DATA SCIENCE: CONCEITOS E APLICAÇÕES PARA PEQUENAS E MÉDIAS EMPRESAS

MONTEIRO, Miguel Angelo[1]
SOUZA, Leandro de[2]

RESUMO

Este artigo aborda os conceitos fundamentais do *data science*, com a exploração dos conceitos de KDD, *data mining* e suas tarefas específicas, traçando um paralelo com a importância destas ferramentas na tomada de decisão em negócios, sendo o objetivo do estudo investigar o impacto e os desafios na adoção do *data science* para tomada de decisões nas Pequenas e Médias Empresas (PMEs). Para atingir o objetivo o método utilizado foi a pesquisa exploratória qualitativa, com a busca de conceitos e resultados em trabalhos acadêmicos gerando uma discussão a respeito dos posicionamentos e definições resultantes da pesquisa. Após elaboração do referencial teórico, foi possível notar que mesmo com diversos desafios em diferentes áreas enfrentados pelas empresas, a adoção de técnicas e processos orientados a dados traz benefícios para as operações e resultados financeiros para as empresas. As principais barreiras que impedem a disseminação do *data science*, conforme observado nos estudos, são a cultura organizacional, a falta de alinhamento estratégico da liderança, além disso é notável o custo financeiro ou de recursos humanos para obter retornos de uma área de ciência de dados, assim como a falta de conhecimento específico da área, evidenciando tais características também como barreiras relevantes.

PALAVRAS-CHAVE: ciência de dados, *data science*, KDD, pequenas e médias empresas, tomada de decisão.

ABSTRACT

This article addresses the fundamental concepts of data science, exploring the concepts of KDD, data mining and their specific tasks, drawing a parallel with the importance of these tools in business decision-making, especially in Small and Medium-sized Enterprises (SMEs). The aim of the study is to investigate the impact and challenges of adopting data science in SMEs. To achieve the objective, it was conducted a qualitative and exploratory research, seeking for definitions and results in academic works to produce a discussion about the concepts and concerns resulting from the research. After drawing up the theoretical framework, it was possible to see that, despite the various challenges faced by companies in different areas, the adoption of data-driven techniques and processes brings benefits to operations and financial results for companies. The main barriers preventing the dissemination of data science, as observed in the studies, are organizational culture, lack of strategic alignment of leadership, in addition, the financial or human resource cost to obtain returns from a data science area is notable, as it is the lack of specific knowledge in the area, highlighting these characteristics also as relevant barriers.

KEYWORDS: data science, KDD, small and medium-sized enterprises, decision-making process.

1 INTRODUÇÃO

Data science é uma área da ciência da computação que tem gerado resultados interessantes para as empresas ao longo do tempo. O termo mais antigo relacionado ao *data science* é *Knowledge Discovery in Databases* (KDD), ou seja, Descoberta de Conhecimentos em Bancos de Dados, que ainda se aplica quando temos mais interesse no processo e nos algoritmos. A definição do KDD tem origens nos trabalhos de Fayyad *et al*. (1996), quando

[1] Acadêmico de Engenharia de Software do Centro Universitário FAG e-mail: mamonteiro5@minha.fag.edu.br
[2] Docente Orientador do curso de Engenharia de Software – e-mail: leandrosouza@fag.edu.br

os resultados em explorar os dados ainda não eram notoriamente utilizados. As etapas conceituadas neste trabalho em questão apresentadas na fundamentação teórica.

Apesar de não ser uma ciência recente e contar com diversos estudos e processos bem definidos, ainda está em evolução, tem diversos desafios e uma grande abrangência, conforme nota-se em Coenen (2011) e Martínez-Plumed *et al.* (2019).

Mesmo com todo o tempo e a evolução que ocorreu na ciência de dados, ainda há um grande potencial de mercado em termos econômicos. Utilizando dados da pesquisa mais recente da Associação Brasileira de Empresas de Software (ABES) (2023), assim como os dados econômicos divulgados pelos IBGE (2023), é possível perceber que o montante financeiro relacionado ao setor de TI corresponde a 2,5% do PIB nacional, ou seja, é um valor pequeno para uma área que está presente direta ou indiretamente em todos os setores de negócios.

Além disso, na pesquisa da ABES (2023) também são descritos *insights* sobre as tendências de crescimento do setor, sendo a previsão de número 6 bastante positiva para automação e inteligência nos negócios, ponderando que estas ferramentas estão se tornando mais robustas, porém ainda há falta de confiança por parte das empresas para delegar funções críticas a cargo da Inteligência Artificial (IA).

Outra questão relevante consiste no fato que as Pequenas e Médias Empresas (PMEs) dispõem de menos recursos financeiros e humanos, o que significa que não têm capacidade própria para armazenar, atualizar e analisar os dados ou até mesmo terceirizar este tipo de serviço. Saeed (2020) corrobora a ideia de que as principais barreiras para estas empresas são as restrições financeiras, de conhecimento e consciência sobre o *data mining*. Por outro lado, se isto pudesse ser superado, os principais benefícios seriam a flexibilidade, redução de custos, aumento de eficiência, qualidade e vantagens competitivas para estes negócios.

Com este pequeno contexto, este artigo faz uma exploração do tema, com a justificativa de que o *data science* pode gerar resultados positivos para um maior número de empresas, fazendo a correlação do valor dos dados com os ambientes e características de negócios, especialmente nos pequenos e médios negócios, que não tem um vínculo tão forte com a área da tecnologia da informação.

