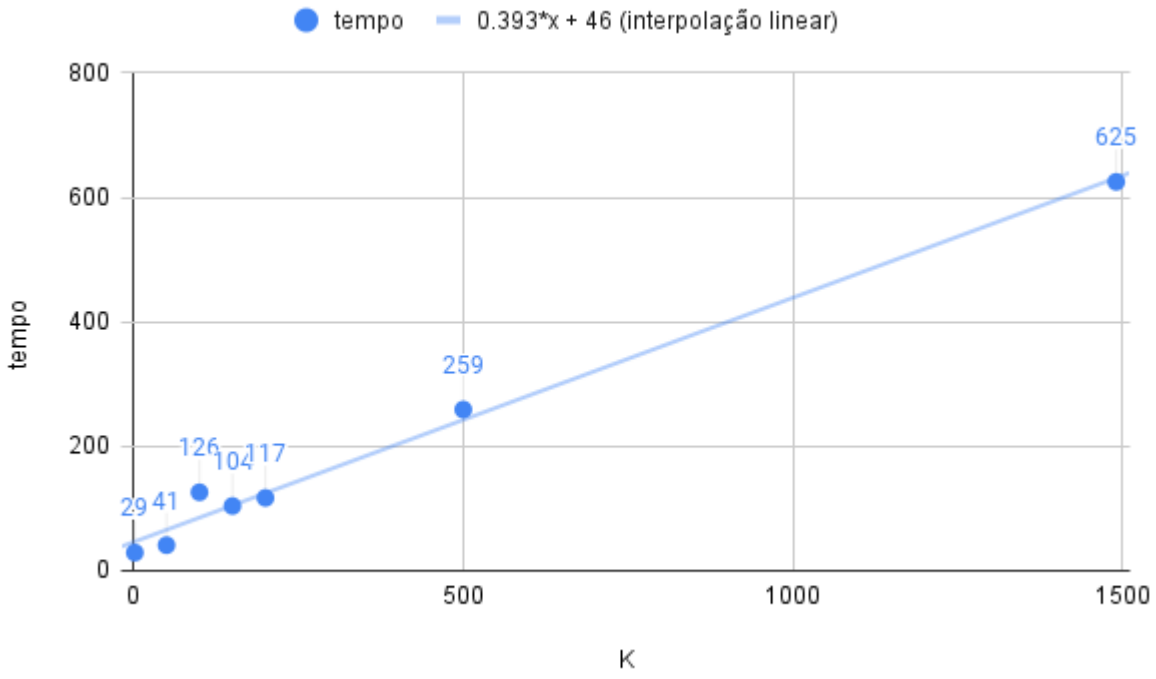
2.2.3 Modelagem de Tópicos

Segundo Roque et al. (2019), o algoritmo LDA utiliza inferência Bayesiana para estimar a distribuição das palavras por tópicos com base nas palavras contidas em cada documento, havendo necessidade de inicialização do número de tópicos K por parte do pesquisador, uma vez que seu valor é desconhecido a priori e, dado a natureza não-supervisionada da modelo, sua medida direta não é aplicável. O autor utiliza a divergência de Kullback-Leibler e o Método de Maximização de Expectativa e ressalta também a importância de analisar a qualidade dos tópicos gerados.

Lee e Lim (2021), utilizaram três métricas para inicialização da medida: similaridade de cossenos, similaridade de Jaccard e divergência de Kullback-Leibler; Asmussen & Møller (2019) utilizou a minimização da perplexidade. Neste estudo, entende-se que o número máximo de tópicos deve aproximar-se do número total de áreas de conhecimento (54) presentes no portfólio (APÊNDICE N). A heurística adotada para este limite deve-se em grande parte ao tempo computacional requerido ao utilizar-se o número total de documentos como número de iterações a serem realizadas. Com apoio da interpolação linear (Figura 11) chegou-se à equação de previsão ($T_i = 0.393 * K_i + 46$, com T_i em segundos) para o tempo computacional para cada K, com K entre 2 e 1490.

Dessa forma, caso houvesse decisão por iterar K até o número limite de documentos (1490), chegar-se-ia a aproximadamente seis dias de processamento com a utilização do equipamento disponível e especificado na Seção 3.1. Em contrapartida, optou-se por aplicar o algoritmo iterativamente para $1 < K < 201$, coletar métricas de coerência e perplexidade para qualitativa da formação de grupos.

Figura 11: Interpolação linear para estimativa de tempo de processamento para cada K



Deve-se ressaltar que $K_{inicial}$ deve ser maior que 1, caso contrário confunde-se com o próprio portfólio. Em segundo lugar, K_{final} igual a 200 (aproximadamente 4 vezes o número de Áreas de Conhecimento) encontra-se dentro de uma faixa viável de processamento (5h) e nos permite visualizar com maior segurança o comportamento das métricas coletadas.

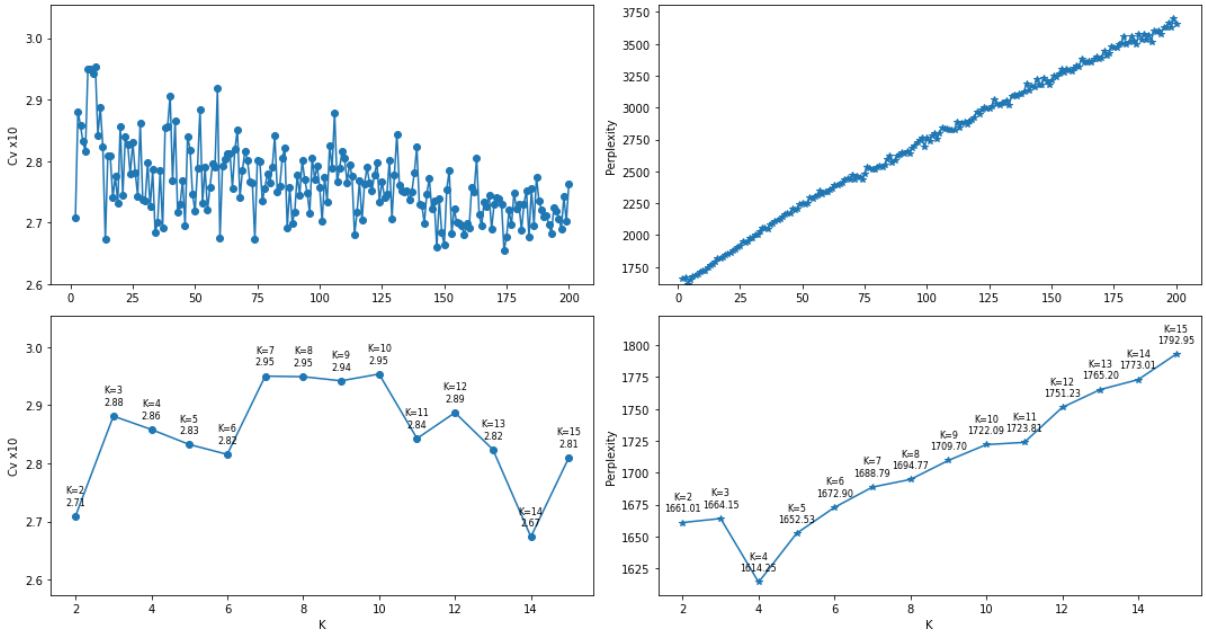
2.2.4 Análise de Similaridade

Como apresentado por (Nicolas et al., 2021), conforme aumenta-se o número de tópicos K, garante-se a máxima coerência intra tópico (C_v), dado que no limite o tópico converte para o próprio documento. Da mesma maneira, entende-se que a perplexidade (*perplexity*) aumenta em virtude do esforço necessário para obter incrementos positivos de coerência. Este estudo utilizou como heurística, incrementar K enquanto houver diferenças positivas de C_v para o menor valor possível de perplexidade.

A Figura 12 evidencia que para K > 6 e K < 11 temos a máxima coerência para o intervalo entre 2 e 200 tópicos. A partir deste ponto, gera-se um

maior consumo de recursos (aumento da perplexidade) sem que para isso haja ganhos de coerência. Logo, é de se esperar que o número ideal de K esteja compreendido entre 7 e 10.

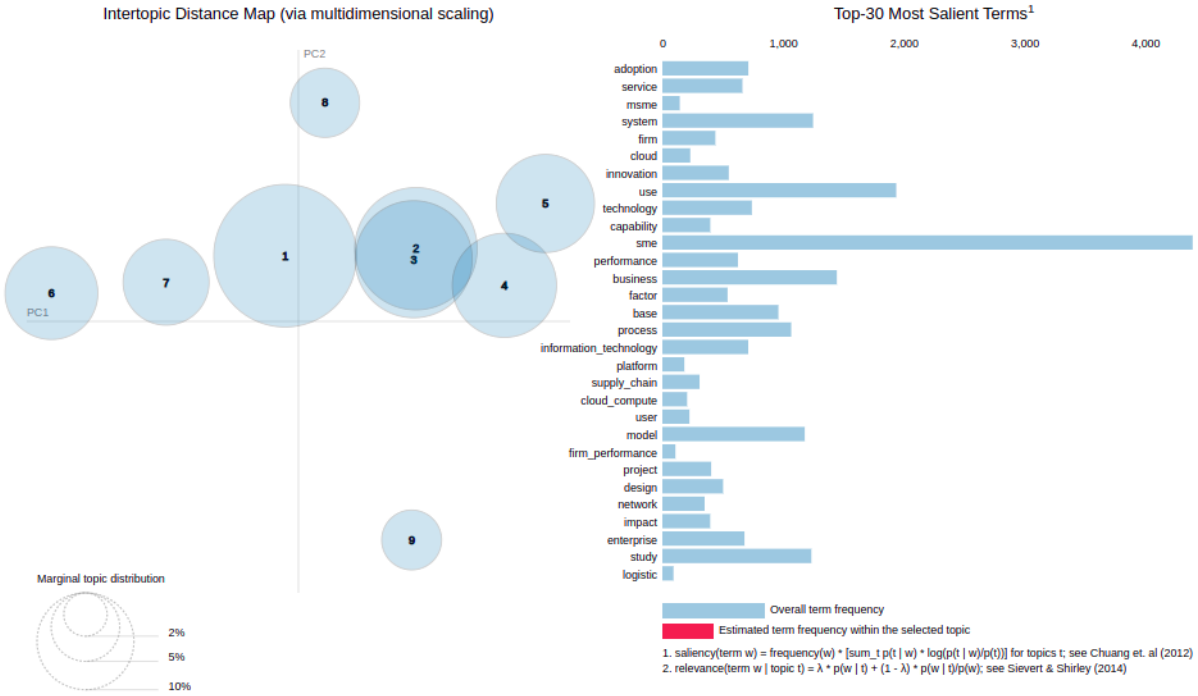
Figura 12: Métricas para seleção do número de tópicos (K)



Outro fator adicionado por (Nicolas et al., 2021) e corroborado por (Sievert; Shirley, 2015) é que quanto maior a distância inter-tópico, maior a coerência percebida entre os documentos do agrupamento. Durante a etapa de análise, além da métrica de coerência, foi utilizado para seleção do número de tópicos mais adequado a biblioteca pyLDAvis em linguagem Python (também disponível na linguagem R). O pacote auxilia na análise das distâncias inter-tópicos como suporte a decisão do número de tópicos ideal para o corpus em questão.

A Figura 13 evidencia que para K=9 tem-se a menor sobreposição de clusters e, portanto, a máxima coerência, o que valida a heurística supracitada que informa a priori que o número ideal de tópicos está compreendido entre 7 e 10. Ainda nesta fase, decidiu-se por consolidar os tópicos 2 e 3 por tratar-se de uma sobreposição completa. A segunda saída fornecida pelo pacote computacional Gensim é a distribuição de cada tópico dentro de cada documento. Esta informação permite identificar o nível de dispersão do conteúdo das publicações que compõem o corpus.

Figura 13: Distâncias Inter tópicos (K = 9)



Contrariamente ao que acontece em artigos de notícia, artigos científicos em geral visam um aprofundamento maior em relação a determinado tema (Lee; Lim, 2021). Logo, objetivando utilizar materiais com maior densidade do tema abordado, adotou-se como tópico final do documento o tópico de maior carga e selecionou-se apenas documentos que apresentaram probabilidade maior ou igual a 80% de um determinado tópico.

A rotulação de tópicos é a fase que define a abordagem de mineração de texto para classificação e seleção de documentos como um processo qualitativo-quantitativo. A etapa de processamento fornece a carga de cada palavra sobre o fator, no entanto, a interpretação do constructo depende diretamente de uma avaliação qualitativa. Dessa maneira, assim como Wagire, Rathore e Jain (2019), a rotulação dos tópicos gerados neste estudo foi realizada por pares em paralelo e em caso de conflitos, prossegue-se à avaliação de um terceiro agente. O objetivo da avaliação distribuída é diminuir o efeito da subjetividade na interpretação dos fatores. A Tabela 3 reúne a distribuição dos documentos para cada tópico já com os rótulos devidamente atribuídos.

Tabela 3: Distribuição de documentos por tópico dominante

Rótulo	Rótulo PT-BR	qtde	%
Sustainability and Business Development	Sustentabilidade e Desenvolvimento Organizacional	26	3.82%
Hardware and Software Applications	Aplicações de Hardware e Software	232	34.11%
Information Architecture	Arquitetura de Informação	53	7.79%
Collaborative Network Information	Rede Colaborativa de Informação	160	23.52%
Digitalization	Digitalização	61	8.97%
Business Performance	Desempenho organizacional	72	10.58%
Decision Suport	Suporte a Decisão	31	4.55%
Optmization Technologies	Otimização	45	6.61%
Total		**680**	**100%**

2.2.5 Análise de Desempenho

De posse dos tópicos rotulados na seção 2.2.4, é possível selecionar os assuntos mais aderentes aos objetivos da agenda de pesquisa no que tange a restrições por parte das Pequenas e Médias Empresas referente à Gestão da informação e dessa forma estabelecer critérios para seleção de elementos para o FCD — *Information Architecture*, *Business Performance, Decision Suport, Collaborative Network Information* e *Digitalization*.

Com objetivo de minimizar os efeitos de publicações recentes não integrarem o portfólio final acerca do tema em relação a métrica de citações, convencionou-se que o critério de inclusão será realizado em relação ao periódico sob a métrica Scopus *Source Normalized Impact per Paper* (SNIP). Segundo a Agência de Bibliotecas e Coleções Digitais (2016), o indicador avalia o impacto de citações por revistas pondera sob um campo de assunto, permitindo comparações entre periódicos em áreas de conteúdo diferentes. Neste estudo, foram tomados apenas documentos com SNIP maior que zero e após uma distribuição quartílica dos artigos remanescentes, selecionou-se apenas os materiais cujo indicador encontrava-se acima do terceiro percentil (SNIP > 1.349). A Tabela 4 apresenta a distribuição do portfólio remanescente.

Tabela 4: Distribuição de tópicos pós filtro de SNIP

Rótulo	Rótulo PT-BR	qtde	%
Information Architecture	Arquitetura de Informação	4	6.15%
Collaborative Network Information	Rede Colaborativa de Informação	22	33.84%
Digitalization	Digitalização	18	27,69%
Business Performance	Desempenho organizacional	20	30.76%
Decision Suport	Suporte a Decisão	1	1.53%
Total		**65**	**100%**

Uma análise dos 65 documentos remanescentes corrobora o potencial de contribuição e qualidade do portfólio frente ao objetivo específico de mapear o fluxo da informação organizacional, cuja maior frequência está em áreas de concentração com ênfase em áreas de conhecimento referente à Gestão de TI — "*stratey and management*", "*information systems" e "management information systems"*. A Tabela 5 reúne os periódicos com mais de uma ocorrência na base e evidencia o nível de concentração da amostra uma vez que cerca de 40% dos documentos estão alocados em 14% dos periódicos presentes.

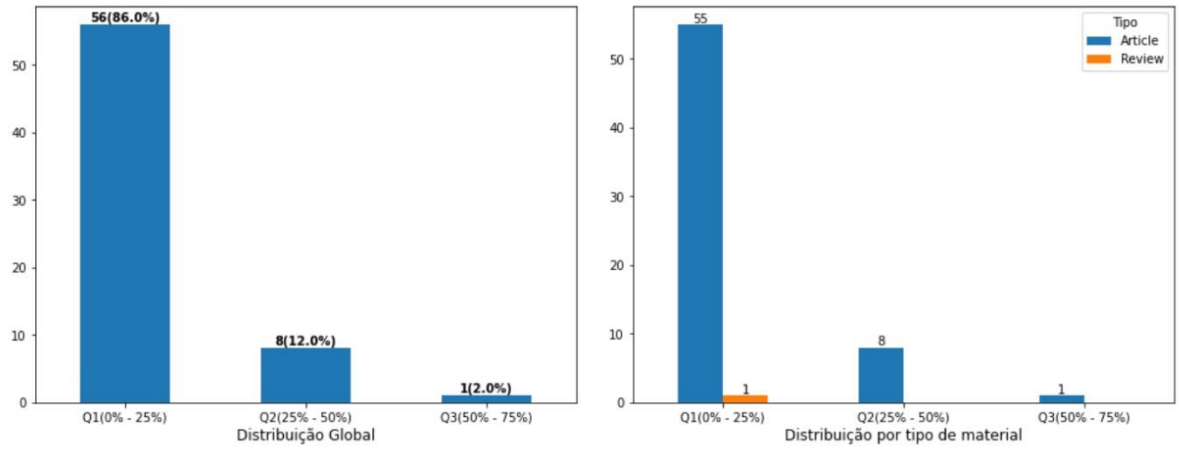
Tabela 5: Número de Publicações do portfólio por Periódico e Área de Conhecimento

Journal	*Área*	#documentos	%documentos
Industrial Management and Data Systems	*computer science applications;industrial and manufacturing engineering;industrial relations;management information systems;strategy and management*	6	9%
International Journal of Information Management	*artificial intelligence;computer networks and communications;information systems;information systems and management;library and information sciences;management information systems;marketing*	5	8%
Technological Forecasting and Social Change	*applied psychology;business and international management;management of technology and innovation*	3	5%
International Journal of Entrepreneurial Behaviour and Research	*business, management and accounting*	2	3%
Journal of Knowledge Management	*management of technology and innovation;strategy and management*	2	3%
Computers in	*computer science; engineering*	2	3%

Journal	*Área*	#documentos	%documentos
Industry			
European Journal of Information Systems	*information systems;information systems and management;library and information sciences;management information systems*	2	3%
Journal of Enterprise Information Management	*decision sciences; information systems;management of technology and innovation*	2	3%
Information Systems Frontiers	*computer networks and communications;information systems;software;theoretical computer science*	2	3%
Total		26	40%

Outro ponto avaliado em relação ao desempenho dos materiais foi sua posição na distribuição quartílica do CiteScore 2020. Da Figura 14, identifica-se que 86% dos artigos encontram-se no primeiro percentil e apenas um documento no terceiro quartil. Importante ressaltar também a prevalência de artigos de periódico nesta etapa, em contraponto ao portfólio da seção 2.1 na qual artigos de conferência também estavam presentes.

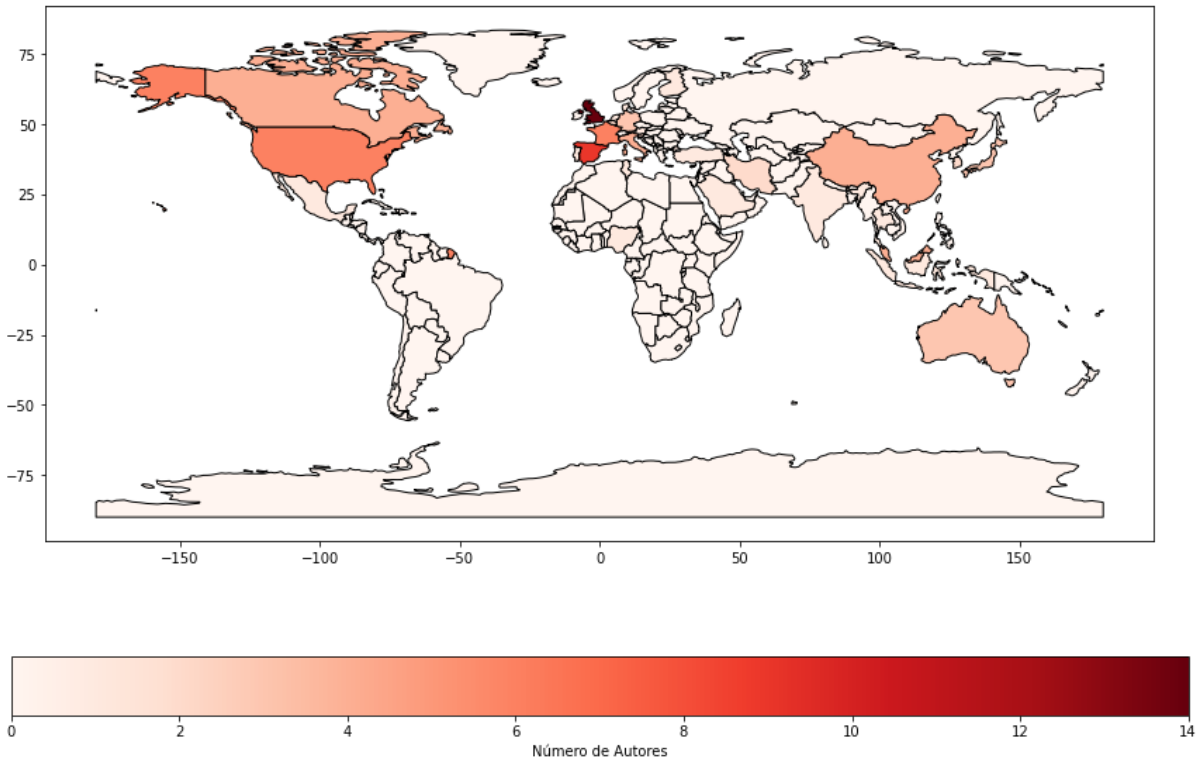
Figura 14: Número de documentos por quartil do CiteScore 2020 Gestão da Informação



Em relação à filiação dos autores, nota-se uma maior participação do Reino Unido com o dobro da segunda posição, os Estados Unidos com seis pesquisadores. Outrossim, como já evidenciado na Figura 4, a Figura 15 aponta que a Europa ainda se configura com um centro de irradiação de conteúdo em relação

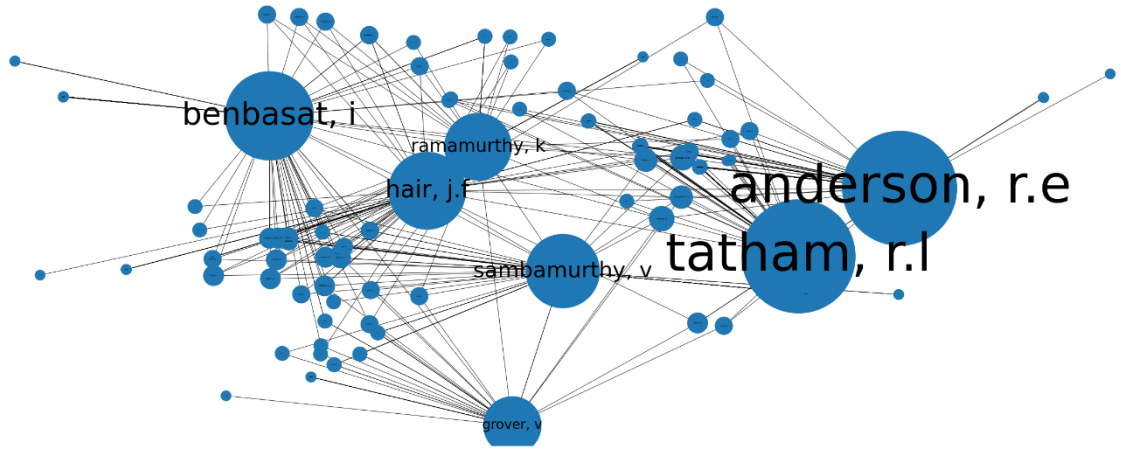
ao contexto abordado neste estudo, principalmente no que tange a conjuntura de Pequenas e Médias Empresas e iniciativas de desenvolvimento desta frente. Autores brasileiros não integraram esta seleção de documentos.

Figura 15: Distribuição geográfica de filiações do portfólio bibliográfico Gestão da Informação



Similar à seção 2.1.2 (Análise de Desempenho), também se buscou aqui identificar os autores de maior expressão acadêmica por meio de suas coocorrências com referências do mesmo portfólio. A seleção dos materiais com presença de tais autores em suas referências possibilita uma análise direcionada aos documentos de maior relevância. No entanto, de forma a manter a coerência com o processo de seleção da seção supracitada, deliberou-se por não utilizar esta análise como critério de inclusão e exclusão de materiais. O relacionamento dos autores e ilustrado pela análise de rede na Figura 16 e as publicações com marcação da presença das principais referências na Quadro 5.