A problemática que este trabalho aborda está relacionada principalmente com a quantidade massiva de dados que são gerados e estão disponíveis de forma bruta nos contextos de negócio de todas as empresas, o que leva à necessidade de um sistema

automatizado para análise de dados de forma inteligente, para que as operações e processos de decisão sejam mais eficientes. Porém, empresas pequenas e médias, principalmente em economias emergentes, têm dificuldades para se adaptar às tecnologias relacionadas ao *data science,* como *Big Data, Business Intelligence, Machine Learning* e *Data Mining* (Saeed, 2020).

Outro aspecto importante a ser analisado, também voltando às companhias pequenas e médias, são as quantidades de empresas que fecham todos os anos. De acordo com o SEBRAE (2023) no grupo das empresas classificadas com micro empreendedores individuais ou (MEIs) 29% fecham após 5 anos de atividade, enquanto no grupo das microempresas (MEs) este valor é de 21,6% e 17% para as empresas de pequeno porte (EPPs). A causa da mortalidade destas empresas apontada na matéria são pouco preparo pessoal, além do planejamento e gestão deficiente do negócio, sendo os dois últimos justamente pontos de melhoria onde o *data science* pode contribuir, buscando responder a seguinte pergunta: De que maneira a adoção de data science está relacionada à melhoria dos processos de negócios?

Para responder esse questionamento propõe-se o seguinte objetivo geral: investigar o impacto e os desafios na adoção de *data science* para tomada de decisões estratégicas em pequenas e médias empresas, focando na melhoria dos processos empresariais e na superação das barreiras comuns enfrentadas por essas empresas.

Este objetivo geral pode ser cumprido por meio de marcos intermediários, determinados pelos objetivos específicos, que seguem:

a. Descrever os conceitos fundamentais relacionados à área de exploração de dados: detalhar o que é *data science,* KDD e *data mining*;
b. Investigar a relação entre a adoção de *data science* e a melhorias dos processos de negócios em PMEs;
c. Explorar as principais barreiras e desafios enfrentados pelas PMEs na adoção de *data science*.

Este trabalho está estruturado da seguinte forma: na introdução é realizada uma contextualização geral do tema, com uma síntese do problema, objetivo e justificativa para o estudo. Na segunda seção é desenvolvida a fundamentação teórica, que relata a pesquisa realizada para responder os objetivos do artigo. Na terceira seção é descrita a metodologia utilizada para produção do trabalho. Na quarta seção são realizadas análises e discussões

sobre as questões e conceitos abordados na fundamentação teórica. Na quinta seção constam as considerações finais.

2 FUNDAMENTAÇÃO TEÓRICA

Neste capítulo serão descritos os conceitos fundamentais a respeito da descoberta de conhecimento em banco de dados, com abordagem de assuntos como *data science* e métodos orientados a dados trazendo correlações para o mundo dos negócios.

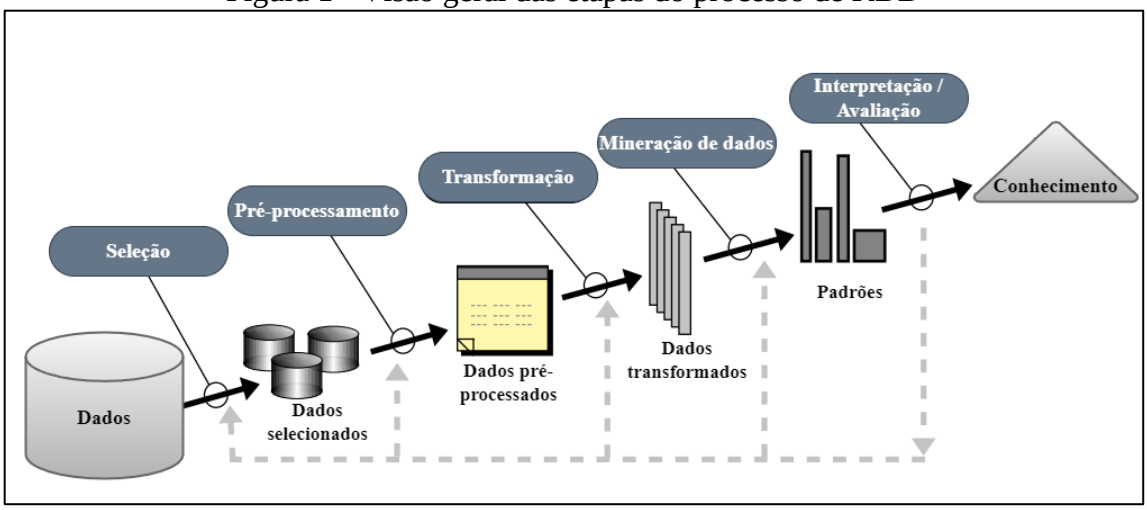
2.1 KDD, DATA MINING E DATA SCIENCE

A ciência de obter informações úteis dos dados é relativamente recente, as origens do *data mining* como uma disciplina das ciências da computação datam do final dos anos 80, quando a comunidade de pesquisa começou a utilizar os métodos e adotar o termo. Mais tarde na década de 90 o termo *data mining* era reconhecido como parte de um processo mais abrangente, denominado *Knowledge Discovery in Databases* (KDD), ou descoberta de conhecimentos em bancos de dados, que consiste em um grupo de atividades mais abrangentes que incluem a coleta e tratamento dos dados para análise (Coenen, 2011).

Para Boscarioli *et al.* (2016), descobrir conhecimento a partir de dados brutos não é uma tarefa trivial, desta afirmação resulta a associação do termo "mineração" para revelar este conhecimento. Para desempenhar esta atividade é necessário conhecimento prévio dos dados, bem como o entendimento do processo de análise e descoberta, as técnicas mais adequadas para "minerar" e dominar as respectivas ferramentas computacionais aplicáveis para cada situação. Ainda são requisitos a percepção correta do ambiente de produção dos dados e os resultados esperados, ou seja, é algo trabalhoso, que necessita de dedicação e tempo.

O processo de KDD é algo mais abrangente, por vezes também confundido com a atividade de *data mining*, sendo essa apenas uma das etapas que compõe o processo. De acordo com Fayyad *et al.* (1996), o processo KDD pode ser definido como um processo não trivial para a identificação de padrões válidos, novos, potencialmente úteis e compreensíveis nos dados. O *data mining* consiste no levantamento dos padrões ou modelos a partir dos dados, sujeito a limitações de eficiência computacional aceitáveis. As etapas do processo de KDD estão representadas na Figura 1.

Figura 1 - Visão geral das etapas do processo de KDD



Fonte: Adaptado de Fayyad *et al.* (1996)

O processo é interativo e iterativo, ou seja, há várias decisões a cargo do usuário sendo que os passos, de forma resumida são (Fayyad, Piatetsky-Shapiro e Padhraic, 1996):

- Aprendizado do domínio da aplicação: inclui conhecimento prévio relevante e os objetivos da aplicação;
- Criar a amostra de dados: seleciona um conjunto para executar a descoberta;
- Limpeza e pré-processamento dos dados: inclui operações básicas de remoção de ruídos ou valores atípicos, além da estratégia para tratamento de informações faltantes ou problemas com o gerenciamento do banco de dados como tipos de dados ou esquemas dos bancos;
- Redução de dados e projeção: encontra recursos para representar os dados, dependendo do objetivo da tarefa, utilizando redução dimensional ou métodos de transformação para redução do número de variáveis;
- Escolha da função de *data mining*: decide o propósito do modelo derivado do algoritmo de mineração de dados (sumarização, classificação, regressão ou agrupamento);
- Escolha do algoritmo de *data mining*: seleção de métodos para buscar pelos padrões nos dados, com a decisão de quais modelos e parâmetros são apropriados e combinam com o critério de KDD, por exemplo, o usuário pode estar mais interessado em entender o modelo do que em sua capacidade preditiva;

- Execução do *data mining*: busca por padrões em uma forma de representação particular, como árvores de classificação ou regras de classificação, regressão, agrupamento, sequenciamento e dependência;
- Interpretação: interpreta os padrões extraídos dos dados, com a possibilidade de retorno a passos anteriores ou visualização dos padrões extraídos, removendo padrões redundantes ou irrelevantes;
- Utilização da descoberta de conhecimento: inclui a incorporação do conhecimento no desempenho do sistema, com a tomada de ações baseada no conhecimento descoberto, ou simplesmente documentação dos resultados.

Com relação às técnicas para extração do conhecimento pode-se afirmar que derivam principalmente do ramo específico da inteligência artificial denominado *Machine Learning*, e são utilizadas junto com estatística para revelar as informações "escondidas" nos dados. É possível encarar o *data mining* como uma aplicação do *Machine Learning* com foco específico nos dados, enquanto o aprendizado de máquina é mais apropriadamente definido com uma tecnologia, focada em mecanismos para que os algoritmos possam aprender (Coenen, 2011).

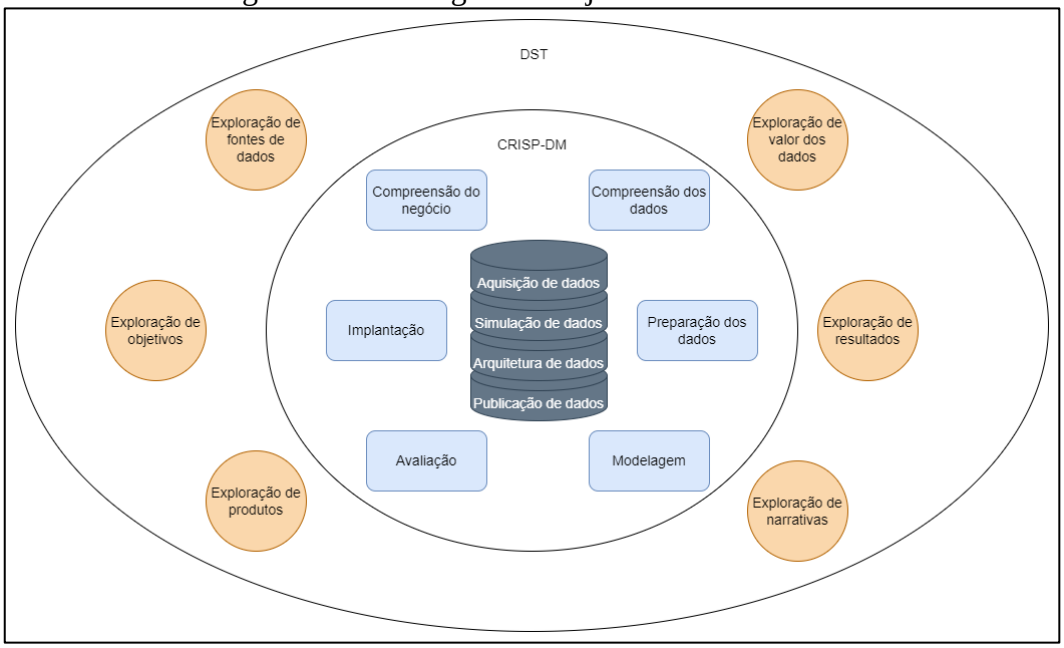
Por outro lado, *data science* é uma abordagem mais abrangente, podendo ser considerada uma disciplina que estuda os dados em todas as suas manifestações, junto com métodos e algoritmos para manipular, analisar, visualizar e enriquece-los. A diferença chave entre os termos, que advém do tempo e evolução da ciência, está na percepção de que o objetivo primário do *data mining* e KDD era o processo em si, enquanto a visão mais moderna do *data science* tem natureza orientada a dados e exploratória (Martínez-Plumed *et al.*, 2019).