Figura 16: Análise de rede em relação as referências do portfólio bibliográfico Gestão da Informação



Quadro 5: Ocorrência de autores representativos no portfólio bibliográfico Gestão da Informação

Referências *	**A**	**B**	**C**	**D**	**E**	**F**	**G**
Dibrell; Davis, Peter S; et al. (2008)						x	
Francalanci; Morabito (2008)	x	x	x		x	x	
Omerzel; Antončič (2008)		x	x				
Nguyen (2009)	x						
Fantazy et al. (2009)						x	
Lopez-Nicolas; Soto-Acosta (2010)		x	x				
Harrigan et al. (2011)		x	x				
Wilkin (2012)					x	x	
Harrigan et al. (2012a)		x	x				
Harrigan et al. (2012b)		x	x				
Oni; Papazafeiropoulou (2014)							x
Dutot et al. (2014)							x
Naudé et al. (2014)		x	x				
Ghobakhloo; Tang (2015)	x				x		x
Ramayah et al. (2016)	x	x					x
Dyerson et al. (2016)	x					x	
Cassia; Magno (2022)							x
Hayajneh et al. (2022)				x			
Soluk; Kammerlander (2021)						x	
Chabbouh; Boujelbene (2020)		x					
Pelletier; Cloutier (2019)	x						
Singh et al. (2019)						x	
Chan et al. (2019)					x	x	x
JosephNg (2018)						x	
Puklavec et al. (2018)	x				x		
Awa et al. (2017)	x				x		x
Neirotti; Raguseo (2017)					x	x	x
Yunis et al. (2017)		x	x				

*A=Benbasat, I; B=Anderson, R.E; C=Tatham, R.L; D=Hair, J.F; E=Sambamurthy, V; F=Grover, V; G=Ramamurthy, K.

2.2.6 Análise de Conteúdo: Requisitos para Adoção de Tecnologias por PMEs

Em relação aos resultados preliminares, Khalil e Belitski (2020) relatam que a adoção de Sistemas de Informação está relacionada ao aumento de desempenho organizacional, aumento na interface com ambiente externo e busca de vantagens competitivas e flexibilidade frente aos demais agentes de mercado. Para os autores, investimentos em computação em nuvem permitem maior fortalecimento do alinhamento tecnológico nos domínios gerenciais e estratégico.

Apontado por Wilkin (2012), a ineficiência nas competências digitais e experiência do quadro colaborativo em tecnologias utilizadas pela organização configuram impedimentos para uma governança eficiente de TI. O autor demonstra que cerca de 78,9% do efetivo tem seis anos ou mais experiência com TI, mas esse número cai para 37,8% se for considerado apenas na própria organização e com foco majoritariamente em atividades operacionais. Esse efeito pode ser mitigado por meio da criação de agendas digitais e políticas públicas para desenvolvimento educacional em competências digitais Khalil e Belitski (2020), proporcionando ao decisor maior percepção da entrega de valor do seu negócio por meio TI.

O Quadro 6 reúne as principais lacunas de pesquisa em práticas de governança e barreiras na implementação de práticas de governança de TI em PMEs. Entender qual a estrutura de governança e qual a influência de seus elementos é um processo fundamental para que PMEs possam direcionar seus recursos de forma racional. Keller et al. (2020) destaca a importância do delineamento de protocolos de governança de dados e como a abordar precauções éticas é fundamental para a eficácia na ciência de dados desde a coleta até sua interpretação e conversão em informação.

Quadro 6: Barreiras e Lacunas de Pesquisa em Governança de TI

	Descrição
Lacunas de pesquisa	Análise cruzada de impacto em diferentes países, domínios de negócio, setores, porte, modelos de governança; potencial de desenvolvimento do domínio Estratégico e Gerencial da governança de TI; estudo longitudinal do impacto da governança de TI no desempenho de PMEs; Relação entre práticas de governança de TI e entrega de

	valor.
Barreiras	Baixa disponibilidade e limitação de lideranças técnicas; baixo investimento em digitalização; baixa disponibilidade de dados longitudinais; concentração de decisões sobre gestor/proprietário; foco em atividades operacionais; Baixa percepção de valor; políticas públicas.

A abordagem descrita em Khalil e Belitski (2020) divide a estrutura de governança em três mecanismos: estrutura, processo e relacional. Cada mecanismo apresenta um ou mais níveis de interação com os três domínios organizacionais — estratégico, gerencial e operacional. Neste estudo, apenas a influência da TI sobre os domínios organizacionais será abordado e como sugestão para estudos futuros a ralação entre TI e mecanismos de governança.

O domínio estratégico está fortemente ligado ao alinhamento estratégico da organização e focado na maximização da entrega de valor do negócio Khalil e Belitski (2020). Neste âmbito, estudo anteriores mostram ainda uma forte relação entre a adoção de TI e processos de inovação, tanto na etapa de desenvolvimento de produto como elemento associativo entre inovação e desempenho Dibrell et al. (2008). Relativo às lacunas pesquisa identificadas neste estudo (Quadro 7) referente a este domínio e a agenda de pesquisa, como ponto mais frequente se destacam as análises cruzadas (Jamali et al. (2015), Khan et al. (2019), Neirotti et al. (2013), Neirotti e Raguseo (2017) e Yunis et al. (2017)), uma vez que em grande parte os trabalhos são realizados em estudos de caso único, não refletindo a diferença de impacto em cenários diferentes.

Quanto ao processo de pesquisa, Kmieciak et al. (2012) e Walsh et al. (2010) identificam a análise de causalidade como ponto a ser explorado, uma vez que as análises de correlação ou análises qualitativas fornecem apenas uma medida indireta de como as variáveis se relacionam limitando as possibilidades de extrapolação. O desenvolvimento e utilização de tecnologia está muitas vezes associado a ganhos intangíveis em comunicação, controle e capacidade de inovativa. Neste contexto, este trabalho aponta duas possibilidades de pesquisa:

impacto da percepção de valor por parte do empreendedor na utilização de TI (Yunis et al., 2017) e avaliação da eficiência de gastos em TI e eficácia na entrega de valor (Jamali et al., 2015).

Para ambientes de alta tecnologia, existe potencial de exploração do impacto de tecnologias "*as a service*" para remoção de barreiras (Neirotti et al., 2013) no alinhamento estratégico a exemplo da turbulência tecnológica (Pratono , 2017), que somado a restrição de custos (Neirotti et al., 2013) atrasa a atualização constante da infraestrutura digital de PMEs.

Quadro 7: Lacunas de pesquisa no domínio estratégico em governança de TI

Referência	**Lacuna de Pesquisa**
Miyamoto, 2015	Impacto das estratégias de TI à luz de forças competitivas
Chege et al. (2020), Jamali et al. (2015), Kmieciak et al. (2012), Lányi et al. (2021), Peña-Vinces et al. (2012) e Qureshil et al. (2009)	Estudo longitudinal em relação a variáveis de desempenho
Jamali et al. (2015), Khan et al. (2019), Neirotti et al. (2013), Neirotti e Raguseo (2017) e Yunis et al. (2017)	Análise cruzada em relação à entrega de valor em diferentes ambientes de competição, geográficos, econômicos e de maturidade tecnológica
Peña-Vinces et al. (2012);	Impacto de políticas públicas na entrega de valor de TI;
Kmieciak et al. (2012) e Walsh et al. (2010)	Análise de causalidade entre o uso de TI e a entrega de valor
Khan et al. (2019)	Delineamento da estratégia de TI em relação à requisitos externos a organização
Yunis et al. (2017)	Análise perceptual do uso de TIC na entrega de valor por parte dos tomadores de decisão e Análise multidimensional sobre variáveis de desempenho a partir do uso de TI
Chege et al. (2020)	Relação entre perfil do empreendedor e inovação em TI
Dutot et al. (2014)	Capacidade da TI em relação a estratégia de internacionalização
Jamali et al. (2015)	Efetividade dos gastos de TI em PMEs e Impacto da TI em ambientes de alta tecnologia
Neirotti et al. (2013)	'as a service' e a remoção de barreiras de adoção de TI,

Referência	Lacuna de Pesquisa
	Apoio de grandes empresas na adoção de TI e Estudo entre a capacidade de TI e o ambiente competitivo
Lányi et al. (2021)	Mineração de conteúdo em relação à PMEs com presença online

Os principais impedimentos identificados neste estudo (Quadro 8) no âmbito estratégico de PMEs estão centrados na figura do gestor/empreendedor e de suas múltiplas reponsabilidades deste o direcionamento estratégico da empresa quanto em atividades de gerenciamento e até operação (Miyamoto (2015) e Rasel (2016)). A falta de capacitação em liderança técnica nesta posição (Dibrell et al. (2008), Neirotti et al. (2013), Peña-Vinces et al. (2012), Qureshil et al. (2009), Rasel (2016) e Yunis et al. (2017)) poderiam ser mitigadas por políticas públicas para transformação digital focados em desenvolvimento do empreendedor (Neirotti e Raguseo (2017) e Setiawan et al. (2015)).

Quadro 8: Barreiras de Implantação no domínio estratégico em governança de TI

Referência	Barreiras de Implantação
(Yunis et al., 2017)	Insuficiência de direcionamento teórico entre inovação, empreendedorismo e desempenho organizacional
Rasel (2016)	Planejamento e execução pelos mesmos profissionais
Miyamoto (2015)	Foco em atividades operacionais
Peña-Vinces et al. (2012); Qureshil et al. (2009) e Yunis et al. (2017)	baixa qualificação em TI do quadro de colaboradores, lideranças e empreendedor
Peña-Vinces et al. (2012)	Regulamentação no processo de comunicação entre fornecedor e cliente
Dibrell et al. (2008)	Baixo investimento em inovação (liderança tecnológica e vantagens no pioneirismo)
Neirotti e Raguseo (2017) e Setiawan et al. (2015)	Políticas públicas de incentivo a adoção de TI (crédito, por exemplo)
Neirotti et al. (2013) e Rasel (2016)	Baixa percepção de valor
Neirotti et al. (2013)	Restrição de recursos
Pratono (2017)	Turbulência tecnológica

Como apontado por (Miyamoto, 2015) e evidenciado pelos resultados deste estudo, a TI está locada majoritariamente em atividades operacionais e gerenciais, com maior campo de exploração no domínio estratégico. Na esfera gerencial (Quadro 9) focada na gestão de riscos e recursos (Khalil e Belitski, 2020), além de estudos longitudinais e cruzados já apontados para o domínio estratégico, questões referentes a taxa de conversão de conhecimento por meio da TI (Lopez-Nicolas e Soto-Acosta, 2010), desempenho sustentável (Memon et al., 2019) e eficácia de processos externos a organização (Migdadi et al., 2012) ainda se encontram pouco exploradas. Estes pontos podem ajudar a remover barreiras como a falta de uma base de conhecimento consolidada (Lopez-Nicolas e Soto-Acosta, 2010), utilização parcial de TI (Grande et al., 2011) e divergências com o alinhamento estratégico (Sandulli et al., 2012).

Quadro 9: Lacunas de pesquisa no domínio gerencial em governança de TI

Referência	**Lacuna de Pesquisa**
Lopez-Nicolas e Soto-Acosta (2010)	Conversão de informação em conhecimento em ambientes de alta e baixa tecnologia
Lopez-Nicolas e Soto-Acosta (2010)	Estudo longitudinal de conversão de informação em conhecimento por meio de TI
Bayo-Moriones et al. (2013) e Grande et al. (2011)	Estudo longitudinal sobre o efeito da TI na performance organizacional
Cantú et al. (2009)	Análise psicossociológica da gestão do conhecimento em TI
Francalanci e Morabito (2008)	Estudo de causalidade cíclica entre vantagem competitiva, capacidade de absorção e integração de TI
Memon et al. (2019)	Contribuição da TI em relação a sustentabilidade em diferentes economias
Migdadi et al. (2012)	Avaliação cruzada geográfica; Avaliação de desempenho organizacional pelo uso de TI na esfera empresa-ambiente externo.

O fortalecimento da base de conhecimento fortalece as capacidades dinâmicas de uma empresa (Chan et al., 2016) que por sua vez impactam

positivamente seu desempenho (Nabeel-Rehman e Nazri, 2019). Coordenar recursos em linha com os objetivos estratégicos permite ainda máxima transferência de valor pelo negócio (Bhaskaran, 2013), redução de riscos de ineficiência inerentes a sua adoção e demais barreiras na adoção de TI neste domínio (Quadro 10).

Quadro 10: Barreiras de Implantação no domínio gerencial em governança de TI

Referência	Barreiras de Implantação
Lopez-Nicolas e Soto-Acosta (2010)	Baixa ocorrência de uma estrutura de conhecimento centralizada
Shiau et al. (2009)	Custo e complexidade tecnológica; Baixa qualificação do quadro profissional alocado; saúde financeira
Rehman et al. (2020)	Atualização constante do quadro profissional
Bayo-Moriones et al. (2013)	necessidade de monitoramento contínuo como forma de garantir máximo aproveitamento
Grande et al. (2011)	Utilização parcial de TI
Sandulli et al. (2012)	Ineficiência frente ao alinhamento estratégico

Referente ao domínio operacional (Quadro 11) apesar de mais difundido em PMEs, as agendas de pesquisa disponíveis estão no contexto de portabilidade entre os sistemas (Ahuja et al. (2010); Corrocher e Fontana (2008)) reduzindo as falhas no processo de implantação) e aderência nos requisitos (Ahuja et al. (2010); Bhaskaran (2013)).

Quadro 11: Lacunas de pesquisa no domínio operacional em governança de TI

Referência	Lacuna de Pesquisa
Sidola et al. (2012)	Formulação de políticas de compartilhamento de informações
Bhaskaran (2013)	Impacto de TI em situações de crise e Estudos longitudinais
Corrocher e Fontana (2008)	Padronização e interoperabilidade de sistemas
JosephNg (2018).	Terceirização de tecnologias

Outro ponto com espaço para pesquisa suportada pelo desenvolvimento da interoperabilidade de sistemas é a terceirização de TI por PMEs (JosephNg, 2018), diminuindo os efeitos do custo médio elevado e retorno sobre o investimento apontados na Quadro 12 (Corrocher e Fontana (2008); Rehman et al. (2020)).

Quadro 12: Barreiras de Implantação no domínio operacional em governança de TI

Referência	**Barreiras de Implantação**
Ahuja et al. (2010)	Dificuldade em medir o desempenho, Superficialidade no uso de TI, Implementações malsucedidas e Qualidade de software
Bhaskaran (2013)	Requisitos não aderentes com a realidade PMEs
Corrocher e Fontana (2008); Rehman et al. (2020)	Incerteza tecnológica, Falta de vantagem relativa e Custos de investimento e substituição
JosephNg (2018)	Tempo médio de retorno elevado

Esta fase permite analisar o panorama de Tecnologias de Informação no contexto de Pequenas e Médias Empresas, identificando lacunas de pesquisa ainda pouco exploradas no âmbito acadêmico e quais as principais barreiras para adoção no ambiente organizacional. Os impedimentos e as lacunas foram segmentados segundo os domínios de governança deixando como sugestão de estudos futuros abordagem semelhante para os mecanismos de governança.

Como principal contribuição à pesquisa até o presente momento, esta fase apresenta um guia de sugestões de agendas de pesquisa e no contexto organizacional, o mapeamento oferece requisitos a serem considerados ainda na etapa de planejamento e fornece sólidos subsídios para fortalecer o alinhamento estratégico, garantindo a máxima entrega de valor às PMEs por meio de Tecnologias de Informação como no caso do FCD proposto neste estudo.

3 MATERIAIS E MÉTODOS

3.1 MATERIAIS

Todo o processamento de dados deste estudo foi realizado em um notebook da marca Dell, modelo Vostro 5490, 16GB de RAM e processador core i5. Assim como Yee *et al.* (2020), a linguagem Python foi amplamente utilizada, como evidenciado no pré-processamento de dados com aplicação dos pacotes computacionais Pandas e Numpy; para o processamento de dados e aplicação do algoritmo LDA os pacotes Gensim e NLTK; a persistência dos modelos treinados em Pickle. Para visualização, utilizou-se Matplotlib, PyLDAvis (Gensim) e Networkx para análise de rede. As tabelas de resultados de cada uma das etapas foram armazenadas em planilhas Microsoft Excel. O portfólio bibliográfico foi gerenciado pela ferramenta Mendeley com intuito principal de controlar as referências aplicadas a este relatório, cuja escrita foi realiza com o editor de texto Microsoft Word.

3.2 ESTRATÉGIA DE PESQUISA

Como já evidenciado na seção 2, Ciência de Dados no âmbito de Pequenas e Médias Empresas é um assunto pouco explorado na literatura. Se de um lado a ausência de conteúdo permite uma gama de possibilidades ao pesquisador, do outro, a coleta de insumos deve ser realizada de forma mais ampla e segmentada. Logo, a estratégia adotada neste estudo realizou o processo de seleção de portfólio em duas etapas —*Data Science Framework* e *SMEs & System Information*.

Ilustrado na Figura 17, evidencia-se o fluxo de estruturação de portfólio como um *framework* genérico de Ciência de Dados adaptado de George et al. (2016) e Nicolas et al. (2021). Para o primeiro caso, uma análise completa sobre os resultados do termo de busca (*query*), enquanto no segundo caso utiliza-se de técnicas de mineração de texto como forma de varredura mais abrangente em detrimento da volumetria de resultados.

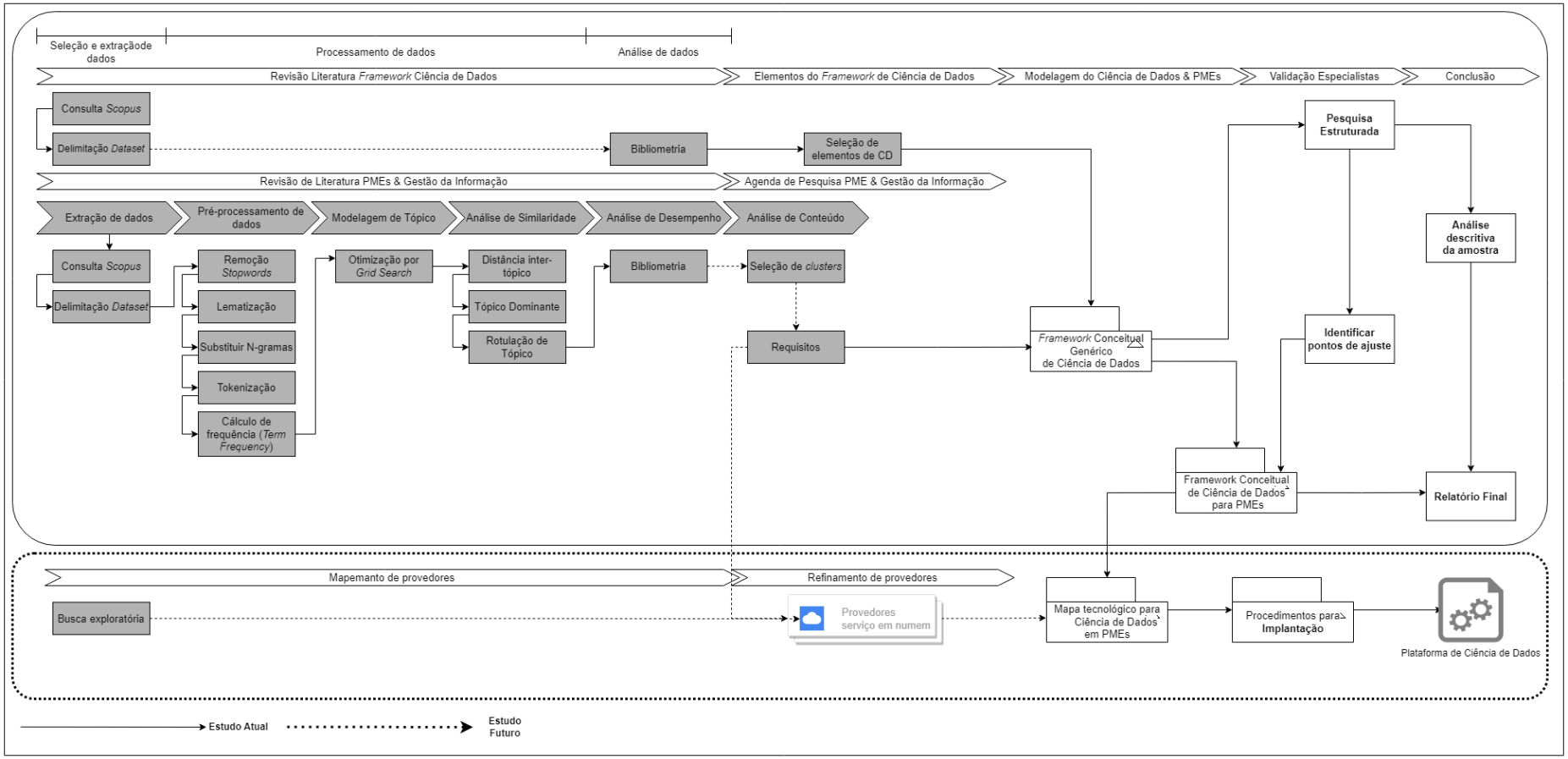
A seção 2.1 apresenta como resultado 49 artigos, os quais são

utilizados e discutidos na seção 4.1 para identificação dos elementos que compõem um FCD baseado no fluxo de informação organizacional, o que permite a elaboração do modelo teórico do FCD. No entanto, como proposta de contribuição, este estudo objetiva uma ferramenta que se enquadra na conjuntura de PMEs segundo a contextualização e justificativa apresentada na seção 1.

Como resultado da seção 2.2, 65 documentos foram selecionados para dar subsídio ao enquadramento da FCD no contexto de PMEs. A análise dos materiais terá como ponto de partida identificar quais são os *stakeholders*, requisitos para implantação, principais barreiras enfrentadas no processo e critérios de sucesso. Tais elementos reunidos permitem uma visão aprofundada das necessidades das PMEs e com isso selecionar somente componentes que em linha com seus atributos, evitando uma estrutura de custo financeiro e operacional elevado garantindo um melhor aproveitamento (eficiência) e maior entrega de valor (eficácia) pela FCD.

A contribuição e relevância acadêmica é evidenciada pelos resultados da abordagem supracitada. No entanto, este estudo busca avançar ainda no modelo prático e fornecer ao empreendedor um guia dos principais serviços em nuvem cuja necessidade de cada componente do FCD teórico pode ser atendida. Para tal, um levantamento exploratório foi realizado paralelamente às revisões de literatura com objetivo de identificar tais tecnologias e principais provedores dos serviços (tendo em vista também as restrições e contexto das PMEs) e constituir o modelo prático do FCD.

Figura 17: Metodologia geral de pesquisa



Pratono (2017) aponta a turbulência tecnológica como barreira na adoção de tecnologias da Informação por Pequenas e Médias Empresas, na qual diversas ferramentas surgem e desaparecem num breve espaço de tempo. Associado a orçamentos reduzidos como apontado por Neirotti et al. (2013) e Shiau et al. (2009), essa classe de empresas não dispõe de margem para implantações mal sucedidas e resiliência financeira para acomodar uma falha de implantação da mesma forma que o fariam empresas de grande porte. A argumentação mostra a relevância de um processo colaborativo com profissionais atuantes no segmento de inteligência de dados para que o FCD seja submetido a uma análise de consistência e comprovar ou refutar sua aplicação no ambiente de negócio; esta informação foi obtida por meio de pesquisa de campo em entrevistas com profissionais que atuem no segmento de TI em diferentes níveis organizacionais.

Cauchick Miguel et al. (2018) apresentam as seguintes etapas para um planejamento de pesquisa: selecionar o tipo de levantamento e o tipo de plano a ser utilizado. Na primeira fase decide-se se será realizado um levantamento populacional ou amostral; este último adotado neste estudo em virtude do custo e viabilidade. Uma vez selecionada a opção de levantamento amostral, deve-se escolher se o plano será probabilístico ou não probabilístico. Neste estudo, por conta de a necessidade de uma experiência prévia por parte do entrevistado ser condição para sua inclusão optou-se por uma amostragem não probabilística por julgamento.

A amostragem inicial contou com cerca de 100 profissionais distribuídos em relação ao tempo de atuação no setor de tecnologia. Ao aplicar a condição de que parte da experiência estivesse relacionada direta ou indiretamente às PMEs e em posição de liderança (gestor, coordenador, líder técnico e afins), chegou-se a 5 entrevistados com potencial de contribuição prática ao modelo. No entanto, entende-se que podem existir fatores de ajuste que contribuem para melhorar a eficiência deste estudo ampliando a amostra para 22 profissionais.