A metodologia para mineração de dados denominada *Cross Industry Standard Process for data mining* (CRISP-DM) trata de uma perspectiva orientada para um objetivo, a qual será detalhado na seção 2.3, ou seja, está bastante preocupada com o processo, suas diferentes tarefas e papéis que levam à sua correta execução. Esta visão trata os dados como um ingrediente para atingir o objetivo, muito importante, porém apenas um componente. Em contrapartida, em *data science* moderno os dados são o elemento principal, enquanto as metodologias de análise e geração do conhecimento deixam de ser prescritivas para algo mais inquisitivo, ou seja, o que deve ser feito, ao invés do que pode ser feito (Martínez-Plumed *et al.*, 2019).

Nesta perspectiva, a tratativa dos dados é vista mais como uma trajetória e menos um processo. Por exemplo, por vezes é possível gerar valor comercial em dados apenas realizando uma coleta e integração para consulta em um repositório do tipo *Online Analytical Processing* (OLAP), ou em uma aplicação muito comum atualmente, onde um aplicativo analisa as informações de localização dos usuários e recomenda rota com base em seus padrões. Essas perspectivas mostram que o produto é o próprio dado e o conhecimento extraído destes dados, muito diferente da abordagem processual focada em etapas que é útil para abordagens orientadas a objetivos (Martínez-Plumed *et al.*, 2019).

Por fim, é importante entender que a ciência evoluiu e os termos se referem a assuntos diferentes. Além disso, nem toda atividade de *data science* é exploratória ou independente de um processo organizado, por isso, o ciclo que será apresentado na seção 2.3 - CRISP-DM, continua válido e está contido em uma abordagem mais genérica, chamada de *Data Science Trajectories* (DST), conforme ilustrado na Figura 2. Esta ilustração ainda deixa implícito, porém proposital, o fato de não discriminar direções, ou seja, não há necessariamente uma ordem padrão para atividades de *data science*, onde pode-se inclusive intercalar atividades com foco em um determinado objetivo com atividades exploratórias (Martínez-Plumed *et al.*, 2019).

Figura 2 – Abordagem de trajetórias de *data science*



Fonte: Adaptado de Martínez-Plumed *et al.* (2019)

2.2 AS TAREFAS DE MINERAÇÃO DE DADOS

Neste capítulo, serão descritas as principais tarefas da mineração de dados, correlacionando com exemplos de negócios e com a descrição de alguns elementos técnicos das tarefas. Esta abordagem é válida porque há um grande número de algoritmos diferentes para aplicação de mineração de dados, entretanto, as tarefas ou objetivos da mineração essencialmente diferentes são apenas algumas, descritas a seguir.

2.2.1 Classificação e estimativa de probabilidade

Para Goldschmidt e Passos (2005) uma das tarefas de mineração de dados mais relevantes é a classificação, que consiste em modelar uma função a partir de dados de treinamento que permita classificar corretamente um sujeito desconhecido, ou seja, o modelo desenvolvido serve para prever a classe em que determinados registros se encontram.

Fawcett e Provost (2016) afirmam ainda que a tarefa de pontuação ou determinação da probabilidade de classe está intimamente ligada à classificação, ou seja, não apenas determinar a qual classe um novo indivíduo pertence, mas também de acordo com uma probabilidade. Por exemplo, ao ofertar um novo plano de telefonia a um cliente, qual a probabilidade de que o mesmo aceite esta oferta, sendo de conhecimento da operadora os atributos que descrevem esse cliente como a faixa etária, duração do contrato, utilização do serviço, renda entre outras características.

Boscarioli et. al (2016) descreve diversas técnicas adequadas para realizar a classificação, com detalhes de aplicações didáticas e práticas. Alguns métodos para determinar o modelo de classificação são a árvore de decisão, redes neurais, k-vizinhos mais próximos ou k-NN e Naive Bayes. A árvore de decisão é um algoritmo bastante simples que tem um bom desempenho, sendo uma das técnicas mais populares para a classificação.

2.2.2 Regressão

A tarefa de regressão busca ajustar uma função aos registros de um banco de dados, seja esta função linear ou não. Desta forma, esta tarefa é adequada para atributos numéricos. A regressão busca quantificar determinado atributo, por exemplo, a predição de risco de determinados investimentos ou a definição do valor de limite para uma linha de crédito de determinado cliente (Goldschmidt e Passos, 2005).

Para obter a função de regressão pode-se recorrer à métodos estatísticos, com base em condições e premissas das distribuições dos dados. Isto implica em fazer uma pré-

visualização dos dados em um gráfico de dispersão, por exemplo, buscando adequar o método de obtenção e a função mais adequada para ajustar o modelo. Estes modelos, além de linear ou não, podem ser simples ou multivariados, a depender da quantidade de atributos que são utilizados para previsão, onde as os modelos simples utilizam apenas um atributo (Boscarioli, Da Silva e Peres, 2016).

2.2.3 Combinação por similaridade

Esta tarefa consiste em agrupar dados que são semelhantes entre si, a afirmação parece bastante óbvia, porém alguns conceitos fundamentais muito importantes estão implícitos nesta frase. Quando a noção de similaridade é conhecida, existe um objetivo para a combinação, ou seja, a busca de uma característica de similaridade, esse fato leva à conclusão que este tipo de tarefa pode ser resolvida com técnicas supervisionadas de *data mining* (Fawcett e Provost, 2016).

A similaridade entre os dados, colocando-se matematicamente, pode ser definida como a distância entre estes dados, por exemplo: quando descrevemos dados com um vetor de característica, podemos estimar a distância euclidiana entre estes vetores, quanto menor esta distância, maior a similaridade entre os dados descritos (Goldschmidt e Passos, 2005).

Algumas situações que exemplificam essa tarefa podem ser, quando uma empresa quer encontrar em uma base de dados outras empresas que são similares a seus melhores clientes, ou seja, há um objetivo bastante claro do que é similar e um critério para a classificação destes elementos (Fawcett e Provost, 2016).

2.2.4 Agrupamento

A análise de agrupamento, de forma diferente da combinação por similaridade, é mais adequada para tarefas de exploração dos dados, como um processo que permite verificar relações entre os dados, ou seja, não há um rótulo pré-determinado. Assim, são adequadas as técnicas não supervisionadas de *data mining*, sendo o objetivo principal maximizar a similaridade intragrupo e minimizar a similaridade entre os grupos resultantes (Boscarioli, Da Silva e Peres, 2016).

Fawcett e Provost (2016) explicitam que o agrupamento pode ser utilizado para identificar grupos naturalmente diferentes. Assim, essa exploração dos dados permite entender melhor quem são os clientes, ou melhorar a forma de aplicação de campanhas de marketing, sendo uma abordagem vantajosa sempre que seja interessante entender quais são

os diferentes grupos para determinado negócio. Diferente da abordagem supervisionada, neste caso, os métodos não se concentram em uma variável alvo.

2.2.5 Agrupamento de coocorrência ou descoberta de associações

Para Goldschimdt e Passos (2005) esta tarefa busca itens que acontecem simultaneamente e de forma frequente em transações de bancos de dados. O exemplo mais clássico para esta tarefa é a aplicação de otimização e marketing com base nos registros de vendas em um supermercado, buscando itens que são frequentemente comprados juntos, com o propósito de otimizar layout e impulsionar o consumo de tais produtos.

Acontece que a associação é uma tarefa que depende do conhecimento do domínio e experiência do analista, afinal, quando há itens populares em comparação pode-se inferir erroneamente uma regra de associação. Para tanto, é necessário definir um parâmetro de suporte e outro de confiança ou força da associação, o primeiro diz respeito a representatividade da análise, ou seja, as regras devem se aplicar ao menos a 1% das transações. Já, a confiança representa um valor mínimo para que a frequência de ocorrência seja considerada uma associação, por exemplo, em 10% ou mais das vezes que um produto A é comprado, também ocorre para o produto B (Fawcett e Provost, 2016).

Para a descoberta de associações há diversos algoritmos específicos aplicáveis, pode-se citar: DHP (Direct Hashing and Pruning), Partition, DIC (Dynamic Itemset Counting), Apriori, Eclat, entre outros. Porém, estes algoritmos possuem alguma estrutura similar ao método Apriori (Goldschmidt e Passos, 2005).

2.2.6 Outras tarefas de mineração

É possível definir ainda algumas tarefas distintas, que de modo geral utilizam combinações e análises contidas nas outras técnicas, são elas:

- Perfilamento: consiste em entender o comportamento de um determinado grupo, população ou indivíduo, por exemplo, ao avaliar uma compra suspeita em um cartão de crédito com base no perfil conhecido do cliente;
- Previsão de vínculo: abordagem que busca indicar um vínculo ou ligação entre os dados de determinada aplicação, por exemplo em sistemas de recomendação de filmes, ainda não há uma ligação entre o usuário e o filme, porém, com base no que o usuário assistiu, deveria existir, sugerindo assim uma recomendação;

- Redução de dados: consiste em substituir um grande conjunto de dados, mantendo a maior parte possível de informações, buscando facilitar a análise ou processamento deste conjunto;
- Modelagem causal: tenta compreender a relação de causalidade, ou seja, quais ações realmente gerou influência em determinado grupo de pessoas. Esta abordagem é útil para diminuir as suposições em determinada análise, porém geram um dilema com relação ao aumento dos investimentos.