A entrevista teve um tempo médio esperado de 50 minutos e seguiu o protocolo proposto no Apêndice A. Dos pontos abordados, inicialmente foi apresentado ao respondente o contexto no qual está inserido o FCD e em seguida avaliou-se aspectos que qualifiquem o respondente (sem identificá-lo) com objetivo de analisar a percepção de valor em profissionais com experiência nessa

modalidade organizacional — Pequenas e Médias Empresas — e vivência na área de TI. Na terceira etapa, o entrevistado foi arguido em relação à sua percepção de valor em relação à FCD para empresas genéricas, lacunas não atendidas pela proposta e principais barreiras na sua utilização como guia de implantação para uma estrutura de Ciência de Dados em uma empresa genérica. A quarta etapa da pesquisa estruturada foi coletar quais fatores devem ser adicionados de forma a garantir adequação ao contexto de PMEs e quais as principais barreiras podem ser enfrentadas por esta classe de empresas com a utilização do FCD. Finalmente, o entrevistado teve um tempo livre para discorrer abertamente sobre temas não abordados *a priori* e que ele julgasse importante adicionar.

4 RESULTADOS

A respeito das referências utilizadas neste estudo, nota-se ampla aplicação do termo *framework* sem que haja asserção categórica sobre sua definição como em Lee; Tsai (2019) e Olmos et al. (2020) e Ramasamy et al. (2021) e Vijaykanth Reddy; Sashi Rekha (2021), por exemplo. No entanto, Shehabuddeen et al. (2000) realiza uma larga discussão sobre o assunto contribuindo com definição, aplicações, propósito e forma de apresentação de um *framework*.

Segundo o último autor supracitado, um *framework* "suporta o entendimento e comunicação da estrutura e relacionamento dentro de um sistema para um propósito definido". Nota-se um forte apelo nesta definição quando a finalidade do termo e não uma definição propriamente dita e o autor traz a luz dois propósitos básicos que identificam os tipos de *framework*:

i. *Know-How*: explanar como um objetivo pode ser atingido;
ii. *Know-What*: descrever uma situação em particular;
iii. Híbrido: parcialmente *Know-How* e parcialmente *Know-What.*

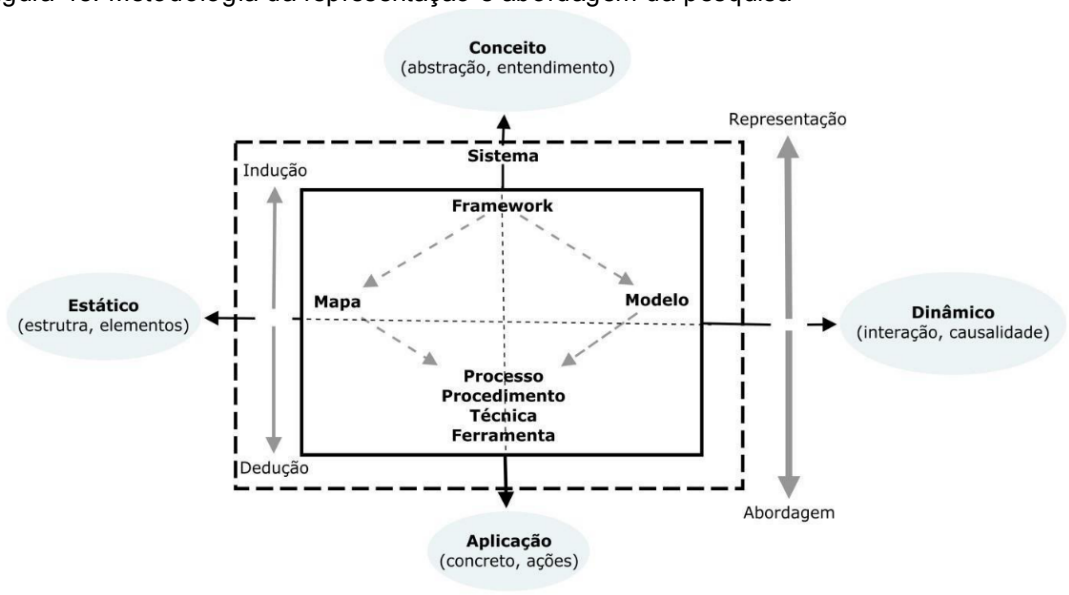
Como já evidenciado neste estudo, o termo *Framework* pode ser empregado em diversas situações, dentro das quais Shehabuddeen et al. (2000) aponta que uma aplicação bem sucedida deve traduzir situações complexas em um formato simples e passível de análise por parte do leitor. Para exemplos de casos de uso, o autor relata:

i. Comunicar ideias e resultados tanto em frentes puramente acadêmicas ou acadêmico-empresarial;
ii. Comparar diferentes situações e abordagens;
iii. Definição das condições de contorno de determinada situação;
iv. Descrição de um contexto ou discutir a validade de resultados;
v. Suportar o desenvolvimento de procedimentos, técnicas, métodos e ferramentas.

Com base no entendimento de PERRONI (2017) ilustrado na Figura 18 em relação ao exposto por Shehabuddeen et al. (2000), nesta pesquisa justifica-se a utilização do termo *framework* por tratar-se de uma representação conceitual de um sistema e que pode derivar para representações também conceituais mais estáticas

como mapas ou dinâmicas como modelos; este desdobramento será discutido na seção de trabalhos futuros. Procedimentos, técnicas, métodos e ferramentas encontram-se na camada mais prática e se apresentam como formas de implementação das representações conceituais.

Figura 18: Metodologia da representação e abordagem da pesquisa



Fonte: PERRONI (2017) apud Shehabuddeen et al. (2000)

Pinto (2016) apresenta cinco atividades chave as quais destinam-se um *framework* conceitual. No primeiro caso, entendimento da situação-problema. Segundo, define-se qual modelagem a ser utilizada e quais os objetivos gerais a serem atingidos. Em terceiro lugar analisar os resultados obtidos no modelo. A quarta atividade listada é identificar quais as entradas do modelo e finalmente qual conteúdo do modelo (maior detalhamento). As asserções da autora estão em consonância com os Elementos de Ciência de Dados apresentados na Tabela 6 corroborando a validade do estudo.

Frente ao exposto acima, este estudo enquadra-se no caso de uso “Suportar o desenvolvimento de procedimentos, técnicas, métodos e ferramentas”, pois uma vez validado, o desdobramento a ser realizado por estudos futuros permite a obtenção de uma lista estruturada de procedimentos para implantação de uma plataforma de Ciência de Dados, bem como suportar o traçado de um mapa para tecnologias elegíveis às necessidades de projeto. No tocante ao tipo, o resultado

acerca-se do grupo *Know-How* por buscar identificar quais processos os dados devem estar sujeito ao longo do fluxo da informação — gerar informação como objetivo geral.

4.1 *FRAMEWORK* GENÉRICO DE CIÊNCIA DE DADOS: GENÉRICO

Prakash et al. (2020) reforçam a utilização de *Framework* de Ciência de Dados como meio de obter predições sobre constructos de dados históricos estruturados e não estruturados. No entanto, o FCD pode ser aplicado como ferramenta de gestão de processos de dados como adotado por Gonzales et al. (2019), sem que haja necessariamente utilização de modelos em alta complexidade. Nesta abordagem a principal vantagem está em definir quais etapas de projeto são necessárias e quais resultados devem ser obtidos em cada uma delas. Seguindo esta a mesma ótica do autor supracitado e admitindo o enquadramento realizado no início do capítulo 4 referentes aos paradigmas de pesquisa, a Figura 19 ilustra o *Framework* Genérico de Ciência de Dados proposto.

De forma macro, utilizou-se a distribuição de atividades em grupos funcionais assim como realizado por Demchenko et al. (2017) — *Data Science Engineering* e *Data Science Analytics*. No entanto, tais grupos devem funcionar de forma sinérgica nas quais cada ação ou decisão tomada em uma camada deve gerar uma necessidade de ajuste ou balanceamento na outra camada. Em outras palavras, esta estrutura sugere um caráter de simbiose na atuação sem que haja subordinação entre as camadas.

Como ponto de partida no fluxo de dados tem-se a **Geração de Eventos**. Esta etapa é definida por uma série de eventos de natureza determinística ou estocástica e que por conseguinte dispara uma ou mais ordens de persistência de dados. O ambiente organizacional não controla os efeitos provocados por tais eventos externos a companhia, mas tem como demanda recorrente desenvolver uma forma de atuação resiliente e absortiva que consiga utilizar eventos passados para conjecturar cenários e adaptar-se as mudanças futuras. Os dados desta fase são armazenados em servidores de arquivos ou banco de dados dedicados a aplicação (ERP, sensoriamento, redes sociais, internet das coisas etc.) e o pré-

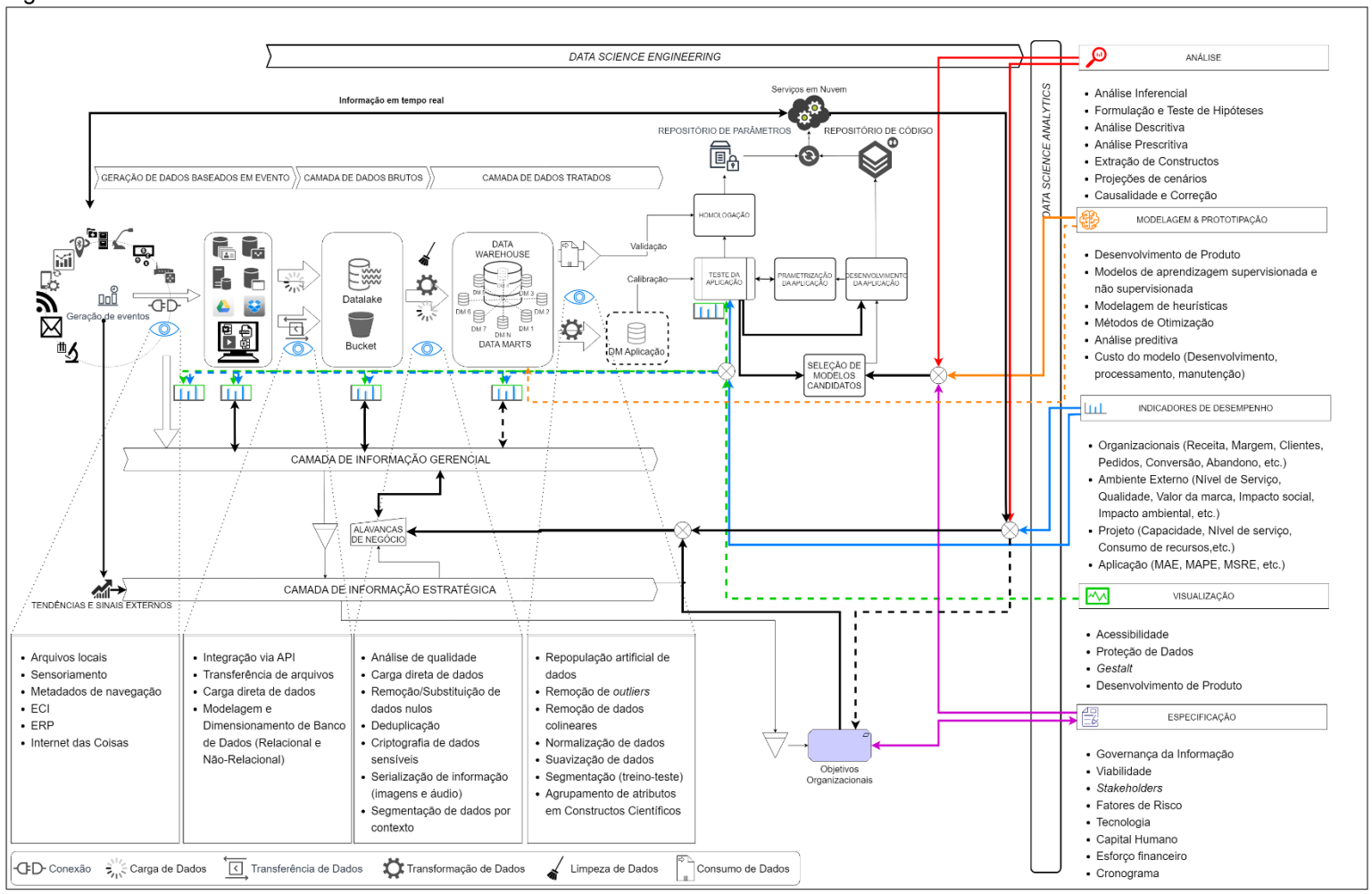
processamento é praticamente inexistente uma vez que não há necessidade de consumo de dados transformados e o foco está na garantia de integridade do dado.

No que tange ao armazenamento de dados de competência da organização, este estudo as divide em duas camadas — **Dados Brutos e Dados Tratados**. No primeiro caso os dados provenientes das fontes de eventos, por meio de integração via API ou carga direta de dados, são organizados em formato padrão que garantem a continuidade dos dados ao longo do tempo (versionamento) e permitem a leitura e manipulação em grandes volumes utilizando linguagens de alto nível como Python, R ou SQL, por exemplo. A esta estrutura dá-se o nome de *Data Lakes* e como apontado por Brous et al. (2020) e já mencionado neste trabalho. Sua utilização atende aos requisitos de velocidade e integridade dos dados, mas adiciona uma camada de complexidade em relação a segurança da informação e governança de dados. O presente trabalho prevê ainda a utilização de *Buckets*, ou seja, serviços em nuvem para armazenamento de arquivos de caráter imutável como imagens e áudios ou arquivos de texto cujo custo de desempenho em bancos de dados relacionais seria elevado.

A segunda camada de armazenamento tem por finalidade disponibilizar dados por contexto (financeiro, cadastro, produção, logística, comercial etc.) em formato simplificado para consumo e geração de informação para tomada de decisão — *Data Warehouse*. DW como são comumente chamados podem adotar diferentes arquiteturas como Relacional (OLAP ou *Online Analytical Processing*), Dados Colunares, *In-Database Analytics*, *In-memory Databases* e Processamento Massivo Paralelo (MPP) (Sherman, 2014). O *Framework* proposto adotou a arquitetura OLAP Relacional com uma abordagem dimensional por contexto (*Data Marts*) cuja integração resulta no *Data Warehouse* Organizacional.

A partir deste ponto surge a necessidade de limpeza e transformação de dados uma vez que o objetivo é fornecer um ambiente de integrado e de fácil consumo onde as inconsistências sistêmicas já foram corrigidas segundo as premissas de negócio. Das atividades chave realizadas aqui lista-se a segmentação de dados por contexto, proteção da informação segundo os preceitos legais, remoção ou substituição de dados nulos, remoção de duplicidades e serialização para imagens e áudios quando não for opção o armazenamento direto em *Buckets*.

Figura 19: *Framework* Genérico de Ciência de Dados



Fonte: Elaborado pelo autor (2023)

Por questão de desempenho de modelos de Ciências de Dados é preciso realizar transformações que vão garantir menor tempo de processamento e maior precisão nos resultados das predições (Chen et al., 2021;Lee et al.,2022). A depender da arquitetura adotada, é possível a criação de um *Data Mart* nativo para aplicação em detrimento do consumo direto dos demais *Data Marts* contextuais (financeiro, produção, logística etc.). Dentre as transformações realizadas destaca-se a repopulação artificial de dados quando o volume de uma determinada amostra é substancialmente menor que as demais e pode causar viés durante o processamento, remoção de valores extremos ou *outliers* e de dados colineares cuja permanência acarreta consumo de recursos sem ganhos de eficácia do modelo, normalização, suavização e agrupamento de atributos.

As camadas supracitadas se apresentam como insumos que serão utilizados na etapa de desenvolvimento de produtos de Ciência de Dados e cujo controle e monitoramento está sujeito a uma **Camada de Informação Gerencial** que traduz os interesses da **Camada de Informação Estratégica** em atividades operacionais e especializadas como Marketing, Comercial, Atendimento ao cliente etc. para **Geração de Eventos** e Suporte, Rede, Tecnologia, Desenvolvimento etc. para **Camada de *Data Science Engineering***.

A **Camada de Informação Estratégica** atua como origem de decisão e a partir do qual toda a estrutura de informação deve ser dimensionada. Deve escutar **Tendências e Sinais Externos** e ser capaz de convertê-los em **Objetivos Organizacionais** que dará início ao ciclo de vida dos produtos de Ciência de Dados. Importante ressaltar que esta camada pode tornar-se um ponto de maior dificuldade para Pequenas e Médias Empresas em detrimento do acúmulo de funções na figura do gestor-executor (Rehman et al., 2020;Sandulli et al.,2012;Shiau et al.,2009). Neste caso os objetivos de longo prazo são prejudicados pelos objetivos de curto prazo, o que gera um perfil organizacional imediatista e altamente reativo frente as flutuações de mercado.

De posse dos objetivos a serem atingidos pela organização, inicia-se o ciclo de vida dos produtos de Ciência de Dados. A primeira fase prevê a utilização do primeiro recurso da **Camada de *Data Science Analytics*** — **Especificação**. Este recurso fará o desdobramento dos objetivos em requisitos de projeto (viabilidade,

capital humano, esforço financeiro e cronograma), seleção de tecnologias, análise de risco e impacto sobre *stakeholders* internos e externos (inclusos aspectos legais), bem como garantir as melhores práticas de governança da informação.

Em seguida, adiciona-se ao subproduto da **Especificação** os esforços de **Modelagem e Prototipação** com intuito de selecionar o modelo mais adequado às questões de negócio. Além de conteúdos técnicos relativos ao desenvolvimento de soluções de Ciências de Dados como aprendizagem supervisionada e não-supervisionada, desenvolvimento de heurísticas e técnicas de otimização, é requirido deste recurso forte formação em desenvolvimento de produto para que sejam estabelecidas condições claras para lançamento, manutenção, melhorias e descontinuação da solução. O recurso deve ser capaz de prover um conjunto de métricas quantitativas e/ou qualitativas relativo ao custo de cada modelo permitindo uma seleção racional e de maior adesão às especificações.

Vale destacar que ainda no processo de seleção de modelos é possível uma associação com o recurso de **Análise** e **Indicadores de Desempenho** para definição de casos de uso que norteiam os processos de Parametrização e Testes da Aplicação e garantem condições mínimas para homologação. Durante todo o ciclo de desenvolvimento da aplicação é fundamental a sincronia com repositórios de código que permitem uma atuação paralela controlada e concerne maior segurança no controle de versões. Faz-se necessário também um repositório de parâmetros para capturar saídas do processo de homologação; mesmo para modelos não-supervisionados o ambiente pode ser utilizado para condições de contorno automatizadas.

Concluída a fase de homologação da solução, é papel da camada de ***Data Science Engineering*** disponibilizar a solução em ambiente produtivo em nuvem, garantir estabilidade e nível de serviço da aplicação. É nesta etapa que se inicia a geração de informação para tomada de decisão conforme eventos externos chegam ao ambiente organizacional. Munidos das saídas da solução de Ciências de Dados, os recursos de **Análise** e **Indicadores de Desempenho** em linha com os objetivos da companhia fornecem à camada de estratégica os subsídios necessários para mover as alavancas de negócio com maior propriedade e segurança. Tais alavancas provocam o balanceamento da estrutura e o processo reinicia reiteradas

vezes até que o ciclo de vida da aplicação termine e uma nova solução seja necessária.

Foi apresentado acima o fluxo da informação em sentido único. Mas como proposto por Gaedke Nomura et al. (2021) existem aprendizados durante as etapas intermediárias que podem ser incorporados e por conseguinte melhorar a eficácia do modelo (*feedback*). O *framework* proposto neste estudo faz uso deste artifício na fase de especificação, desenvolvimento e após o lançamento da aplicação. Na primeira atividade é fundamental o processo de *feedback* para garantir máxima adesão com os objetivos organizacionais e desencadear ajustes em detrimento das avaliações de viabilidade. No segundo processo as etapas de teste podem adicionar informações para adequação e substituição do modelo em casos mais críticos. Finalmente, a retroalimentação das saídas do modelo já em ambiente produtivo pode provocar mudanças diretamente nos objetivos da companhia e, portanto, em toda a estrutura proposta.

A **Visualização** é se apresenta no *framework* proposto como um recurso de utilização indireta da camada de ***Data Science Analytics*** e que não está diretamente ligada com a disponibilização da solução, podendo haver consumo do mesmo pela camada de ***Data Science Engineering*** para controle e monitoramento dos processos de coleta e armazenamento de dados. No entanto, assim como apontado por Joshi et al. (2018) e Saraee; Silva (2018), pode-se identificar padrões que melhoram a precisão do modelo na fase calibração da aplicação.

Faz parte da rotina deste recurso além do estudo de formas (*gestalt*), a percepção humana da visualização e questões de acessibilidade. Apesar de garantias legais de proteção da informação já estarem contempladas em fases anteriores, esta última camada deve estar atenta e garantir que uma engenharia reversa não permita precisar o local de origem de informações sensíveis (um gráfico com 100% sobre determinada opinião em uma pesquisa anônima com somente um respondente torna público a opinião do respondente, por exemplo). Da mesma forma que o recurso de **Modelagem e Prototipação**, uma formação com ênfase em desenvolvimento de produto garante maior diferencial e eficiência neste tipo de solução.

4.2 PROPOSIÇÃO DE AJUSTES DO FCD POR ESPECIALISTAS

Da mesma forma como apresentado no protocolo de pesquisa (Seção 3.2), esta etapa dos resultados foi dividida em quatro blocos — qualificação de entrevistado, percepção de valor e ajustes do FCD genérico, ajustes do FCD ao contexto e PMEs e possíveis barreiras de adoção, informações complementares.

4.2.1 Qualificação dos entrevistados

O objetivo central desta etapa é situar o leitor quanto ao perfil do entrevistado e suas experiências nos diversos setores e segmentos, permitindo desta forma traçar de paralelos que validem a proposta aqui apresentada em diferentes realidades guardadas as devidas particularidades. Os entrevistados foram agrupados inicialmente segundo o **"nível organizacional de atuação"**. Esta proposta busca enquadrá-lo referente ao seu nível de responsabilidade no direcionamento da organização que atua no momento deste estudo.

O nível **Acadêmico** foi utilizado quando os entrevistados estão vinculados a instituições de ensino, **Estratégico** quando responsável por definir os objetivos da organização, **Tático** quando responsável pelo desdobramento dos objetivos em atividades e **Operacional** quando executor das atividades. Um fato comum em profissionais de tecnologia é o enquadramento em mais de um nível em detrimento de um modelo de gestão mais horizontal (estratégico/tático e tático/operacional), nestes casos foi considerado o maior nível de responsabilidade.

Em seguida, coletou-se a **Idade** e **Tempo de Experiência com TI** dos entrevistados. Reunidos na Tabela 6, os dados apontam maior dispersão em ambos os atributos na camada estratégica em virtude principalmente pelo aumento do empreendedorismo em profissionais mais jovens em início de carreira. Outro destaque está na alta similaridade entre valores da camada tática e operacional. Este fato dissocia a relação intrínseca entre tempo de experiência e responsabilidades gerenciais, havendo profissionais com formação direcionada (4 anos de experiência) para atuação tática e profissionais com carreira sólida na camada operacional (19 anos de experiência), o que nos mostra que esta última

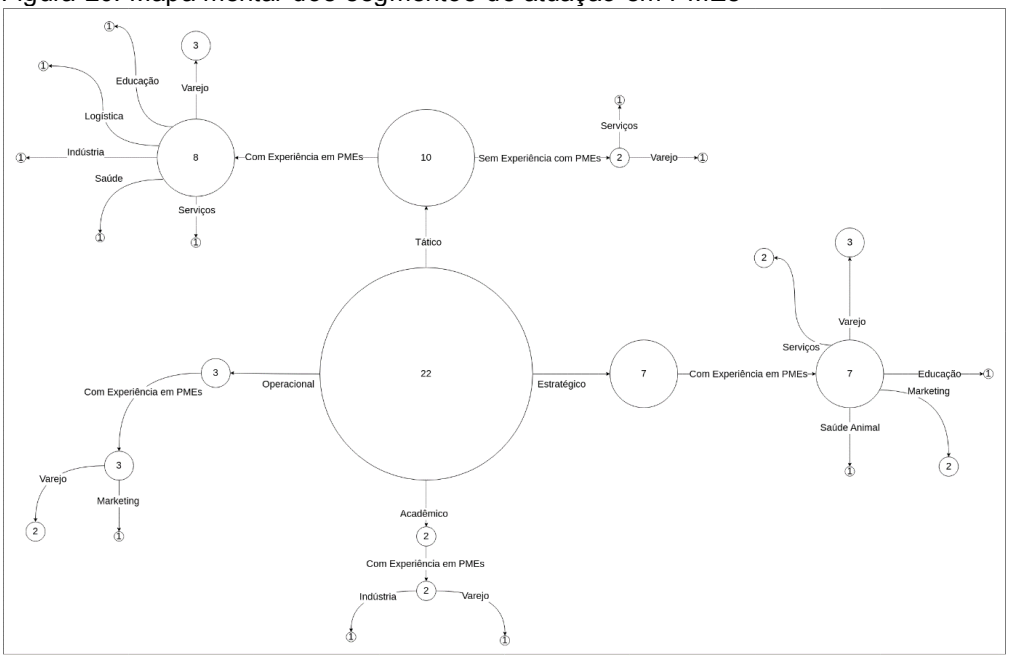
necessariamente não se apresenta como fase inicial para chegar-se a primeira.

Tabela 3: Dispersão de idade e tempo de experiência por nível organizacional de atuação

Nível	Número respondentes	Mínimo (anos)		Médio (anos)		Máximo (anos)	
		Idade	Experiência	Idade	Experiência	Idade	Experiência
Acadêmico	2	42	13	42,50	18,00	43	23
Estratégico	7	31	6	39,57	20,00	56	40
Tático	10	26	4	33,20	11,60	39	22
Operacional	3	28	5	32,67	11,33	38	19
Total geral	22	26	4	36,00	14,82	56	40

Na Figura 20, buscou-se identificar dentro da amostra de entrevistados com e sem experiência com PMEs quais os setores de maior expressão. Nota-se um desequilíbrio quanto a seleção de entrevistados em relação ao nível organizacional, o que sugere para estudos futuros um aprofundamento sobre as camadas **Operacional** e **Acadêmico**, aqui com participação de 14% e 9% respectivamente. Referente ao atributo **Experiência com PMEs** nenhuma conclusão pôde ser apontada pois apenas dois entrevistados no nível tático reportaram não ter contato com PMEs. Nestes casos as contribuições coletadas subsidiaram os ajustes no FCD genérico, não inviabilizando suas respostas.

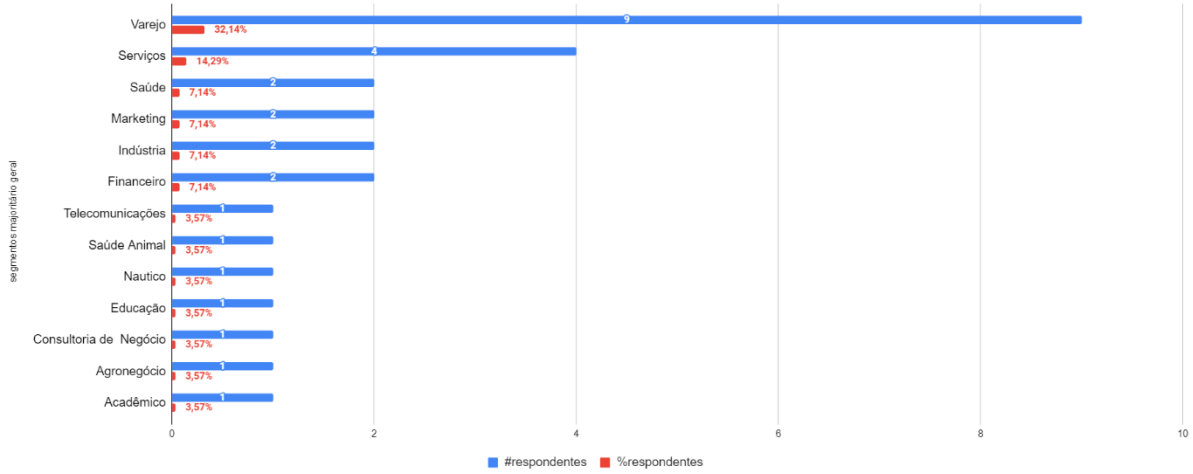
Figura 20: Mapa mental dos segmentos de atuação em PMEs



Fonte: Elaborado pelo autor (2023)

Em relação a área geral de atuação dos entrevistados ilustrados na Figura 21, nota-se uma preponderância para Varejo e Serviços da mesma forma que evidenciado na Figura 20 para apenas atuações em Pequenas e Médias Empresas. O atributo **Segmento de Atuação** foi arguido em relação a finalidade a qual destina-se seu trabalho, sendo Tecnologia da Informação uma área meio, sua aplicação tem por finalidade suporte a negócio na área de varejo, saúde, educação etc. Nesta pesquisa não foi utilizado Classificação Nacional de Atividades Econômicas (CNAE) para identificar o ramo de atividade da organização do entrevistado, mas somente uma abordagem qualitativa.

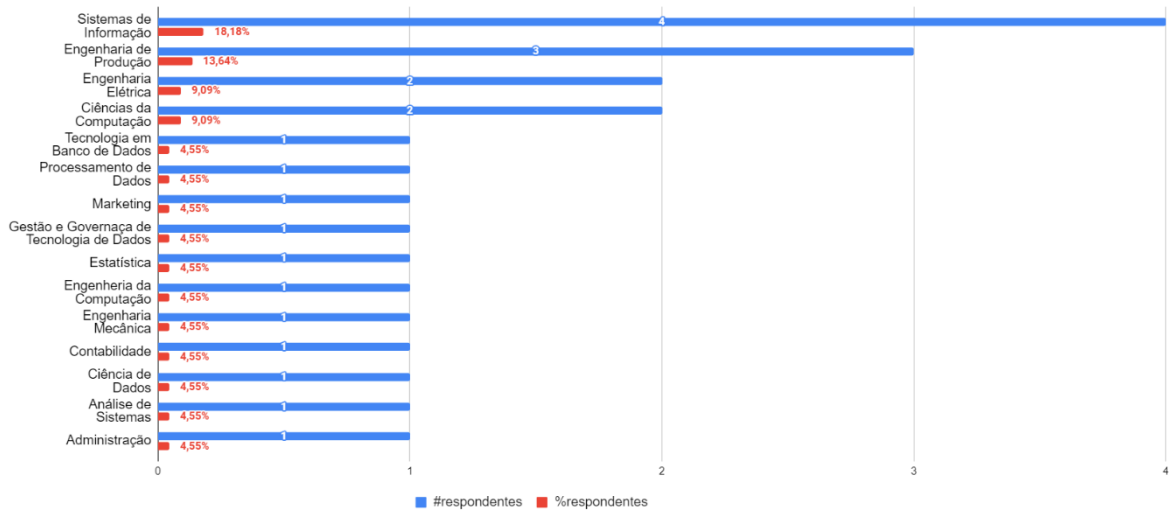
Figura 21:Distribuição de entrevistados por segmento de atuação geral



Fonte: Elaborado pelo autor (2023)

Referente a **Formação Acadêmica**, a Figura 22 nos mostra uma preponderância de Engenharias (43,92%) e Sistemas de Informação (18,18%) na amostra pesquisada — O primeiro caso por apresentar um caráter curricular voltado essencialmente para desenvolvimento de projetos, pode explicar parcialmente a alocação de profissionais com menor tempo de experiência com atuação na camada tática e estratégica. Outro fato relevante ainda sobre a formação dos respondentes é a maior concentração em áreas técnicas (Ciências Exatas), com apenas dois entrevistados (9%) provenientes de Ciências Humanas (Administração) e Sociais Aplicadas (Contabilidade).

Figura 22: Distribuição de entrevistados por Formação Acadêmica



Fonte: Elaborado pelo autor (2023)

Por trata-se de uma amostragem por julgamento não foi possível realizar qualquer inferência sobre os atributos avaliados na pesquisa, mas somente descrever o cenário de atuação e expertise dos perfis entrevistados a título de corroborar sua participação neste estudo.

4.2.2 Percepção de Valor e Ajustes em Relação ao FCD Genérico

Após a realização da primeira etapa do protocolo de pesquisa — apresentação do FCD — o entrevistado recebeu a seguinte questão: **O FCD proposto pode ser utilizado para direcionar a adoção de processos de Ciência de Dados por uma empresa genérica?** Todos os participantes responderam afirmativamente corroborando a validade da proposta nos diferentes âmbitos organizacionais.

Uma vez que a proposta de valor foi assegurada, os entrevistados foram questionados sobre **quais pontos de atrito podem ser mitigados com a adoção do FCD para orientar processos de Ciência de Dados**. A ideia central é mapear quais as regiões de maior demanda no esforço de projeto cuja entrega de valor terá impacto mais significativo e que poderia receber, portanto, maior priorização. A maior percepção de valor mencionado foi **Melhoria na Governança da Informação**, com 45% de ocorrência. Com alta relevância surgiu ainda menções sobre **Melhoria na eficiência dos processos** e **Aumento no desempenho da**

equipe, indicando que os entrevistados acreditam no impacto positivo da proposta sobre essas áreas. Outras menções significativas incluem **Clareza no Escopo de Projeto/Produto e Tomada de Decisões** e **Melhoria na Governança de TI**.

De forma geral, os entrevistados avaliaram como positivo a clareza, orientação e suporte fornecidos pelo *Framework* Genérico de Ciência de Dados nas diversas fases de um projeto de Ciência de Dados. Além dos fatores já mencionados como governança da informação, tomada de decisões e eficiência dos processos, o FCD pode contribuir para **melhoria na comunicação**, **colaboração e coordenação dentro dos projetos**, **incorporação de *feedbacks***, **visão holística**, **maior segurança da informação** e **aprimorar a proteção e confidencialidade dos dados**. A **redução de custos** e **escalabilidade** foram mencionadas, sugerindo que os respondentes veem também potenciais benefícios econômicos e vantagens de escalabilidade na utilização da proposta. A seguir, os grupos de respostas e descrição de impacto:

a) Melhoria na governança da informação: ajuda a aprimorar a gestão, controle e integridade da informação dentro de suas organizações;
b) Redução da dispersão de recursos: reduzir a dispersão de recursos, o que pode levar a uma melhor utilização e eficiência dos recursos;
c) Melhoria na eficiência dos processos: contribui para a otimização dos processos, resultando em uma melhor eficiência em seus projetos ou organizações.
d) Aumento no desempenho da equipe: melhora o desempenho da equipe por meio de uma melhor colaboração, comunicação ou alocação de recursos.
e) Maior clareza no escopo do projeto/produto: compreensão mais clara do escopo do projeto ou produto, ajudando-os a definir metas e objetivos de forma mais eficaz.
f) Melhoria na governança de TI: contribui para uma melhor governança de TI, que envolve alinhar estratégias de TI com objetivos empresariais, melhorar os processos e garante conformidade na tomada de decisão.
g) Maior clareza na tomada de decisões: auxilia na tomada de decisão, fornecendo insights, informações ou orientações mais claras.
h) Visão holística: possibilita uma compreensão abrangente do processo de

Ciência de Dados e ajuda a otimizar suas atividades relacionadas a dados.

i) Fornece um processo de feedback já especificado: o processo de feedback integrado auxilia na melhoria contínua e refinamento de projetos.
j) Proporciona um fluxo claro da jornada da informação: facilita um fluxo de informação suave e bem definido, melhorando a comunicação e coordenação em seus projetos.

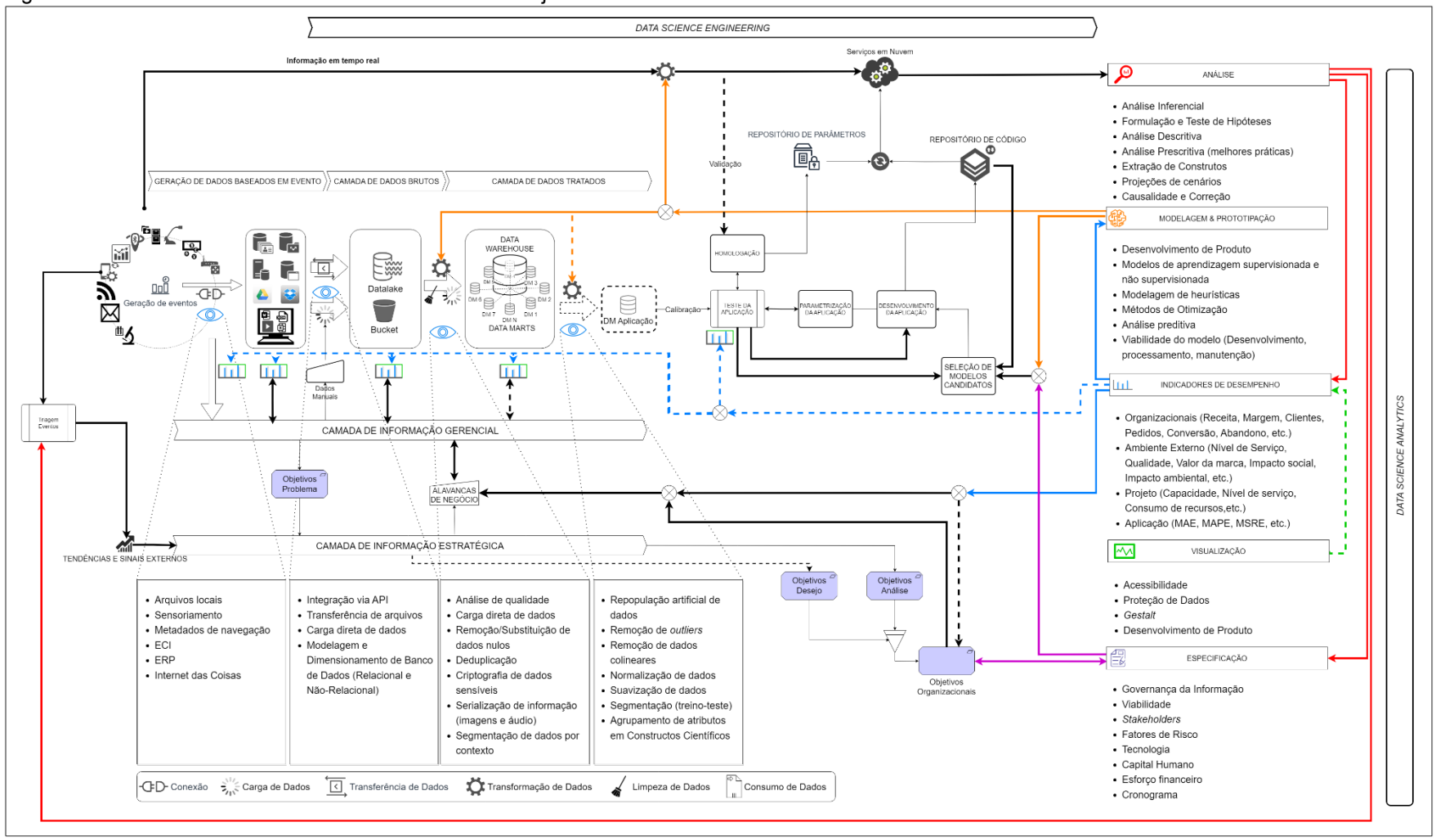
A segunda questão da etapa de percepção de valor foi avaliar **quais pontos/elementos o respondente entende que não foram contemplados pelo FCD** e que o entrevistado julga pertinente para robustez da proposta. No ponto de **Geração de Eventos** foi apontado a necessidade de um ponto de triagem para categorização, associação e identificação de padrões em eventos que dê suporte a camada estratégica para elaboração dos objetivos estratégicos. Outro ponto de destaque nesta questão foi a entrada de dados manuais referentes a novos aprendizados e condições conjunturais particulares a organização (metas, desabastecimento, campanhas, sazonalidade, vendas corporativas etc.).

Foi sugerido ainda que a camada de **Análise** seja o primeiro ponto de contato com a camada de **Especificação** fornecendo subsídios para o fechamento de requisitos primários da aplicação, antecipando pontos de atrito que só seriam percebidos no processo de **Seleção de Modelos**. O respondente apontou ainda a necessidade de integração de múltiplas aplicações, no formato atual a cadeia proposta prevê a disponibilização de apenas uma solução e de sua substituição quando da mudança dos objetivos estratégicos.

Um dos entrevistados discorreu a respeito dos objetivos de uma empresa de forma geral agrupando-os em três categorias em relação a origem: “problema”, quando dissidente de uma falha e tem caráter mais reativo; “análise”, quando a organização identifica por meio de análise um comportamento ou uma oportunidade (visão de longo prazo); “desejo”, quando proveniente de uma ideia cuja necessidade não está relacionada diretamente com a sobrevivência da organização, mas que pode conferir-lhe vantagens em alguma frente específica.

A proposta deste estudo prevê apenas a segunda modalidade (“análise”), no entanto, as demais também são realidade e precisam ser consideradas a fim de garantir o máximo de informações de entrada e maior

Figura 23: Framework Genérico de Ciência de Dados Ajustado



Fonte: Elaborado pelo autor (2023)

aderência da solução. A Figura 23 ilustra o resultado após inclusão dos ajustes sugeridos.

4.2.3 Ajustes do FCD Genérico ao Contexto de PMEs e Barreiras de Adoção

Esta fase do protocolo teve por objetivo utilizar a experiência dos entrevistados para identificar **quais ajustes devem ser realizados a título de garantir adequação do FCD ao contexto de PMEs**. O Quadro 13 reúne os pontos de ajuste considerados válidos.

Quadro 13: Pontos de Ajuste do FCD ao contexto de PMEs

Categoria	**Pontos de Ajuste**
Governança	- Concatenação das camadas de informação gerencial e estratégica
	- Simplificar processos para cada contexto de negócio (segmento/porte)
	- Modularização do framework (solução *end-to-end* para armazenamento, *end-to-end* para digitalização etc.)
	- Simplificação de apresentação do framework para uma PME
	- Ajuste do desenho para modelo caixa preta (entrada-saída)
	- Adicionar processo para definições de digitalização
	- Faseamento da aplicação em relação a maturidade da empresa
	- Selecionar pontos chave e conciliar com gestão ágil para simplificação no desenvolvimento do projeto e garantir entregas cíclicas
	- Incorporar bases de conhecimento ao longo da jornada da informação
	- Adicionar camada de aprendizagem entre o operador de CD e o problema/objetivo organizacional
Desenvolvimento	- Suprimir etapas de homologação e testes
	- Desenvolvimento em ambiente produtivo diretamente
Armazenamento	- Simplificação da entrada de dados
	- Centralização de armazenamento de dados
	- Utilização de *Lakehouse* integrando a camada de dados brutos e tratados
	- Direção de dados tratados do DW para derivar os DMs (para evitar falhas de comunicação e relacionamento de dados)
	- Armazenamento de dados brutos somente em *Bucket* ou somente em *Datalake*
	- Substituir *Data Warehouse* por *Feature Store*

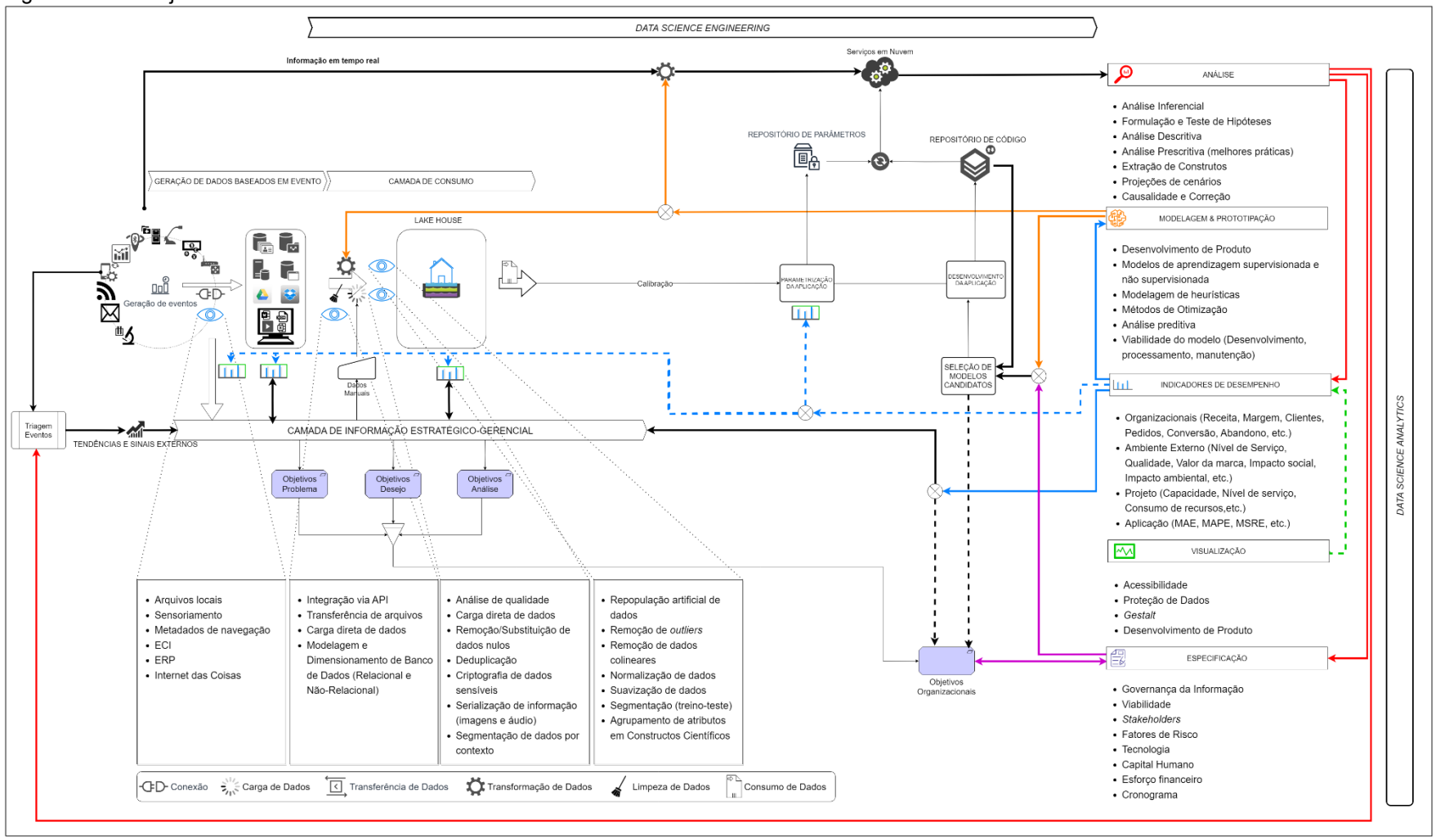
Determinados apontamentos já haviam sido contemplados na seção anterior — processo de triagem para classificação de eventos e suporte da camada de análise para identificação de padrões pelo Camada de Informação Estratégica. Quatro entrevistados não adicionaram pontos de ajuste, dois não haviam tido contato com PMEs e dois julgaram a proposta adequada a qualquer porte organizacional.

As respostas foram categorizadas em três grupos: Governança, Desenvolvimento e Armazenamento. No primeiro caso as contribuições buscam simplificar e modularizar o modelo de apresentação e melhorar o entendimento da proposta de valor e mitigar o efeito de diferenças de qualificação do quadro profissional. Outrossim está na figura do Gestor/Executor, não havendo necessidade da quebra em uma camada estratégica e uma gerencial.

A segunda categoria está associada ao processo de desenvolvimento da aplicação. Na realidade de grande parte das PMEs brasileiras, como já apontado neste estudo, é a verticalização de múltiplas funções por um quadro colaborativo reduzido. Dessa maneira, como forma de dar mais agilidade ao processo de desenvolvimento, sugeriu-se a supressão do ambiente de teste e homologação disponibilizando a solução direto em ambiente produtivo. Apesar das ressalvas quanto ao modelo de negócio, dois entrevistados afirmaram utilizar essa abordagem justificando que o custo de correção de falhas é inferior ao de revisar completamente as aplicações e manter três ambientes em uma empresa pequena e de médio porte.

Por fim, as modificações estruturais mais expressivas surgiram na categoria de Armazenamento de Dados. O ponto de vista defendido pelos entrevistados é que com a baixa disponibilidade de capital humano qualificado e recursos financeiros escassos é inviável a manutenção de serviços para dados brutos e para dados tratados separadamente. A centralização dos dados, mesmo que num primeiro momento, acarreta redução nos custos de infraestrutura e facilita o consumo da informação. A sugestão por parte dos especialistas é adotar uma solução de *Lakehouse* (*Data Lake* + *Data Warehouse*) que possibilita o armazenamento tanto dados estruturados como não-estruturados como entrada de aplicações de Ciência de Dados e para informações gerenciais. A Figura 24 ilustra o resultado após incorporação dos critérios apontados.