2.3 DATA SCIENCE ORIENTADO PARA NEGÓCIOS

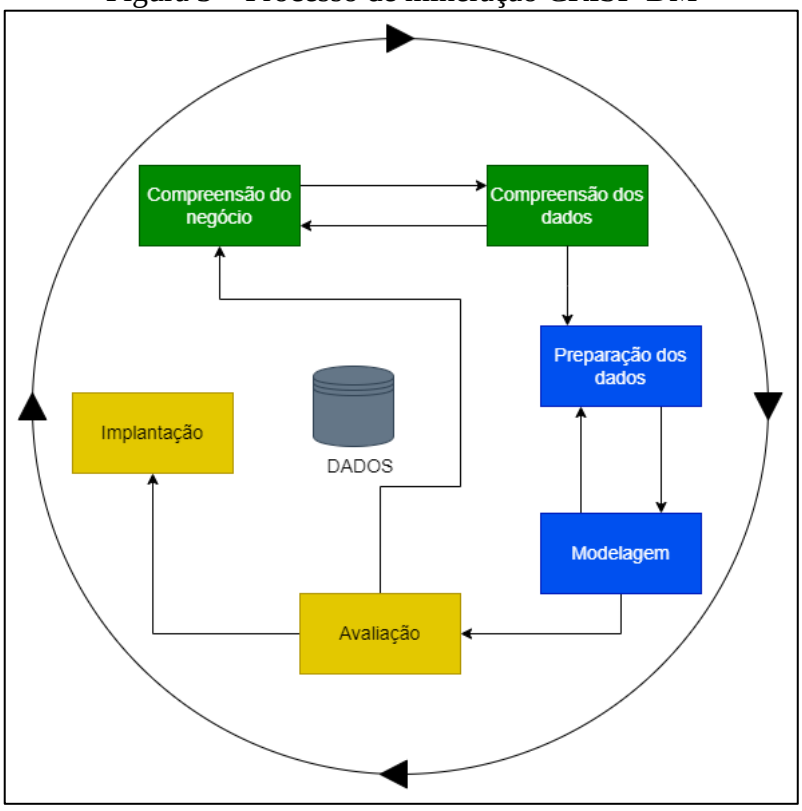
Segundo Fawcett e Provost (2016) o interesse pelo domínio do *data science* como ferramenta para extrair conhecimento e utilidade dos dados decorre da alta disponibilidade encontrada nos ambientes de negócios, tanto das informações estratégicas que afetam cada negócio, como as atividades de concorrentes, notícias e tendências, como os próprios dados gerados pelo negócio. Uma das aplicações mais evidentes do *data mining* atualmente está no marketing, por meio de análises em que é possível prever o comportamento de um cliente e obter o maior valor possível das negociações com este cliente.

Os termos envolvidos nas atividades de descoberta de conhecimento a partir de dados, mesmo que referindo-se às mesmas técnicas, têm semântica diferente. Para Fawcett e Provost (2016), o objetivo principal do *data science* é aprimorar a tomada de decisão, que é muito interessante para os negócios, funcionando como um suporte e também se sobrepondo a tomadas de decisão.

Entretanto KDD e *data mining* são ciências derivadas do aprendizado de máquina, que tem uma preocupação mais abrangente com todas as fases do processo de análise de dados: pré-processamento e coleta, modelos de aprendizagem e avaliação dos resultados. Uma empresa que tem uma equipe de *data science*, busca responder perguntas como: “Este novo cliente em particular será lucrativo? Quanto rendimento eu devo esperar que esse cliente gere?” (Fawcett e Provost, 2016).

Para atingir esse objetivo vários conceitos fundamentais são necessários. Uma premissa básica é encarar os dados como um ativo, que tem um custo para aquisição e podem gerar valor para um determinado negócio. Fawcett e Provost (2016) trazem o processo de mineração como um ciclo para os negócios, conforme a Figura 3 abaixo:

Figura 3 – Processo de mineração CRISP-DM



Fonte: Adaptado de Fawcett e Provost (2016)

Este processo denomina-se *Cross Industry Standard Process for data mining* (CRISP-DM) ou Processo Padrão de Indústria Cruzada para Exploração de Dados. A imagem deixa claro que o processo é iterativo, sendo que a repetição e característica cíclica fazem parte do refinamento da solução objetivo. Fawcett e Provost (2016) descrevem estas etapas como segue nos itens a seguir:

a) A compreensão do negócio busca traduzir um problema da área de negócios para o *data science*, analisando questões como o cenário de uso e o problema a ser resolvido. Nesta etapa as perguntas são: “O que queremos fazer? Como faríamos, exatamente? Quais partes do cenário de uso constituem possíveis modelos de mineração de dados?” (Schroer, Kruse e Gómez, 2021).

b) A compreensão dos dados tem relação direta com o entendimento do problema, de forma a perceber a limitação e as características fortes contidas nos dados. Geralmente os dados vem de bases de difícil relacionamento e exigem um grande trabalho para limpeza e correlação (Fawcett e Provost, 2016).

c) A preparação dos dados implica em converter dados para formato tabular, remover ou inferir valores ausentes, executar conversões de tipo e

normalizações para valores numéricos de acordo com a técnica específica de mineração de dados que será aplicada (Schroer, Kruse e Gómez, 2021).