Figura 24: FCD ajustado ao contexto de PMEs



Fonte: Elaborado pelo autor (2023)

Um dos especialistas sugeriu a substituição completa da camada de tratamento de dados por uma *Feature Store* na qual são armazenados atributos e transformações necessárias para uma aplicação de Ciência de Dados. As *Features* podem ser reaproveitadas em outras soluções e consomem direto da camada de dados bruta. A viabilidade desta proposta é deixada a estudos futuros uma vez que aproposta do *Lakehouse* adiciona valores secundários aos propósitos da Ciência de Dados como uma camada de informações gerenciais, por exemplo, mostrando-se mais adequada a proposta deste estudo.

Na Seção 2.2.6 foi coletado por meio de revisão de literatura quais barreiras podem ser enfrentadas na adoção de Tecnologias de Informação. Da mesma maneira, com objetivo de identificar limitações no contexto de Ciência de Dados, os entrevistados foram questionados sobre **quais as principais barreiras que podem ser enfrentadas na utilização do FCD como insumo para um roteiro de implantação de uma estrutura de Ciência de Dados em PMEs**. Como resultado, a falta de capital humano qualificado e a baixa percepção de valor foram mencionadas repetidamente como as principais barreiras. Outros pontos de atenção incluem limitações financeiras, falta de organização da empresa, cultura organizacional e disponibilidade tecnológica insuficiente. Todas as respostas foram agrupadas e descritas a seguir:

a) Complexidade: complexidade das camadas de informação dificultam o uso efetivo do *framework*.
b) Velocidade de design: velocidade de design de um produto de Ciência de Dados é um desafio, tanto na criação ou personalização de soluções.
c) Capital humano: A falta de recursos humanos ou insuficiência de pessoal capacitado.
d) Gestor executor: falta de foco em atividades gerenciais e alta reatividade.
e) Baixa capacidade técnica: A falta de conhecimentos técnicos ou habilidades tecnológicas necessárias para utilização do *framework*.
f) Baixa disponibilidade tecnológica: A falta de recursos tecnológicos ou infraestrutura adequada.
g) Cultura organizacional: A cultura ou falta de organização existente na empresa pode ser uma barreira à adoção e utilização efetiva do *framework*.

h) Baixa percepção de valor: Entendimento de que a proposta pode adicionar valor ao negócio.
i) Limitações financeiras: Orçamentos curtos associados a baixa percepção de valor diminui iniciativas de inovação.

Em resumo, camadas de informação complexas, velocidade de design, limitações financeiras, falta de infraestrutura tecnológica e falta de disponibilidade de recursos humanos qualificados indicam desafios técnicos e tático/operacionais enfrentados por PMEs. Enquanto falta de apoio dos gestores, cultura organizacional, falta de organização da empresa e a falta de clareza sobre a adição de valor foram mencionadas como obstáculos estratégicos na adoção do FCD por Pequenas e Médias Empresas.

Importante retomar neste ponto as barreiras de implantação reunidas nos quadros 8, 10 e 12, bem como confrontar tais pontos retirados da literatura com as opiniões de especialistas no tema. Do ponto de vista estratégico todas as opiniões são corroboradas pela revisão de literatura exceto pelas **Políticas públicas de incentivo a adoção de TI** abordada por Neirotti e Raguseo (2017) e Setiawan et al. (2015), o que pode sinalizar uma falta de alinhamento entre entidades públicas e privadas a ponto de dessa última gerar expectativas sobre os resultados da primeira. Da camada gerencial e operacional, chama a atenção o fato dos especialistas não apontarem a estrutura de conhecimento com um fator relevante, como apontado na literatura, esta camada é fundamental no processo de adoção de tecnologias de informação com um todo e mensuração de resultados.

4.2.4 Informações Complementares

Durante o processo de pesquisa podem surgir elementos ou discussões não previstas durante a elaboração do protocolo e por tratar-se de uma entrevista estruturada corre-se o risco de não inclusão de tais contribuições. Desta forma, foi adicionado ao protocolo esta seção livre na qual o entrevistado é arguido se **existe algo a mais que queira adicionar cujas perguntas anteriores não tenham contemplado**.

De fato, tal situação pôde ser mapeada embora determinadas opções

foram descartadas *a priori* em detrimento do enquadramento de pesquisa. O primeiro comentário sugeriu apontamento no FCD elementos sobre interoperabilidade de sistemas e níveis de acesso durante o fluxo de dados. No entanto, conforme definido na introdução do capítulo, este estudo é uma representação mais abstrata do sistema e conforme o seu desdobramento em procedimentos ou ferramentas tais papeis serão definidos.

Em relação ao escopo, uma proposta que pode ser avaliada por estudos futuros é a segmentação do FCD por porte de empresa (Micro, Pequena e Média) por haver grandes discrepâncias na realidade de cada uma delas e/ou por categoria de problemas — Framework de Ciência de Dados para roteirização em Microempresas varejistas, por exemplo. Argumentou-se que conforme houver aumento de complexidade nas informações, os resultados podem não ser positivos para a proposta atual apesar dos ajustes. Outro comentário sugere ainda que a depender do contexto, o framework deve ser utilizado por completo para garantir máxima adesão normativa (legislação, auditoria etc.) — empresas financeiras, por exemplo — convertendo-se em um instrumento de validação do modelo de negócio.

Ainda sobre definições escopo, uma proposta de entrega de valor em fases baseadas no nível de maturidade da empresa prevendo inclusive processo de digitalização. Uma contribuição de alta relevância dado que este estudo não fez referência sobre o nível de informação das empresas inicialmente sugerindo capacidade minimamente instalada. No mesmo sentido, sugeriu-se a utilização de *Business Feature*, conceito que prevê a especificação da plataforma tecnológica com base em modelos de negócio ou necessidades de negócio pré-formatados.

Finalmente, como informado inicialmente no título desta proposta, a solução não especificou o cenário de utilização do *framework* como particular a uma única empresa ou um serviço a múltiplas empresas de pequeno e médio porte. No entanto, a percepção corrente entre os entrevistados foi de que o FCD poderia ser utilizado como modelo de negócio para consultoria junto a PMEs fornecendo um guia de como estruturar seu fluxo de dados e aumentar a eficiência na tomada de decisão.

5 CONSIDERAÇÕES FINAIS

5.1 CONCLUSÃO

O estudo tem como delimitação o contexto de Pequenas e Médias Empresas (PMEs) brasileiras e como evidenciado ao longo de seu desenvolvimento, mesmo com papel importante na economia representando uma parcela significativa dos empregos gerados, muitas delas enfrentam ainda desafios na adoção de novas tecnologias, devido às limitações de recursos financeiros e de capital humano, falta de conscientização sobre capacidades digitais, padronização de processos e segurança da informação. A falta de uma estrutura qualificada de informações e decisões orientadas a dados é uma questão comum para muitas PMEs, mesmo para empresas de comércio eletrônico que fazem uso intensivo de ferramentas de gestão da informação. Além disso, diferentes setores enfrentam desafios específicos, como interoperabilidade, acesso à informação, políticas de saúde, segurança da informação, validação em tempo real e natureza e complexidade dos dados.

A formação de profissionais na área de análise de dados é apontada como uma necessidade crucial para impulsionar o desempenho econômico, industrial e acadêmico. A Ciência de Dados exige uma abordagem multidisciplinar, e a modernização da economia requer uma reformulação nos modelos tradicionais de ensino. Frente a importância das PMEs na economia e as dificuldades que enfrentam na adoção de tecnologias, os esforços de investimento em capital humano e infraestrutura de informação pode impulsionar seu crescimento e inovação.

A discussão dos resultados sobre a utilização de um Framework Genérico de Ciência de Dados (FCD) em Pequenas e Médias Empresas (PMEs foi dividida em quatro blocos: qualificação dos entrevistados, percepção de valor e ajustes do FCD genérico, ajustes do FCD ao contexto de PMEs e possíveis barreiras de adoção. O primeiro bloco mostrou maior dispersão na relação idade/tempo de experiência na camada estratégica, devido ao aumento do empreendedorismo entre profissionais mais jovens. Além disso, houve uma alta similaridade entre os valores na camada tática e operacional, mostrando que a experiência não corresponde

necessariamente às responsabilidades gerenciais.

Quanto a percepção de valor e ajustes do FCD genérico os entrevistados afirmaram que o FCD poderia ser utilizado para direcionar a adoção de processos de Ciência de Dados em empresas genéricas. A melhoria na governança da informação foi mencionada como o ponto de maior valor, seguida pela melhoria na eficiência dos processos e aumento no desempenho da equipe. Os entrevistados também destacaram a clareza no escopo do projeto/produto, tomada de decisões, melhoria na governança de TI e outros benefícios proporcionados pelo FCD. Os ajustes do FCD ao contexto de PMEs foram sugeridos na camada de governança, desenvolvimento e armazenamento; destaques para a necessidade de triagem para classificação de eventos, a entrada de dados manuais e a integração de múltiplas aplicações, bem como a adoção de uma solução de *Lakehouse* para o armazenamento de dados central visando reduzir custos e facilitar o consumo da informação.

A falta de capital humano qualificado e a baixa percepção de valor foram apontadas como as principais barreiras. Outros obstáculos mencionados incluem limitações financeiras, falta de organização da empresa, cultura organizacional e disponibilidade tecnológica insuficiente. Em resumo, os resultados indicam a validade do uso do FCD genérico em PMEs, destacando os benefícios e ajustes necessários para sua adoção. As barreiras e informações complementares como a necessidade de elementos sobre interoperabilidade de sistemas, a segmentação do FCD por porte de empresa e a entrega de valor em fases baseadas no nível de maturidade da empresa fornecem insights importantes para futuros estudos e desdobramento do *framework* em ferramentas e/ou procedimentos de dados.

5.2 SUGESTÕES DE TRABALHOS FUTUROS

Um dos pontos deixados em aberto por este estudo reside no esclarecimento e validação do ganho de desempenho entre um método de revisão de literatura autônomos ou semiautônomos frente aos métodos de revisão tradicionais. As aplicações com este propósito estão em franca expansão no período

de elaboração deste relatório (Marjanovic; Dinter, 2018; OMara-Eves; Thomas, J.; McNaught, J.; Miwa, M., 2016); Watanabe et al., 2018), no entanto, **quais pontos ou documentos deixa-se de considerar em detrimento de uma avaliação completa do portfólio comparando-se métodos tradicionais e abordagens por mineração de texto?** Esta é apenas uma das questões neste âmbito que pode ser explorada e com potencial de contribuição acadêmica.

Outrossim que pode ser avaliado em conjunto na proposta supracitada é a própria **medida de eficácia na atribuição de rótulos em modelagem de tópicos**. Uma proposta seria uma pesquisa probabilística na qual o respondente é apresentado a um conjunto aleatório de tópicos e ao texto de referência (resumo da publicação, por exemplo) e deve selecionar qual tópico representa a ideia central do texto. Caso haja relativa dispersão entre os rótulos recebidos para um mesmo tópico sua qualidade pode estar comprometida.

Da mesma forma que as propostas de Perroni (2017) e Pinto (2016), este estudo se enquadra em um paradigma de pesquisa de alta abstração e cujo objetivo é direcionar a elaboração de procedimentos para implantação de uma proposta técnica. No contexto abordado, poder-se-ia utilizar o *Framework* de Ciência de Dados proposto para **elaboração de procedimentos para implantação de uma plataforma de Ciência de Dados tanto para PMEs como para empresas em geral**.

Este estudo reuniu diversas barreiras na utilização e implantação de Tecnologia da Informação em Pequenas e Médias Empresas (Tabelas 6,8,10 e 12). No contexto de dados as barreiras vão de limitações financeiras à disponibilidade dos dados em si. Como forma de mitigar tais efeitos, uma **proposta de plataforma colaborativa de ciência de dados para PMEs** colocaria no mapa empresas que de outra forma não teriam acesso a tecnologias de vanguarda. Esta sugestão pode ser desenvolvida posteriormente a **elaboração de procedimentos para implantação de uma plataforma de Ciência de Dados tanto para PMEs**, entendendo-se que essa última é um requisito para a primeira.

REFERÊNCIAS

AGÊNCIA DE BIBLIOTECAS E COLEÇÕES DIGITAIS. Indicadores e Métricas. Disponível em: <https://www.abcd.usp.br/apoio-pesquisador/indicadores-pesquisa/lista-indicadores-bibliometricos/>. .

AHUJA, V.; YANG, J.; SKITMORE, M.; SHANKAR, R. An empirical test of causal relationships of factors affecting ICT adoption for building project management: An Indian SME case study. **Construction Innovation**, v. 10, n. 2, p. 164–180, 2010.

ALBAYATI, M. B.; ALTAMIMI, A. M. An empirical study for detecting fake facebook profiles using supervised mining techniques. **Informatica (Slovenia)**, v. 43, n. 1, p. 77–86, 2019. Slovene Society Informatika. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85066744264&doi=10.31449%2Finf.v43i1.2319&partnerID=40&md5=8b44f61d676fe3751d2ca40a9a07840a>. .

ALBAZ, A.; DONDI, M.; RIDA, T.; SCHUBERT, J. Unlocking growth in small and medium enterprises | McKinsey. **McKinsey&Company Public & Social Sector**, , n. June, p. 1–20, 2020. Disponível em: <https://www.mckinsey.com/industries/public-and-social-sector/our-insights/unlocking-growth-in-small-and-medium-size-enterprises>. .

AWA, H. O.; UKO, J. P.; UKOHA, O. An Empirical Study of Some Critical Adoption Factors of ERP Software. **International Journal of Human-Computer Interaction**, v. 33, n. 8, p. 609–622, 2017.

AZEVEDO, A.; ALMEIDA, A. H. Grasp the challenge of digital transition in smes—a training course geared towards decision-makers. **Education Sciences**, v. 11, n. 4, p. 1–20, 2021.

BANCO CENTRAL DO BRASIL. Conversor de Moedas. Disponível em: <https://www.bcb.gov.br/conversao>. Acesso em: 21/2/2023.

BARRETTO, L. Micro e pequenas empresas geram 27% do PIB do Brasil. , p. 1–3, 2015. Disponível em: <www.sebrae.com.br>. .

BAYO-MORIONES, A.; BILLÓN, M.; LERA-LÓPEZ, F. Perceived performance effects of ICT in manufacturing SMEs. **Industrial Management and Data Systems**, v. 113, n. 1, p. 117–135, 2013.

BERINATO, S. Data Science and the Art of Persuasion. **Harvard Business Review**, p. 126–137, 2019. Disponível em: <https://hbr.org/2019/01/data-science-and-the-art-of-persuasion>. .

BERTO, R. M. V. D. S.; NAKANO, D. N. Metodologia da Pesquisa e a Engenharia de Produção. **Encontro Nacional de Engenharia de Produção**, p. 7, 1998.

BHASKARAN, S. Structured case studies: Information communication technology adoption by small-to-medium food enterprises. **British Food Journal**, v. 115, n. 3, p. 425–447, 2013.

BLEI, D. M.; NG, A. Y.; JORDAN, M. I. Latent Dirichlet Allocation. **Journal of Machine Learning Research**, v. 3, p. 139–159, 2002.

BOONSIRITOMACHAI, W.; MCGRATH, G. M.; BURGESS, S. Exploring business intelligence and its depth of maturity in Thai SMEs. **Cogent Business and Management**, v. 3, n. 1, 2016. Cogent. Disponível em: <http://dx.doi.org/10.1080/23311975.2016.1220663>. .

BRASIL. **Lei 123 de 14 de Dezembro de 2006**. Presidência da República, 2006.

BRASIL. **Lei 11.638 de 28 de Dezembro de 2007**. Presidência da República, 2007.

BROUS, P.; JANSSEN, M.; KRANS, R. Data Governance as Success Factor for Data Science. (S. H. P. I. D. Y. K. M. M. Hattingh M. Matthee M., Org.)**Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)**, v. 12066 LNCS, p. 431–442, 2020. Springer. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85084927389&doi=10.1007%2F978-3-030-44999-5_36&partnerID=40&md5=d67a886a1699c7eb2de7f02297a3b502>. .

CANTÚ, L. Z.; CRIADO, J. R.; CRIADO, A. R. Generation and transfer of knowledge in IT-related SMEs. **Journal of Knowledge Management**, v. 13, n. 5, p. 243–256, 2009.

CASSIA, F.; MAGNO, F. Cross-border e-commerce as a foreign market entry mode among SMEs: the relationship between export capabilities and performance. **Review of International Business and Strategy**, 2022.

CAUCHICK MIGUEL, P. A.; FLEURY, A.; PEREIRA MELLO, C. H.; et al. **Metodologia de Pesquisa em Engenharia de Produção e Gestão de Operações**. 2018.

CHABBOUH, H.; BOUJELBENE, Y. Open innovation in SMEs: The mediating role between human capital and firm performance. **The Journal of High Technology Management Research**, v. 31, n. 2, p. 100391, 2020. Elsevier Inc. Disponível em: <http://dx.doi.org/10.1016/j.jcrc.2015.01.006>. .

CHAN, C. M. L.; TEOH, S. Y.; YEOW, A.; PAN, G. Agility in responding to disruptive digital innovation: Case study of an SME. **Information Systems Journal**, v. 29, n. 2, p. 436–455, 2019.

CHAN, Y. E.; DENFORD, J. S.; JIN, J. Y. Competing Through Knowledge and Information Systems Strategies: A Study of Small and Medium-Sized Firms. **Journal of Information and Knowledge Management**, v. 15, n. 3, 2016.

CHANIOTAKIS, V.; KOUMAKIS, L.; KONDYLAKIS, H.; et al. Predictive analytics based on open source technologies for acute respiratory distress syndrome. In: S. L. K. B. T. A. S. P. O. J. L. Almeida J.R. Gonzalez A.R. (Org.); Proceedings - IEEE Symposium on Computer-Based Medical Systems. **Anais...** . v. 2021-June, p.68–73, 2021. Institute of Electrical and Electronics Engineers Inc. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85110917953&doi=10.1109%2FCBMS52027.2021.00019&partnerID=40&md5=f5ca1

d23f8d86a105abae408eaf2a552>. .

CHEGE, S. M.; WANG, D.; SUNTU, S. L. Impact of information technology innovation on firm performance in Kenya. **Information Technology for Development**, v. 26, n. 2, p. 316–345, 2020. Taylor & Francis. Disponível em: <https://doi.org/10.1080/02681102.2019.1573717>. .

CHEN, J. C.; CHEN, T.-L.; LIU, W.-J.; CHENG, C. C.; LI, M.-G. Combining empirical mode decomposition and deep recurrent neural networks for predictive maintenance of lithium-ion battery. **Advanced Engineering Informatics**, v. 50, 2021. Elsevier Ltd. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85115657777&doi=10.1016%2Fj.aei.2021.101405&partnerID=40&md5=08a9a818040554fd1f8c5b4ce87b77a8>. .

CORREIA, C.; PORTELA, F.; SANTOS, M. F.; SILVA, Á. Data science analysis of healthcare complaints. (R. A. R. L. P. Adeli H. Costanzo S., Org.)**Advances in Intelligent Systems and Computing**, v. 747, p. 176–185, 2018. Springer Verlag. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85045306677&doi=10.1007%2F978-3-319-77700-9_18&partnerID=40&md5=b0779c82773bc5aca49d86ba5d58b084>. .

CORROCHER, N.; FONTANA, R. Objectives, obstacles and drivers of ICT adoption: What do IT managers perceive? **Information Economics and Policy**, v. 20, n. 3, p. 229–242, 2008.

DELGADO, A.; MAROTTA, A.; GONZÁLEZ, L.; TANSINI, L.; CALEGARI, D. Towards a data science framework integrating process and data mining for organizational improvement. In: M. L. M. L. van Sinderen M. Fill H.-G. (Org.); ICSOFT 2020 - Proceedings of the 15th International Conference on Software Technologies. **Anais...** . p.492–500, 2020. SciTePress. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85091743406&partnerID=40&md5=2f50100649a8dc503b5c727cce3168c4>. .

DEMCHENKO, Y.; BELLOUM, A.; LAAT, C. D.; et al. Customisable data science educational environment: From competences management and curriculum design to virtual labs on-demand. Proceedings of the International Conference on Cloud Computing Technology and Science, CloudCom. **Anais...** . v. 2017-Decem, p.363–368, 2017. IEEE Computer Society. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85044322882&doi=10.1109%2FCloudCom.2017.59&partnerID=40&md5=eb893e3a3b1326ff51adc08a240e824b>. .

DIBRELL, C.; DAVIS, PETER S; CRAIG, J. Fueling innovation through information technology in SMEs. **Journal of Small Business Management**, v. 46, n. 2, p. 203–218, 2008.

DIBRELL, C.; DAVIS, PETER S.; CRAIG, J. Fueling innovation through information technology in SMEs. **Journal of Small Business Management**, v. 46, n. 2, p. 203–218, 2008.

DIETRICH, D.; HELLER, B.; YANG, B. **Data Science & Big Data Analytics**.

Indianapolis, IN, USA: John Wiley & Sons, Inc, 2015.

DIVYA, K.; SIDDHARTHA, B. S.; NIVEDITHA, N. M.; DIVYA, B. M. An Interpretation of Lemmatization and Stemming in Natural Language Processing. **Journal of University of Shanghai for Science and Technology** , v. 22, n. 10, p. 351, 2020. Disponível em: <https://www.researchgate.net/publication/348306833>. .

DONG, J. Q.; YANG, C. H. Business value of big data analytics: A systems-theoretic approach and empirical test. **Information and Management**, v. 57, n. 1, p. 103124, 2020. Elsevier. Disponível em: <https://doi.org/10.1016/j.im.2018.11.001>. .

DUTOT, V.; BERGERON, F.; RAYMOND, L. Information management for the internationalization of SMEs: An exploratory study based on a strategic alignment perspective. **International Journal of Information Management**, v. 34, n. 5, p. 672–681, 2014.

DYERSON, R.; SPINELLI, R.; HARINDRANATH, G. Revisiting IT readiness: An approach for small firms. **Industrial Management and Data Systems**, v. 116, n. 3, p. 546–563, 2016.

EDISON. EDISON Data Science Framework (EDSF). Disponível em: <https://edison-project.eu/edison/edison-data-science-framework-edsf/>. Acesso em: 22/2/2023.

EUROPEAN COMMISSION. Internal Market, Industry, Entrepreneurship and SMEs. SME Performance Review. Disponível em: <https://single-market-economy.ec.europa.eu/smes/sme-definition_en>. .

FALAGAS, M. E.; PITSOUNI, E. I.; MALIETZIS, G. A.; PAPPAS, G. Comparison of PubMed, Scopus, Web of Science, and Google Scholar: strengths and weaknesses. **The FASEB Journal**, v. 22, n. 2, p. 338–342, 2008.

FANTAZY, K. A.; KUMAR, V.; KUMAR, U. An empirical study of the relationships among strategy, flexibility, and performance in the supply chain context. **Supply Chain Management**, v. 14, n. 3, p. 177–188, 2009.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. The KDD Process for Extracting Useful Knowledge from Volumes of Data. **Communications of the ACM**, v. 39, n. 11, p. 27–34, 1996.

FIESP. Perfil da MPEs – FIESP. Disponível em: <https://www.fiesp.com.br/perfil-da-mpes/>. Acesso em: 13/10/2022.

FRANCALANCI, C.; MORABITO, V. IS integration and business performance: The mediation effect of organizational absorptive capacity in SMEs. **Journal of Information Technology**, v. 23, n. 4, p. 297–312, 2008.

GAEDKE NOMURA, A. T.; DE ABREU ALMEIDA, M.; JOHNSON, S.; PRUINELLI, L. Pain Information Model and Its Potential for Predictive Analytics: Applicability of a Big Data Science Framework. **Journal of Nursing Scholarship**, v. 53, n. 3, p. 315–322, 2021. Blackwell Publishing Ltd. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85102647920&doi=10.1111%2Fjnu.12648&partnerID=40&md5=12866976fa36baf7b

e9bd2c836ee9674>. .

GARCÍA, M. DEL M. R.; GARCÍA-NIETO, J.; ALDANA-MONTES, J. F. An ontology-based data integration approach for web analytics in e-commerce. **Expert Systems with Applications**, v. 63, p. 20–34, 2016.

GAUZELIN, S.; BENTZ, H. An examination of the impact of business intelligence systems on organizational decision making and performance: The case of France. **Journal of Intelligence Studies in Business**, v. 7, n. 2, p. 40–50, 2017.

GEORGE, G.; OSINGA, E.; LAVIE, D.; SCOTT, B. Big data and data science methods for management research. **Academy of Management Journal**, v. 59, n. 5, p. 1493–1507, 2016.

GHOBAKHLOO, M.; TANG, S. H. Information system success among manufacturing SMEs: case of developing countries. **Information Technology for Development**, v. 21, n. 4, p. 573–600, 2015.

GONZALES, A. D. C.; VILLANTOY, F. L. P.; SANCHEZ, D. S. M. Data science model for the evaluation of customers of rural savings banks without credit history. Proceedings - 2019 7th International Engineering, Sciences and Technology Conference, IESTEC 2019. **Anais...** . p.329–334, 2019. Institute of Electrical and Electronics Engineers Inc. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85078129279&doi=10.1109%2FIESTEC46403.2019.00067&partnerID=40&md5=cf1867e465f531a82366d5af8decd257>. .

GRANDE, E. U.; ESTÉBANEZ, R. P.; COLOMINA, C. M. The impact of accounting information systems (AIS) on performance measures: Empirical evidence in spanish SMEs. **International Journal of Digital Accounting Research**, v. 11, n. February, p. 25–43, 2011.

GUPTA, V.; AGGARWAL, V.; GUPTA, S.; et al. Visualization and Prediction of Heart Diseases Using Data Science Framework. Proceedings of the 2nd International Conference on Electronics and Sustainable Communication Systems, ICESC 2021. **Anais...** . p.1199–1202, 2021. Institute of Electrical and Electronics Engineers Inc. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85116619475&doi=10.1109%2FICESC51422.2021.9532790&partnerID=40&md5=7181f4c4417e05457e5d0ccfe4c3af20>. .

HAIRUDDIN, H.; NOOR, N. L. M.; MALIK, A. M. A. Why do Microenterprise Refuse to Use Information Technology: A Case of Batik Microenterprises in Malaysia. **Procedia - Social and Behavioral Sciences**, v. 57, p. 494–502, 2012. Elsevier B.V.

HAN, H.; TRIMI, S. Towards a data science platform for improving SME collaboration through Industry 4.0 technologies. **Technological Forecasting and Social Change**, v. 174, n. January 2021, p. 121242, 2022. Elsevier Inc. Disponível em: <https://doi.org/10.1016/j.techfore.2021.121242>. .

HANSEN, E. B.; BØGH, S. Artificial intelligence and internet of things in small and medium-sized enterprises: A survey. **Journal of Manufacturing Systems**, v. 58, n. October 2019, p. 362–372, 2021.

HARRIGAN, P.; RAMSEY, E.; IBBOTSON, P. Critical factors underpinning the e-CRM activities of SMEs. **Journal of Marketing Management**, v. 27, n. 5–6, p. 503–529, 2011.

HARRIGAN, P.; RAMSEY, E.; IBBOTSON, P. Exploring and explaining SME marketing: Investigating e-CRM using a mixed methods approach. **Journal of Strategic Marketing**, v. 20, n. 2, p. 127–163, 2012a.

HARRIGAN, P.; RAMSEY, E.; IBBOTSON, P. Entrepreneurial marketing in SMEs: The key capabilities of e-CRM. **Journal of Research in Marketing and Entrepreneurship**, v. 14, n. 1, p. 40–64, 2012b.

HAYAJNEH, J. A. M.; ELAYAN, M. B. H.; ABDELLATIF, M. A. M.; ABUBAKAR, A. M. Impact of business analytics and π-shaped skills on innovative performance: Findings from PLS-SEM and fsQCA. **Technology in Society**, v. 68, n. January, p. 101914, 2022. Elsevier Ltd. Disponível em: <https://doi.org/10.1016/j.techsoc.2022.101914>. .

JAMALI, M. A.; VOGHOUEI, H.; MOHD NOR, N. G. Information technology and survival of SMEs: an emprical study on Malaysian manufacturing sector. **Information Technology and Management**, v. 16, n. 2, p. 79–95, 2015.

JOSEPHNG, P. S. EaaS Optimization: Available yet hidden information technology infrastructure inside medium size enterprise. **Technological Forecasting and Social Change**, v. 132, n. April 2017, p. 165–173, 2018.

JOSHI, M.; HAZELA, B.; SINGH, V. An application of IoT on Hungarian database using data mining techniques: A collaborative approach. Proceedings - 2017 3rd International Conference on Advances in Computing, Communication and Automation (Fall), ICACCA 2017. **Anais...** . v. 2018-Janua, p.1–6, 2018. Institute of Electrical and Electronics Engineers Inc. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85049743442&doi=10.1109%2FICACCAF.2017.8344676&partnerID=40&md5=bfff920d2be4290b1e36ec21807462e5>. .

KELLER, S. A.; SHIPP, S. S.; SCHROEDER, A. D.; KORKMAZ, G. Doing Data Science: A Framework and Case Study. **Harvard Data Science Review**, v. 2, n. 1, 2020. Disponível em: <https://hdsr.mitpress.mit.edu/pub/hnptx6lq>. .

KHALIL, S.; BELITSKI, M. Dynamic capabilities for firm performance under the information technology governance framework. **European Business Review**, v. 32, n. 2, p. 129–157, 2020.

KHAN, N. U.; LI, S.; KHAN, S. Z.; ANWAR, M. Entrepreneurial orientation, intellectual capital, IT capability, and performance. **Human Systems Management**, v. 38, n. 3, p. 297–312, 2019.

KMIECIAK, R.; MICHNA, A.; MECZYNSKA, A. Innovativeness, empowerment and IT capability: Evidence from SMEs. **Industrial Management and Data Systems**, v. 112, n. 5, p. 707–728, 2012.

LÁNYI, B.; HORNYÁK, M.; KRUZSLICZ, F. The effect of online activity on SMEs'

competitiveness. **Competitiveness Review**, v. 31, n. 3, p. 477–496, 2021.

LEE, C.-Y.; CHOU, B.-J.; HUANG, C.-F. Data science and reinforcement learning for price forecasting and raw material procurement in petrochemical industry. **Advanced Engineering Informatics**, v. 51, 2022. Elsevier Ltd. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85118495402&doi=10.1016%2Fj.aei.2021.101443&partnerID=40&md5=d68d01172d043bba2afa203547843399>. .

LEE, C.-Y.; TSAI, T.-L. Data science framework for variable selection, metrology prediction, and process control in TFT-LCD manufacturing. **Robotics and Computer-Integrated Manufacturing**, v. 55, p. 76–87, 2019. Elsevier Ltd. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85050872315&doi=10.1016%2Fj.rcim.2018.07.013&partnerID=40&md5=e71db1c8960f584f96cdc051bc5fa3cb>. .

LEI, H.; O'CONNELL, R.; EHWERHEMUEPHA, L.; et al. Agile clinical research: A data science approach to scrumban in clinical medicine. **Intelligence-Based Medicine**, v. 3–4, 2020. Elsevier B.V. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85112231323&doi=10.1016%2Fj.ibmed.2020.100009&partnerID=40&md5=ac5ff500db38dd67e822cabcdc54db93>. .

LOPEZ-NICOLAS, C.; SOTO-ACOSTA, P. Analyzing ICT adoption and use effects on knowledge creation: An empirical investigation in SMEs. **International Journal of Information Management**, v. 30, n. 6, p. 521–528, 2010. Elsevier Ltd. Disponível em: <http://dx.doi.org/10.1016/j.ijinfomgt.2010.03.004>. .

MARJANOVIC, O.; DINTER, B. Learning from the history of business intelligence and analytics research at HICSS: A semantic text-mining approach. **Communications of the Association for Information Systems**, v. 43, n. 1, p. 775–791, 2018.

MARRAPU, H. K.; MARAM, B.; REDDI, P. New Analytic Framework of Public Mental Health Prediction Using Data Science. 1st IEEE International Conference on Smart Technologies and Systems for Next Generation Computing, ICSTSN 2022. **Anais...** , 2022. Institute of Electrical and Electronics Engineers Inc. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85130008819&doi=10.1109%2FICSTSN53084.2022.9761324&partnerID=40&md5=6ee9239ab810983b0d136e607d4a34d3>. .

MARUNGO, F.; ROBERTSON, S.; QUON, H.; et al. Creating a data science platform for developing complication risk models for personalized treatment planning in radiation oncology. In: S. R. H. Bui T.X. (Org.); Proceedings of the Annual Hawaii International Conference on System Sciences. **Anais...** . v. 2015-March, p.3132–3140, 2015. IEEE Computer Society. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-84944269566&doi=10.1109%2FHICSS.2015.378&partnerID=40&md5=f7762bf96d1ba62576d69319e40ee75d>. .

MEMON, A.; AN, Z. Y.; MEMON, M. Q.; YAN, B. IT capability, capital availability and

firm performance. **Human Systems Management**, v. 38, n. 3, p. 221–233, 2019.

MIGDADI, M.; ZAID, M. K. A.; HUJRAN, O. S. The impact of collaborative technology on organisational performance through intranet use orientations. **Journal of Information and Knowledge Management**, v. 11, n. 1, p. 1–14, 2012.

MINISTÉRIO DA ECONOMIA. Mais de 1,3 milhão de empresas são criadas no país em quatro meses. , p. 20–22, 2022.

MIYAMOTO, M. Application of competitive forces in the business intelligence of Japanese SMEs. **International Journal of Management Science and Engineering Management**, v. 10, n. 4, p. 273–287, 2015.

NABEEL-REHMAN, R.; NAZRI, M. INFORMATION TECHNOLOGY CAPABILITIES AND SMES PERFORMANCE: AN UNDERSTANDING OF A MULTI- MEDIATION MODEL FOR THE MANUFACTURING SECTOR. **Interdisciplinary Journal of Information, Knowledge, and Management**, v. 14, p. 253–276, 2019.

NAUDÉ, P.; ZAEFARIAN, G.; NAJAFI TAVANI, Z.; NEGHABI, S.; ZAEFARIAN, R. The influence of network effects on SME performance. **Industrial Marketing Management**, v. 43, n. 4, p. 630–641, 2014.

NEIROTTI, P.; PAOLUCCI, E.; RAGUSEO, E. Is it all about size? Comparing organisational and environmental antecedents of IT assimilation in small and medium-sized enterprises. **International Journal of Technology Management**, v. 61, n. 1, p. 82–108, 2013.

NEIROTTI, P.; RAGUSEO, E. On the contingent value of IT-based capabilities for the competitive advantage of SMEs: Mechanisms and empirical evidence. **Information and Management**, v. 54, n. 2, p. 139–153, 2017. Elsevier B.V. Disponível em: <http://dx.doi.org/10.1016/j.im.2016.05.004>. .

NGUYEN, T. H. Information technology adoption in SMEs: an integrated framework. **International Journal of Entrepreneurial Behavior & Research**, v. 15, n. 2, p. 162–186, 2009. Disponível em: <https://www.emerald.com/insight/content/doi/10.1108/13552550910944566/full/html>. .

NICOLAS, C.; KIM, J.; CHI, S. Natural language processing-based characterization of top-down communication in smart cities for enhancing citizen alignment. **Sustainable Cities and Society**, v. 66, 2021.

OLMOS, L. E.; TADEO, M. S.; VLACHOGIANNIS, D.; et al. A data science framework for planning the growth of bicycle infrastructures. **Transportation Research Part C: Emerging Technologies**, v. 115, p. 102640, 2020. Elsevier Ltd. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85083312816&doi=10.1016%2Fj.trc.2020.102640&partnerID=40&md5=65f1a4e8dc0f670c49a8dd2708912fa6>. .

OMARA-EVES, J. THOMAS, J. MCNAUGHT, M. MIWA, S. A. Using text mining for study identification in systematic reviews: a systematic review of current approaches. **SIGGRAPH 2016 - ACM SIGGRAPH 2016 Posters**, p. 1–22, 2016.

OMERZEL, D. G.; ANTONČIČ, B. Critical entrepreneur knowledge dimensions for the SME performance. **Industrial Management and Data Systems**, v. 108, n. 9, p. 1182–1199, 2008.

ONI, O.; PAPAZAFEIROPOULOU, A. Diverse views on IT innovation diffusion among SMEs: Influencing factors of broadband adoption. **Information Systems Frontiers**, v. 16, n. 4, p. 729–747, 2014.

PALANIVEL RAJAN D., P. D@GMAIL. CO.; BASWARAJ, D.; VELLIANGIRI, S.; KARTHIKEYAN, P. Next Generations Data Science Application and Its Platform. Proceedings - International Conference on Smart Electronics and Communication, ICOSEC 2020. **Anais...** . p.891–897, 2020. Institute of Electrical and Electronics Engineers Inc. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85094210314&doi=10.1109%2FICOSEC49089.2020.9215245&partnerID=40&md5=92177514170e889a2fb22374ba2cc217>. .

PATTERSON, E.; MCBURNEY, R.; SCHMIDT, H.; et al. Dataflow representation of data analyses: Toward a platform for collaborative data science. **IBM Journal of Research and Development**, v. 61, n. 6, 2017. IBM Corporation. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85038632457&doi=10.1147%2FJRD.2017.2736278&partnerID=40&md5=6e5770c2098ff328424e5990d4584e5d>. .

PELLETIER, C.; CLOUTIER, L. M. Conceptualising digital transformation in SMEs: an ecosystemic perspective. **Journal of Small Business and Enterprise Development**, v. 26, n. 6–7, p. 855–876, 2019.

PEÑA-VINCES, J. C.; CEPEDA-CARRIÓN, G.; CHIN, W. W. Effect of ITC on the international competitiveness of firms. **Management Decision**, v. 50, n. 6, p. 1045–1061, 2012.

PERRONI, M. G. **ABORDAGEM DE PROCESSOS PARA A MEDIÇÃO E CONTROLE DO DESEMPENHO ENERGÉTICO EM MANUFATURA**, 2017. PONTIFÍCIA UNIVERSIDADE CATÓLICA DO PARANÁ. Disponível em: <https://ejournal.poltektegal.ac.id/index.php/siklus/article/view/298%0Ahttp://repositorio.unan.edu.ni/2986/1/5624.pdf%0Ahttp://dx.doi.org/10.1016/j.jana.2015.10.005%0Ahttp://www.biomedcentral.com/1471-2458/12/58%0Ahttp://ovidsp.ovid.com/ovidweb.cgi?T=JS&P>. .

PHILEMOTTE, C. The Advent of Data Science Platforms. **IEEE Potentials**, v. 39, n. 6, p. 28–31, 2020. Institute of Electrical and Electronics Engineers Inc. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85096400164&doi=10.1109%2FMPOT.2020.3014596&partnerID=40&md5=572eaa78578860f9132e5ce5bd1b004d>. .

PINTO, M. M. A. **Proposta de framework de uma cadeia de suprimentos verde a partir da transferência de tecnologia**, 2016. UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ PROGRAMA.

PRAKASH, C. S.; MADHU BALA, M.; RUDRA, A. Data Science Framework - Heart

Disease Predictions, Variant Models and Visualizations. 2020 International Conference on Computer Science, Engineering and Applications, ICCSEA 2020. **Anais...** , 2020. Institute of Electrical and Electronics Engineers Inc. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85089774847&doi=10.1109%2FICCSEA49143.2020.9132920&partnerID=40&md5=b895f1add4bf3ffa19b8a4145fad40db>. .

PRATONO, A. H. Strategic orientation and information technological turbulence: Contingency perspective in SMEs. , 2017.

PUKLAVEC, B.; OLIVEIRA, T.; POPOVIČ, A. Industrial Management & Data Systems Article information : Understanding the Determinants of Business Intelligence System Adoption Stages : An Empirical Study of SMEs. **Industrial Management & Data Systems**, v. 118, n. 1, p. 236–261, 2018.

QIN, Y.; YANG, H.; GUO, M.; GUO, M. An Advanced Data Science Model Based on Big Data Analytics for Urban Driving Cycle Construction in China. ACM International Conference Proceeding Series. **Anais...** . p.1–7, 2020. Association for Computing Machinery. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85086264502&doi=10.1145%2F3398329.3398330&partnerID=40&md5=c11f752e169a37026a6ea14fae805c29>. .

QURESHIL, S.; KAMAL, M.; WOLCOTT, P. Information technology interventions for growth and competitiveness in micro-enterprises. **International Journal of Enterprise Information Systems**, v. 5, n. 2, p. 72–95, 2009.

RAMASAMY, A.; SISAY, B.; BAHIRU, A. A Data Science Framework for Data Quality Assessment and Inconsistency Detection. **International Journal of Advanced Computer Science and Applications**, v. 12, n. 4, p. 605–613, 2021. Science and Information Organization. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85105812928&doi=10.14569%2FIJACSA.2021.0120476&partnerID=40&md5=f500f8a8bd2b8c7b05aa0ec4bc2dba57>. .

RAMAYAH, T.; LING, N. S.; TAGHIZADEH, S. K.; RAHMAN, S. A. Factors influencing SMEs website continuance intention in Malaysia. **Telematics and Informatics**, v. 33, n. 1, p. 150–164, 2016. Elsevier Ltd. Disponível em: <http://dx.doi.org/10.1016/j.tele.2015.06.007>. .

RASEL, F. Combining Information Technology and Decentralized Workplace Organization: SMEs versus Larger Firms. **International Journal of the Economics of Business**, v. 23, n. 2, p. 199–241, 2016.

REHMAN, N.; RAZAQ, S.; FAROOQ, A.; ZOHAIB, N. M.; NAZRI, M. Information technology and firm performance: mediation role of absorptive capacity and corporate entrepreneurship in manufacturing SMEs. **Technology Analysis and Strategic Management**, v. 32, n. 9, p. 1049–1065, 2020. Taylor & Francis. Disponível em: <https://doi.org/10.1080/09537325.2020.1740192>. .

ROQUE, C.; LOURENÇO CARDOSO, J.; CONNELL, T.; SCHERMERS, G.; WEBER, R. Topic analysis of Road safety inspections using latent dirichlet allocation:

A case study of roadside safety in Irish main roads. **Accident Analysis and Prevention**, v. 131, n. July, p. 336–349, 2019.

SANDULLI, F. D.; FERNÁNDEZ-MENÉNDEZ, J.; RODRÍGUEZ-DUARTE, A.; LÓPEZ-SÁNCHEZ, J. I. The productivity payoff of information technology in multimarket SMEs. **Small Business Economics**, v. 39, n. 1, p. 99–117, 2012.

SARAEE, M.; SILVA, C. A new data science framework for analysing and mining geospatial big data. ACM International Conference Proceeding Series. **Anais...** . p.98–102, 2018. Association for Computing Machinery. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85055289855&doi=10.1145%2F3220228.3220236&partnerID=40&md5=1f909c0288a03e2aa607e022eac77994>. .

SEBRAE. Data Sebrae. Disponível em: <https://datasebrae.com.br/totaldeempresas-11-05-2020/>. Acesso em: 21/2/2023.

SETIAWAN, M.; INDIASTUTI, R.; DESTEVANIE, P. Information technology and competitiveness: Evidence from micro, small and medium enterprises in Cimahi District, Indonesia. **International Journal of Entrepreneurship and Small Business**, v. 25, n. 4, p. 475–493, 2015.

SHEHABUDDEEN, N.; PROBERT, D.; PHAAL, R.; PLATTS, K. Representing and approaching complex management issues. **Centre for Technology Management Working Paper Series**, p. 1–20, 2000.

SHERMAN, R. **Business Intelligence Guidebook: From Data Integration to Analytics**. 1º ed. 2014.

SHIAU, W. L.; HSU, P. Y.; WANG, J. Z. Development of measures to assess the ERP adoption of small and medium enterprises. **Journal of Enterprise Information Management**, v. 22, n. 1–2, p. 99–118, 2009.

SIDOLA, A.; KUMAR, P.; KUMAR, D. System dynamics investigation of information technology in small and medium enterprise supply chain. **Journal of Advances in Management Research**, v. 9, n. 2, p. 199–207, 2012.

SINGH, R. K.; LUTHRA, S.; MANGLA, S. K.; UNIYAL, S. Applications of information and communication technology for sustainable growth of SMEs in India food industry. **Resources, Conservation and Recycling**, v. 147, n. April, p. 10–18, 2019. Elsevier. Disponível em: <https://doi.org/10.1016/j.resconrec.2019.04.014>. .

SMITS, P.; SAGOENIE, Y.; CUPPEN, H. Small and Medium-sized Enterprises in Brazil Small and Medium-sized Enterprises in Brazil Introduction: Small and Medium-sized Enterprises in Brazil. , 2018.

SOLUK, J.; KAMMERLANDER, N. Digital transformation in family-owned Mittelstand firms: A dynamic capabilities perspective. **European Journal of Information Systems**, v. 30, n. 6, p. 676–711, 2021. Taylor & Francis. Disponível em: <https://doi.org/10.1080/0960085X.2020.1857666>. .

TAVANA, M.; SHAABANI, A.; JAVIER SANTOS-ARTEAGA, F.; RAEESI VANANI, I.

A review of uncertain decision-making methods in energy management using text mining and data analytics. **Informatik-Spektrum**, v. 42, n. 6, p. 385–386, 2020.

VIJAYKANTH REDDY, T.; SASHI REKHA, K. Deep Leaf Disease Prediction Framework (DLDPF) with Transfer Learning for Automatic Leaf Disease Detection. Proceedings - 5th International Conference on Computing Methodologies and Communication, ICCMC 2021. **Anais...** . p.1408–1415, 2021. Institute of Electrical and Electronics Engineers Inc. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85106023779&doi=10.1109%2FICCMC51019.2021.9418245&partnerID=40&md5=1d1b97aef5d0695ecd0b656359ae26fa>. .

WALSH, G.; SCHUBERT, P.; JONES, C. Enterprise system investments for competitive advantage: An empirical study of Swiss SMEs. **European Management Review**, v. 7, n. 3, p. 180–189, 2010.

WATANABE, R.; FUJII, N.; KOKURYO, D.; et al. A study on support method of consulting service using text mining. **Procedia CIRP**, v. 67, p. 569–573, 2018. Elsevier B.V. Disponível em: <http://dx.doi.org/10.1016/j.procir.2017.12.262>. .

WATSON, A.; BABU, D. S. V; RAY, S. Sanzu: A data science benchmark. In: S. T. G. R. N. R. W. C. Z. H. B.-Y. R. B.-Y. R. H. X. K. J. C. A. T. J. T. M. Nie J.-Y. Obradovic Z. (Org.); Proceedings - 2017 IEEE International Conference on Big Data, Big Data 2017. **Anais...** . v. 2018-Janua, p.263–272, 2017. Institute of Electrical and Electronics Engineers Inc. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85047810708&doi=10.1109%2FBigData.2017.8257934&partnerID=40&md5=7adedcb9265f4a2e1de081e34e08696d>. .

WILKIN, C. L. The role of it governance practices in creating business value in smes. **Journal of Organizational and End User Computing**, v. 24, n. 2, p. 1–17, 2012.

WORLD BANK SME FINANCE. Small and medium enterprises (SMEs) Finance. Disponível em: <https://www.worldbank.org/en/topic/smefinance>. .

YEE, J.; LOW, C. Y.; KOH, C. T.; et al. Data science platform for smart diagnosis of upper limb spasticity. Procedia Manufacturing. **Anais...** . v. 52, p.250–257, 2020. Elsevier B.V. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85100744966&doi=10.1016%2Fj.promfg.2020.11.042&partnerID=40&md5=28a416c509f04e80c37ad7b7ad1bcb41>. .

YUNIS, M.; EL-KASSAR, A. N.; TARHINI, A. Impact of ICT-based innovations on organizational performance: The role of corporate entrepreneurship. **Journal of Enterprise Information Management**, v. 30, n. 1, p. 122–141, 2017.