d) A modelagem consiste em aplicar as técnicas de mineração de dados para obtenção de uma relação entre os dados e o atributo alvo. Por exemplo, ao aplicar a análise preditiva para classificação, obtém-se a árvore de decisão para predizer a classe de um novo sujeito com base em suas características (Fawcett e Provost, 2016).
e) A etapa de avaliação consiste em validar os resultados da mineração, garantir que o modelo satisfaça os objetivos do negócio, buscando resolver o problema. Nesta fase o modelo é testado de forma rigorosa para verificar sua aplicabilidade e eficiência, de forma qualitativa e quantitativa (Schroer, Kruse e Gómez, 2021).
f) Por fim a etapa de implantação consiste em colocar a solução desenvolvida em uso real, como parte de um sistema de informação ou processo de negócios. Por vezes, também é plausível a implantação da própria mineração de dados, em situações em que as condições são muito dinâmicas como em sistemas de detecção de fraude ou invasão ou ainda quando determinado negócio tem muitas tarefas de modelagem (Schroer, Kruse e Gómez, 2021).

É normal que o processo volte à etapa inicial, pois o processo como um todo produz maior conhecimento sobre o problema abordado, sendo que uma segunda iteração pode levar à produção de uma solução melhor. Além disso, podem ser revistas as metas e critérios de desempenho, além de gerar ideias para novos empreendimentos ou linhas de negócios (Fawcett e Provost, 2016).

2.4 *DATA SCIENCE* NAS PEQUENAS E MÉDIAS EMPRESAS (PMES)

Neste capítulo serão discutidos alguns autores que trazem as barreiras para adoção do *data science* nas PMEs, e estudos que apontam as melhorias de desempenho dos negócios que adotam processos de *data science*, com base na validação estatística de hipóteses.

2.4.1 As barreiras para adoção do *data science* por PMEs

No estudo de Willets *et al.* (2020) são descritas sessenta e nove barreiras para adoção do *data science* por PMEs. O autor utiliza *frameworks* teóricos para análise dessas barreiras e agrupamento em pilares de acordo com suas características. Os *frameworks* ou

metodologias teóricas utilizadas foram a *Technology-organization-environment* (TOE), *Technology-fit* (HOT-fit) e *Information Systems Strategy Triangle* (ISST).

Inicialmente as barreiras foram agrupadas em seis temas, sendo eles: dados, conhecimento e habilidades, regulatórios, recursos técnicos e organizacionais. Após a utilização das metodologias já citadas (TOE, HOT-fit e ISST), 7 barreiras levantadas foram descartadas e as demais foram organizadas em 5 pilares, mesclando algumas descrições de barreiras similares, conforme segue (Willetts, Atkins e Stanier, 2020):

- Pilar de negócios: agrupa as barreiras financeira e falta de casos de negócios. A primeira se relaciona ao investimento necessário em gerar, armazenar e processar os dados, além de outros custos como a segurança e treinamento. A falta de casos de negócios se refere à ausência de histórias de sucesso que documenta PMEs que adotaram *data science*;
- Pilar ambiental: agrupa as barreiras de preocupações éticas, falta de habilidade de identificação de riscos digitais, questões regulatórias e a falta de padrões comuns.
- Pilar humano: agrupa as barreiras de falta de experiência em análise de dados na empresa e falta de serviços de consultoria.
- Pilar organizacional: agrupa sete barreiras, o gerenciamento de mudanças, cultura, volume insuficiente de dados para análise, falta de habilidade e consciência gerencial, falta de suporte à alta direção, gerenciamento da tecnologia e dos talentos.
- Pilar tecnológico: agrupa as barreiras de complexidade dos dados, escalabilidade dos dados, silos de dados, prontidão da infraestrutura, falta de software adequado e qualidade ruim dos dados.

Rautenbach *et al.* (2022) faz uma revisão bibliográfica estruturada a respeito da implementação de *data science* em PMEs de países desenvolvidos e em desenvolvimento, com ênfase nos desafios e oportunidades associadas a este processo específico. A respeito das barreiras para implementação de *data science* em PMEs de países em desenvolvimento foram abordadas sete categorias, conforme segue abaixo:

- Qualidade dos dados: o autor menciona que poucas publicações relatam problemas relacionados à qualidade dos dados, visto que este fator é mais relevante quando os problemas de infraestrutura são superados;

- Infraestrutura insuficiente: neste item é ponderada a dificuldade de infraestrutura para utilizar *Big Data*, que acaba se tornando uma barreira para adoção do *data science;*
- Restrições financeiras: é vista como uma das principais barreiras para implantação de *data science*, pelo alto custo da infraestrutura adequada e pessoas adequadas;
- Preocupações com privacidade e segurança dos dados: por vezes os dados são armazenados em nuvem, podendo ser explorados caso não sejam armazenados com segurança, isso traz requisitos de habilidades jurídicas, que também se torna uma barreira;
- Desafios sociais: neste item a ênfase é a cultura organizacional, onde pode surgir a resistência a mudanças;
- Acesso ao software adequado: pode se tornar um problema sob a perspectiva que os custos das ferramentas são altos, além de requerer profissionais com conhecimento para gerar os resultados;
- Falta de habilidade: o autor menciona que este item é o mais frequente e barreira comum para adoção do *data science*. Por vezes os profissionais são muito requisitados e raros, sendo sua contratação inviável por PMEs.