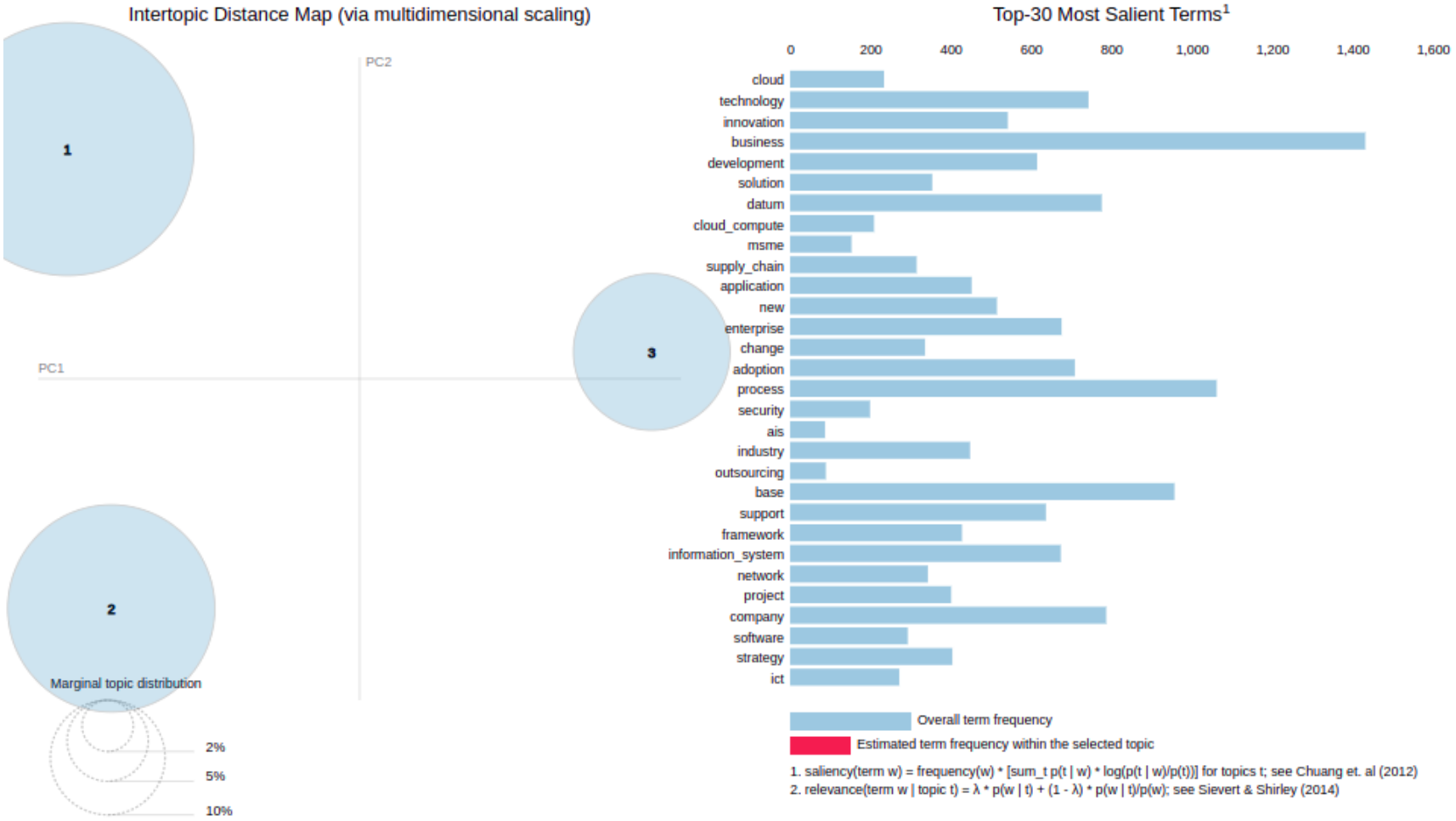
APÊNDICES

APÊNDICE A – Protocolo de entrevista estruturada com especialistas

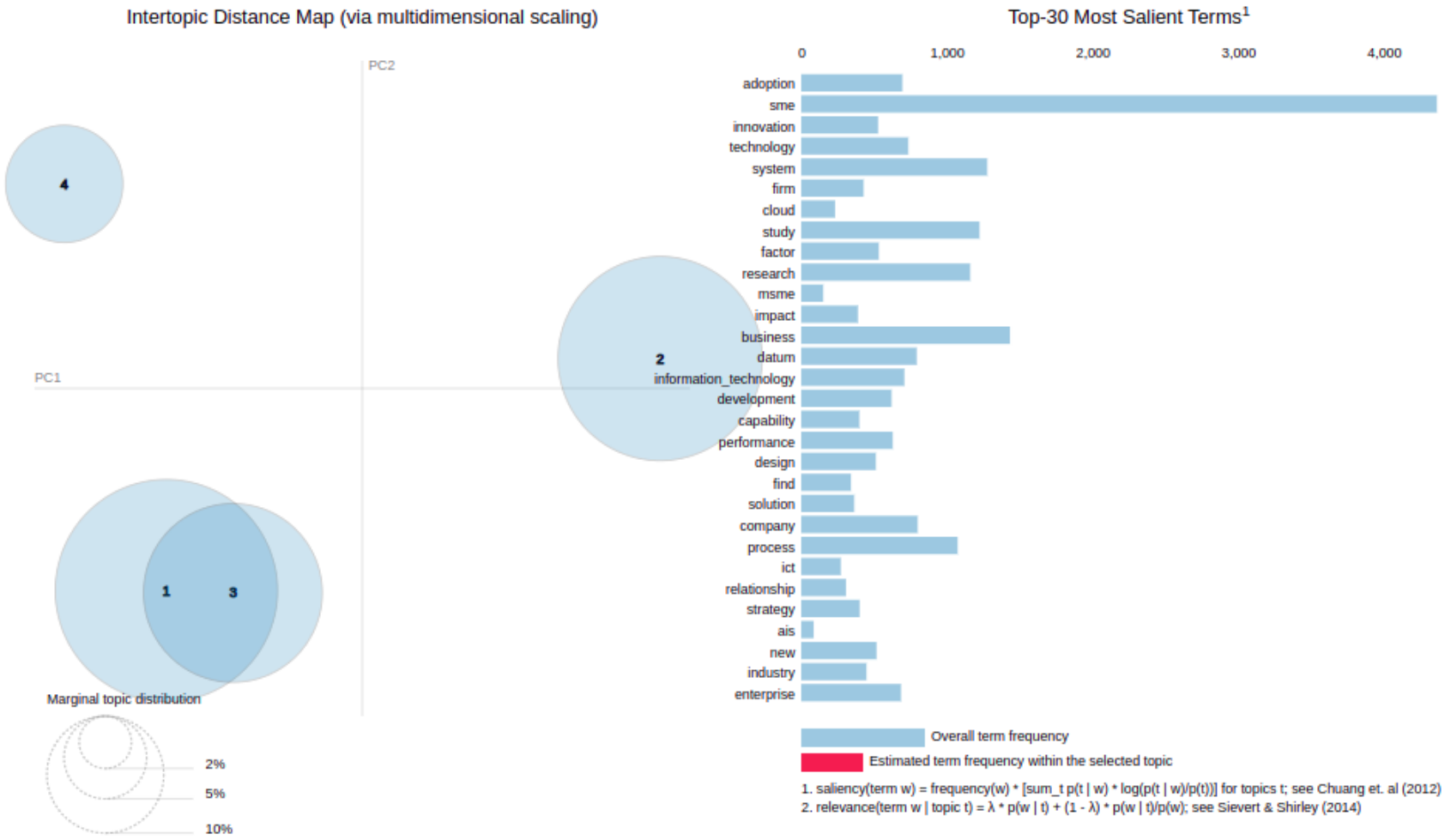
Objetivo geral da entrevista	A revisão de literatura fornece subsídios a cerca de um FDS genérico sem avaliar o ponto de vista brasileiro. Outrossim é a proposta do encurtamento entre o modelo teórico e vivência prática dos problemas enfrentados pelo empreendedor na adoção de tecnologias de vanguarda e que adicionem vantagem competitiva a seu modelo de negócio. Dessa forma, a entrevista tem por objetivo coletar informações que suportem o ajuste do modelo ao contexto de empresas brasileiras e avaliar adequações sob a ótica de profissionais com larga experiência de mercado.
Tempo Total Estimado	50m
Contextualização do FDS (10 minutos)	Para registro, informar ao respondente que todas as suas informações pessoais serão resguardadas e não se fará qualquer menção nesta entrevista à empresa na qual ele atua ou dirige, mas somente sobre sua atuação e experiência profissional com objetivo de lapidar o modelo proposto. Em seguida, far-se-á uma breve introdução explanatória a certa do FDS e direcionamento da pesquisa.
Tópico 1 (10 minutos): Qualificação do	Por favor, informar sua idade, relatar

respondente.	formação acadêmica, experiência profissional (sem mencionar empresa), atuação em PMEs, Segmento majoritário de atuação e Segmento majoritário de atuação somente em PMEs.
Tópico 2 (10 minutos): Percepção de valor do FDS.	Uma vez esclarecido a proposta do FDS, questionar sobre a percepção de valor de entrevistado no caso de adoção por uma empresa genérica e quais pontos de atrito podem ser mitigados com a adoção do FDS para direcionamento de processos de ciência de dados, bem como elementos que não foram contemplados.
Tópico 3 (10 minutos): Adequação à PMEs	Caso tenha experiência com PMEs, quais ajustes devem ser realizados a título de garantir adequação do FCD ao contexto de PMEs.
Tópico 4 (10 minutos): Barreiras de implantação do FDS por PMEs.	Em detrimento da aplicação do FDS como *roadmap* para implantação de uma estrutura de Ciência de Dados em uma PME, quais as principais barreiras, na visão do entrevistado, poder-se-ia enfrentar.
Tópico 5 (10 minutos): Contribuições complementares	Existe ainda algum ponto não avaliado *a priori* pelo entrevistador, mas que o entrevistado entenda como importante mencionar para ajuste do modelo e garantia de adesão?

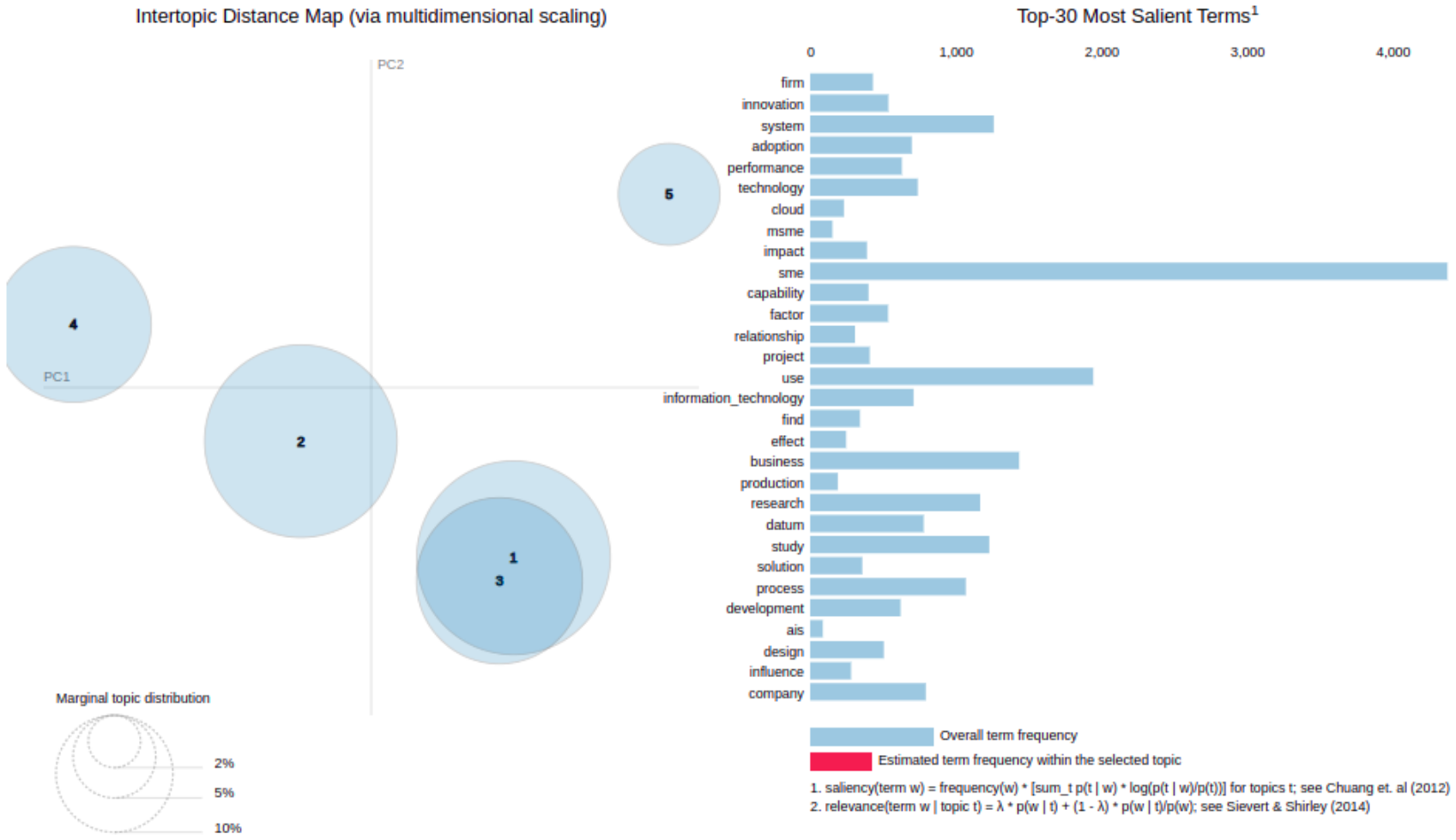
APÊNDICE B – Distâncias Inter tópicos (K = 3)



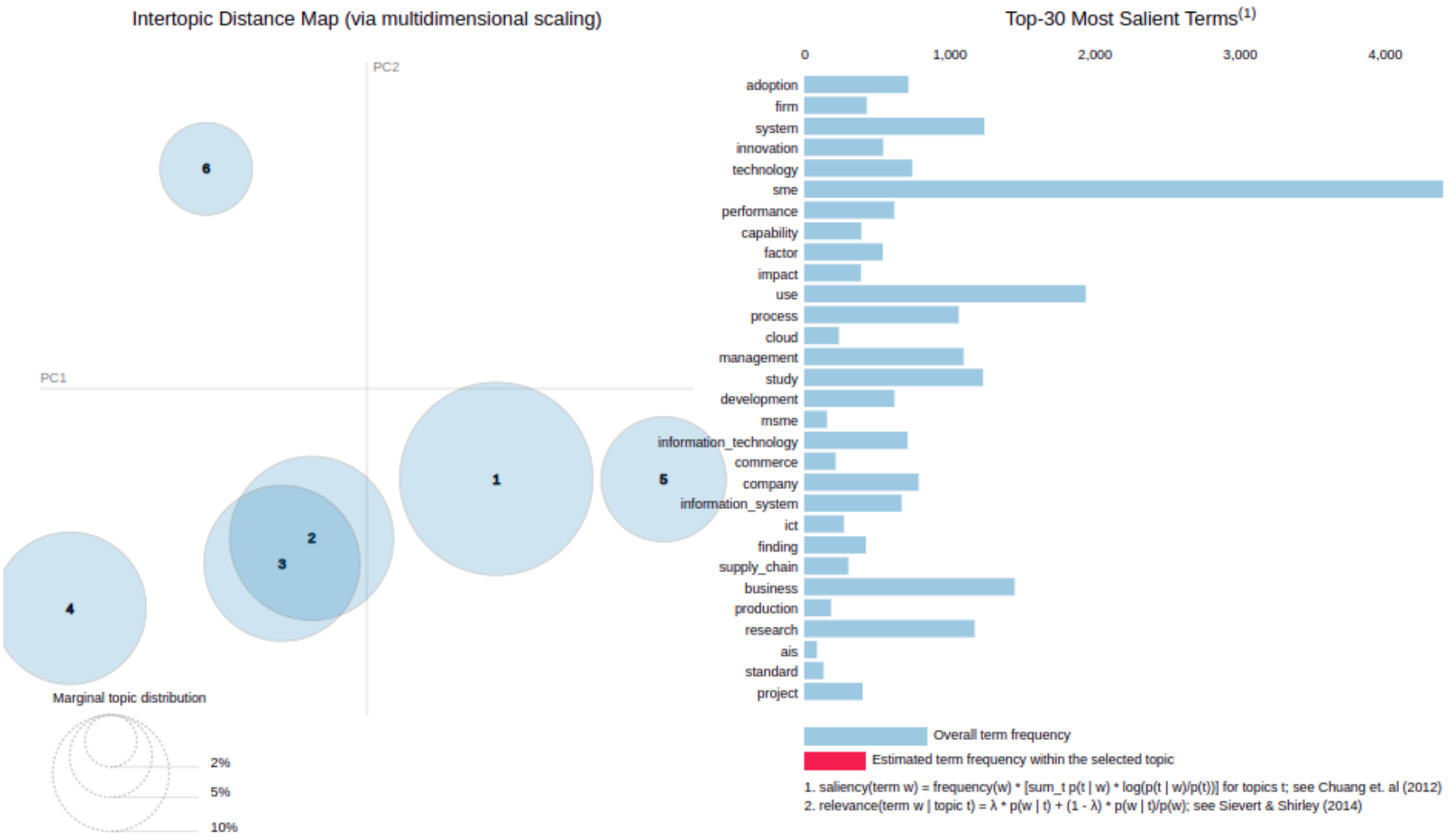
APÊNDICE C – Distâncias Inter tópicos (K = 4)



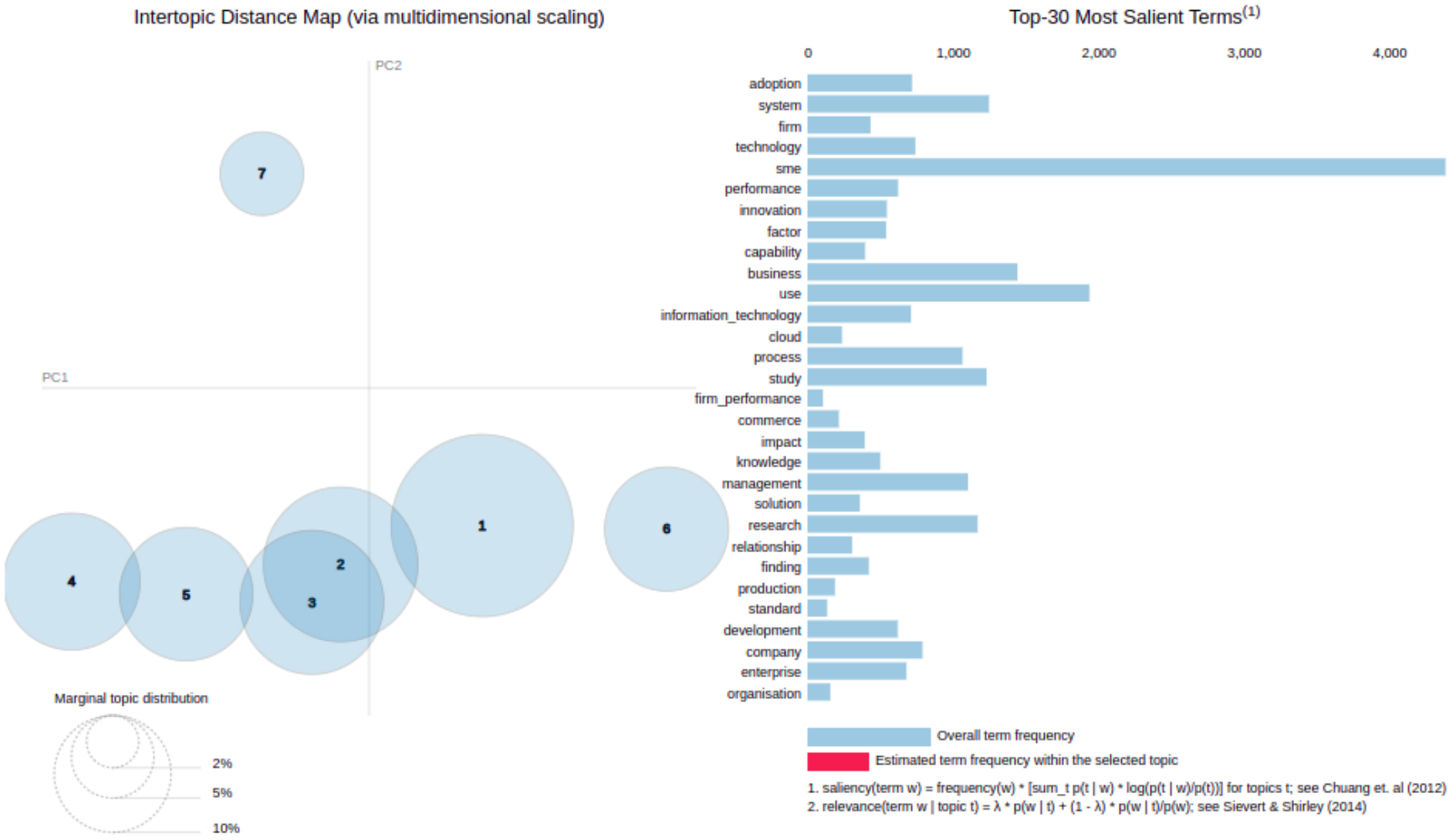
APÊNDICE D – Distâncias Inter tópicos (K = 5)



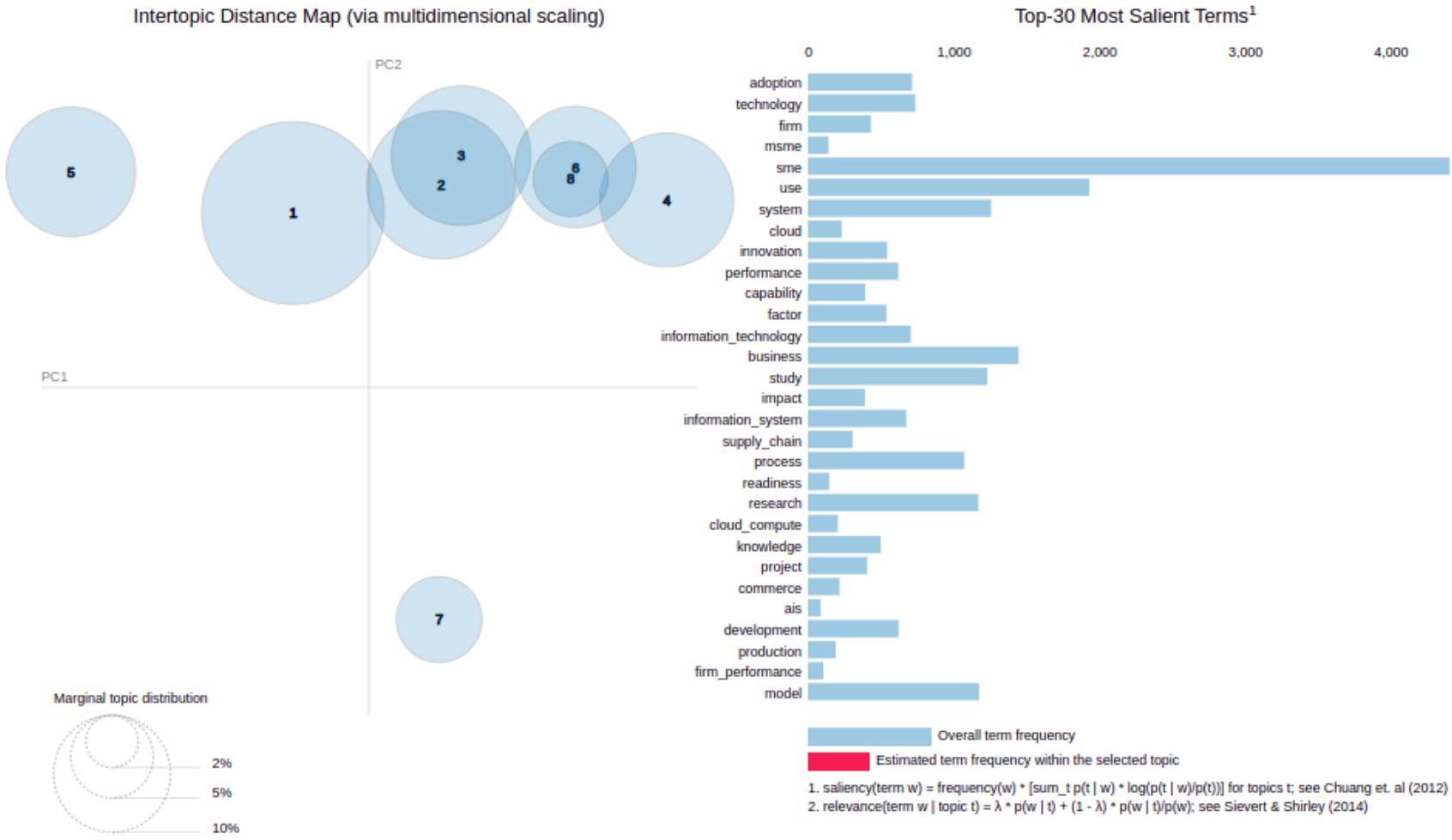
APÊNDICE E – Distâncias Inter tópicos (K = 6)



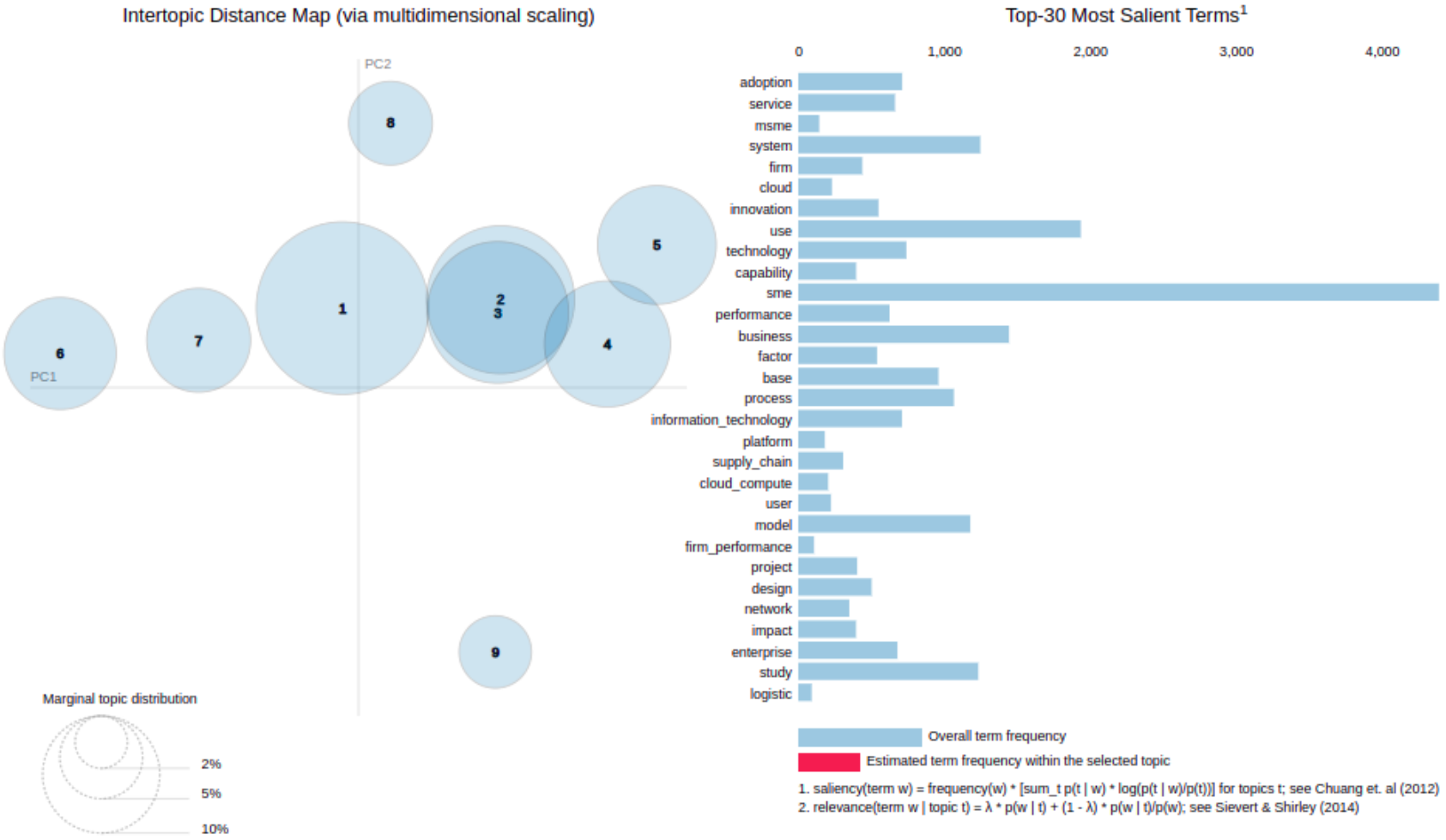
APÊNDICE F – Distâncias Inter tópicos (K = 7)



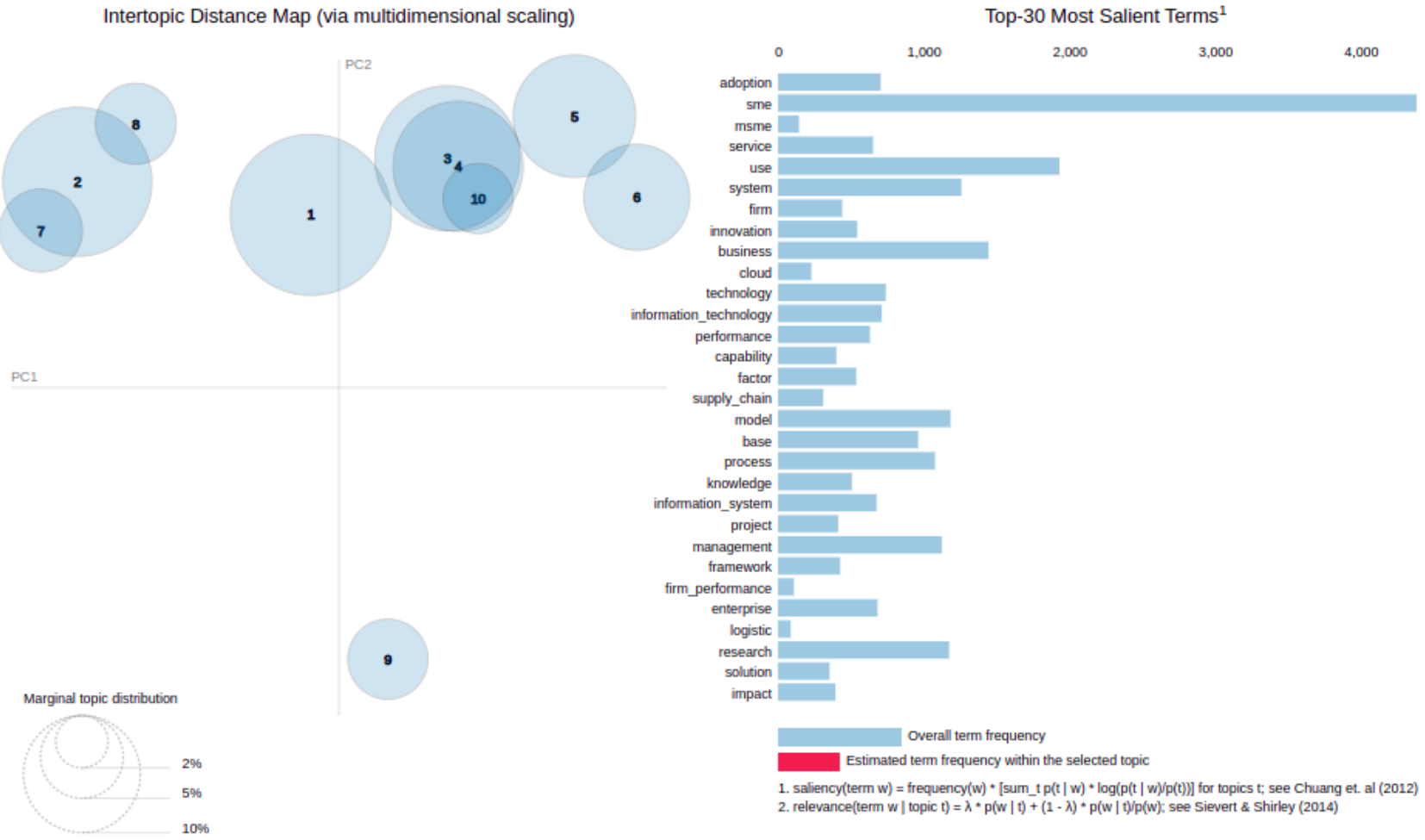
APÊNDICE G – Distâncias Inter tópicos (K = 8)



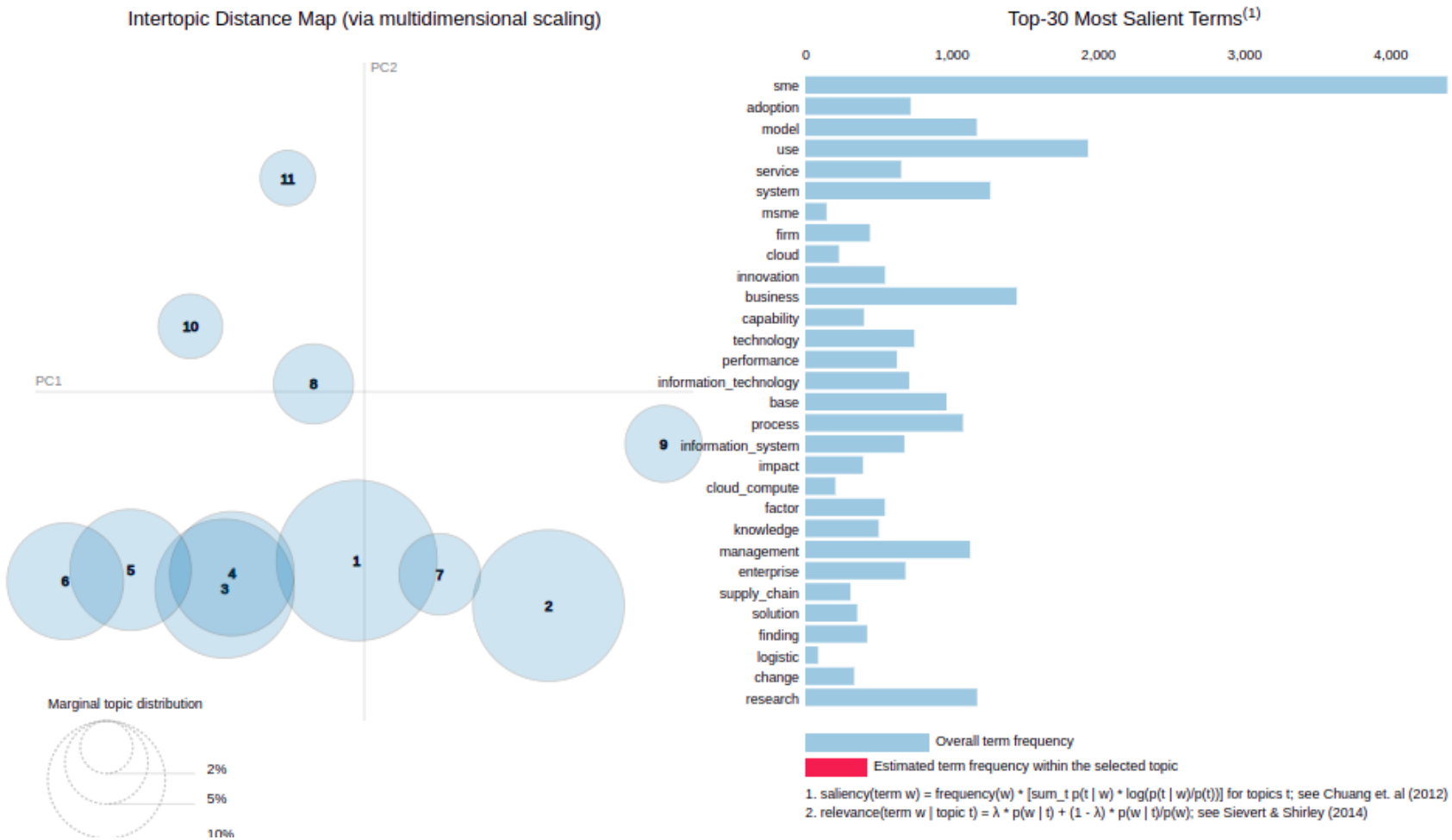
APÊNDICE H – Distâncias Inter tópicos (K = 9)



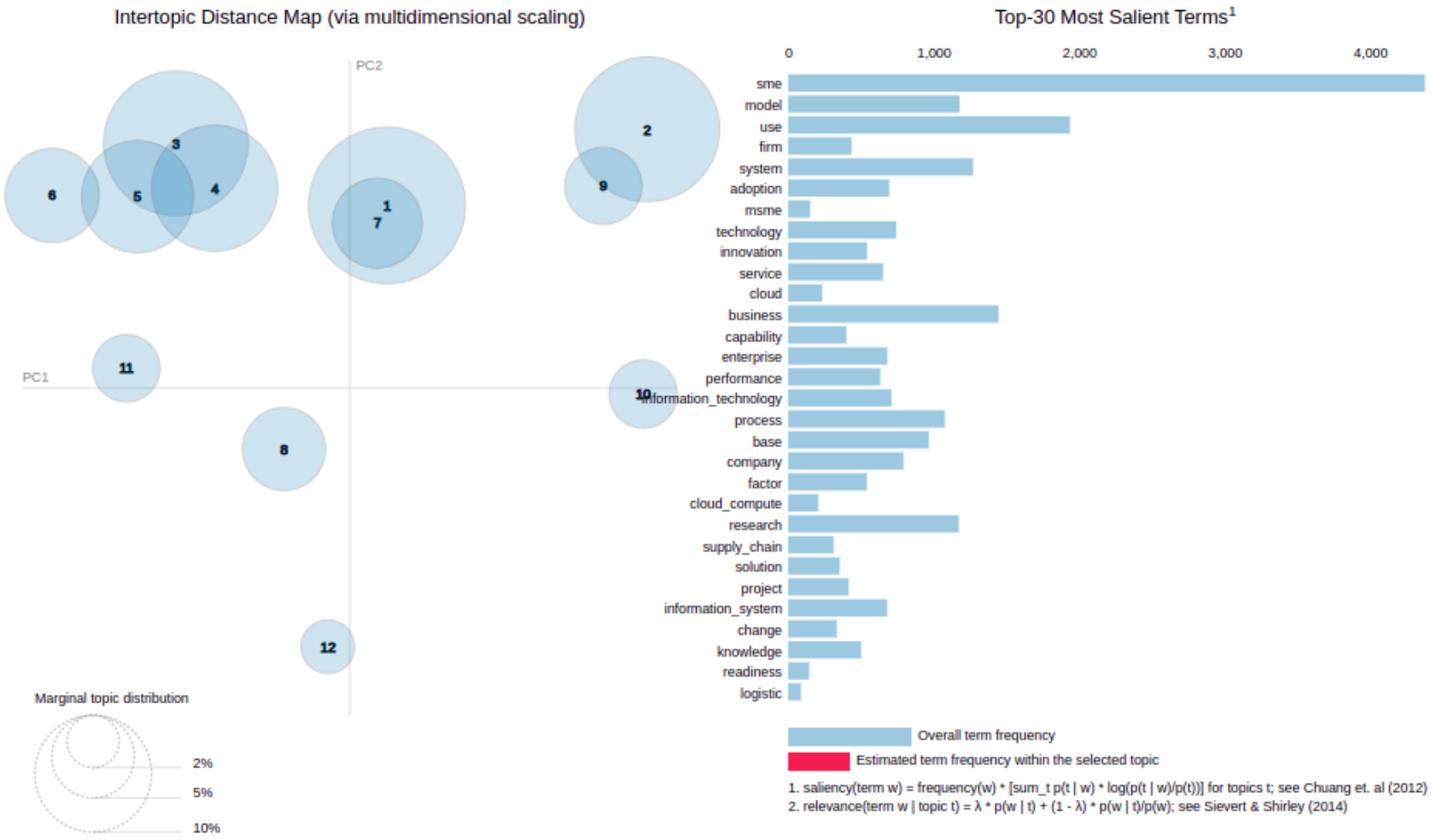
APÊNDICE I – Distâncias Inter tópicos (K = 10)



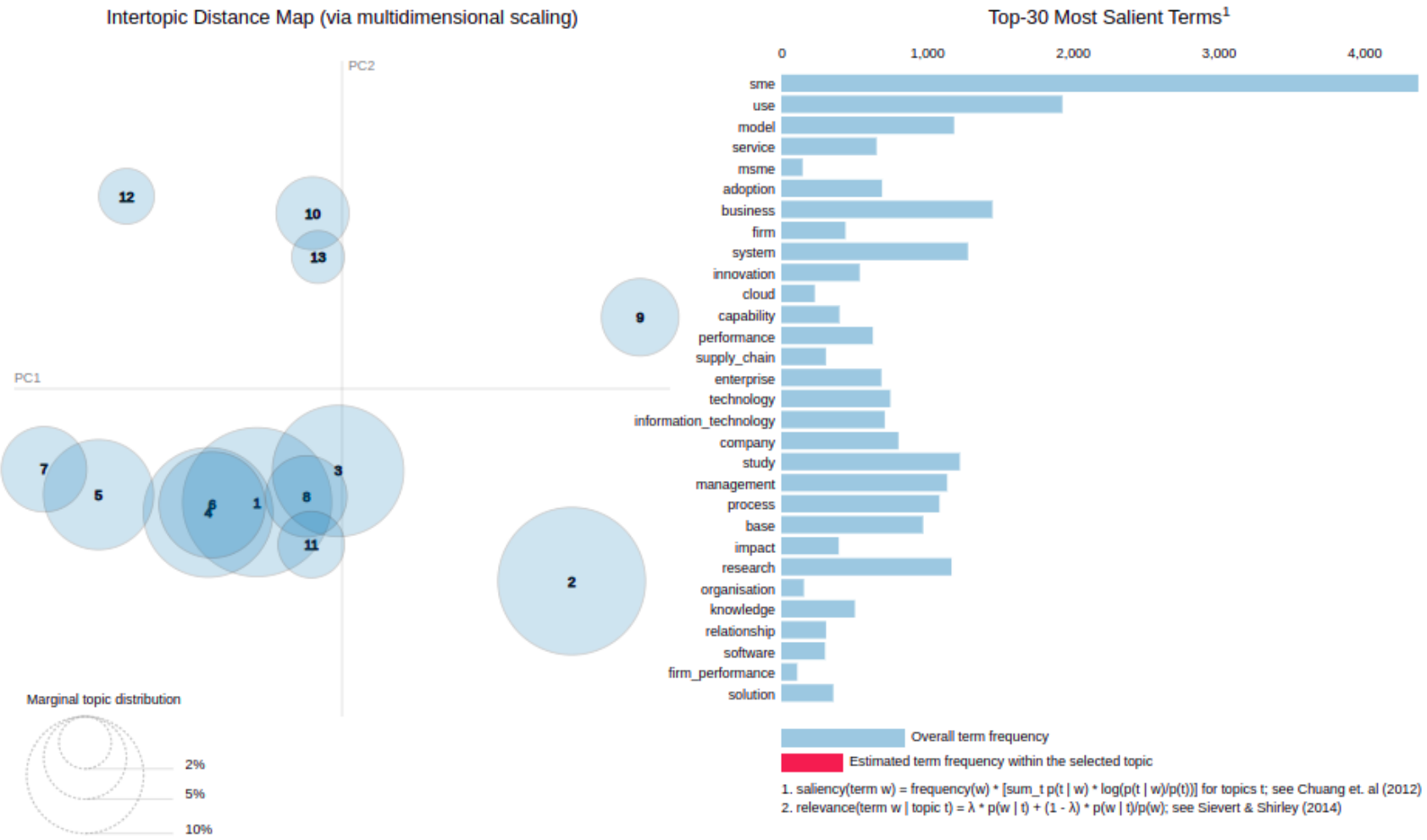
APÊNDICE J – Distâncias Inter tópicos (K = 11)



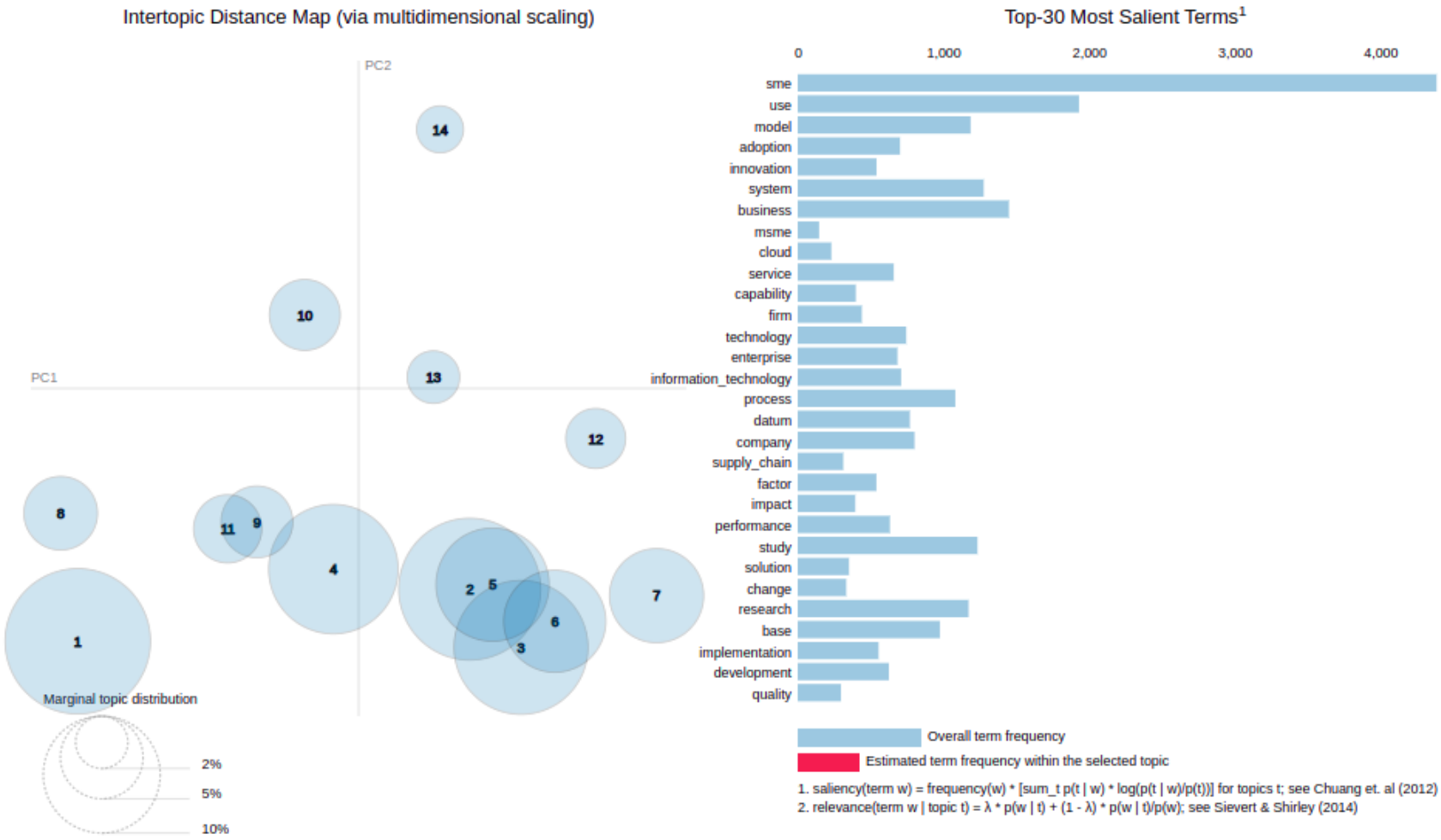
APÊNDICE K – Distâncias Inter tópicos (K = 12)



APÊNDICE L – Distâncias Inter tópicos (K = 13)



APÊNDICE M – Distâncias Inter tópicos (K = 14)



APÊNDICE N – Áreas de Conhecimento Eixo *SMEs & Information Management*

Índice	Áreas	#ocorrências	%ocorrências
1	agronomy and crop science	1	2%
2	animal science and zoology	1	2%
3	applied psychology	3	5%
4	artificial intelligence	8	12%
5	building and construction	1	2%
6	business and international management	6	9%
7	business, management and accounting	5	8%
8	business, management and accounting	2	3%
9	catalysis	1	2%
10	chemistry	1	2%
11	civil and structural engineering	1	2%
12	communication	2	3%
13	computer networks and communications	11	17%
14	computer science	3	5%
15	computer science applications	14	22%
16	control and systems engineering	2	3%
17	decision sciences	2	3%
18	development	2	3%
19	economics and econometrics	3	5%
20	education	1	2%
21	electrical and electronic engineering	2	3%
22	engineering	5	8%
23	engineering	1	2%
24	environmental science	1	2%
25	food science	1	2%
26	geography, planning and development	1	2%
27	hardware and architecture	2	3%
28	human factors and ergonomics	2	3%
29	human-computer interaction	2	3%
30	industrial and manufacturing engineering	9	14%
31	industrial relations	7	11%
32	information systems	19	29%
33	information systems and management	11	17%
34	law	1	2%
35	library and information sciences	10	15%
36	management information systems	16	25%
37	management of technology and innovation	10	15%
38	management science and operations research	4	6%
39	management, monitoring, policy and law	2	3%
40	marketing	10	15%
41	mathematics	1	2%

Índice	Áreas	#ocorrências	%ocorrências
42	mechanical engineering	1	2%
43	media technology	1	2%
44	plant science	1	2%
45	public administration	1	2%
46	public health, environmental and occupational health	1	2%
47	renewable energy, sustainability and the environment	2	3%
48	safety research	1	2%
49	safety, risk, reliability and quality	1	2%
50	sociology and political science	1	2%
51	software	8	12%
52	strategy and management	19	29%
53	theoretical computer science	2	3%
54	waste management and disposal	1	2%