Já para os países desenvolvidos são abordadas quatro categorias (Rautenbach, de Kock e Grobler, 2022):

- Desafios regulatórios: neste item é mencionada a falta de conhecimento sobre legislação específica, além dos requisitos de serviços, privacidade e segurança dos dados;
- Desafios econômicos: este item acaba impactando em todas as outras barreiras, pois sempre existem restrições orçamentárias que pesam mais em alguns itens, por exemplo, infraestrutura e profissionais qualificados;
- Desafios técnicos: existem alguns desafios relacionados à aquisição, armazenamento e analise de grandes volumes de dados, assim como a precisão e rastreabilidade de processos decisórios com uso de Inteligência Artificial;
- Desafios organizacionais: geralmente a alta direção adota *data science* como projetos temporários, ao invés da transformação permanente do processo de decisão baseado em dados, essa abordagem leva a expectativas não atendidas e a descontinuidade dos processos;

- Falta de habilidade: assim como para os países em desenvolvimento a escassez de mão de obra qualificada também é um problema nos países desenvolvidos.

Tawil *et al.* (2023) em seu trabalho traz a análise de 85 PMEs que passaram pelo processo de digitalização e adoção de processos decisórios baseados em dados, dos principais desafios e lições aprendidas pode-se citar os seguintes pontos:

- Falta de dados específicos de cada setor;
- Pouco conhecimento a respeito do valor de dados abertos ou fechados e seu potencial;
- Integração dos dados aos sistemas das e PMEs;
- Recursos financeiros limitados;
- Percepção de que é mais fácil continuar os negócios de forma tradicional;
- Falta de analistas de dados especialistas em determinados domínios;
- Falta ou conhecimento desatualizado;
- Conhecimento insuficiente sobre as ferramentas para análise de dados e as fontes de financiamento disponíveis;
- Limitação de capacidade em *data science* e *machine learning;*

2.4.2 Resultados na implantação de *data science* em PMEs

De acordo com Bhardwaj (2022) quando são relacionados os assuntos *data science* para PMEs, há quatro temas principais que surgem em publicações e podem ser identificados: os fatores que estimulam a adoção, as barreiras ou forças restritivas, que tipos de empresas estão investindo em *data science* e os indicadores de performance.

Os tipos de empresas que mais investem em tecnologias de dados são as empresas de manufatura/produção, software, logística e marketing. Além disso, é possível notar que alguns países tem maior foco de desenvolvimento das PMEs como a Coreia, África do Sul e Tailândia, além dos países com forte orientação à inovação como Estados Unidos e Japão. O sucesso da adoção de novas tecnologias depende fortemente das condições de suporte organizacional e corporativo, das questões comportamentais e da maturidade da empresa para entender a tecnologia. Na maioria dos casos as tecnologias são aplicadas em setores da cadeia de valor como compras, logística, marketing, vendas e operações (Bhardwaj, 2022).

Para as empresas que aplicam com sucesso as tecnologias de análise de dados, é possível notar maior desempenho e lucratividade. Um dos indicadores é o financeiro, quando o *data science* é aplicado neste processo as empresas notam ganhos de produtividade e

maximização de receita. Além disso, quando aplicada ao processo contábil resulta em maior eficiência e acuraria do processo. Outro indicador é o desempenho de mercado, em que normalmente as empresas investem em publicidade de diversas formas, e as tecnologias de *data science* têm ajudado as PMEs a perceber o mercado, entender o costume do consumidor e manter a reputação. Assim, decisões baseadas em dados têm mostrado impacto significativo em vendas, visibilidade, rotatividade de clientes entre outras medidas de performance de mercado (Bhardwaj, 2022).

Por último há o indicador de desempenho organizacional. Conforme citado anteriormente, quando *data science* é utilizado em processos funcionais e operacionais cruciais para a organização, ajuda a desenvolver as capacidades internas de um negócio. O desenvolvimento de habilidades, práticas de compartilhamento de conhecimento, uso avançado da tecnologia, engajamento da inovação em processos internos e externos são alguns efeitos que as empresas testam empiricamente e o uso colaborativo das tecnologias ajudam na integração e simplificação das organizações (Bhardwaj, 2022).

Em termos numéricos Medeiros et. al (2020) afirma que as organizações que utilizam tecnologias de *big data* e *data science* podem alcançar 5 a 6% maior produtividade. Em uma outra análise de 814 empresas durante os anos de 2008 a 2014 mostrou que os ganhos são de 3 a 7%. Um terceiro trabalho que compara o crescimento anual possibilitou notar que as empresas que utilizam *data science* cresceram 7,2% no ano seguinte à adoção das tecnologias em comparação de 3% com o setor.

No mesmo trabalho, são apresentados resultados de uma pesquisa realizada com 211 pessoas onde vários benefícios inerentes ao processo de adoção de práticas de *data science* para negócios são relatados e categorizados. Um total de 46 pessoas forneceram respostas relacionadas à qualidade de dados, como: os processos ajudam a criar uma cultura orientada a dados, reduzem incertezas e ampliam a visão da gestão da empresa. Outras 59 respostas foram relacionadas à inteligência analítica, onde os benefícios são: as tecnologias fornecem análise, avaliação, predição e suporte à melhoria, além de acelerar o processo de geração de insights (de Medeiros, Maçada e Hoppen, 2020).

Outras duas dimensões categorizadas pelo estudo foram a capacidade dinâmica e as vantagens competitivas. No primeiro caso, 34 pessoas afirmaram que *data science* facilita a integração e comunicação, geração do conhecimento e ajuda a entender os ambientes de negócio e percepção de oportunidades. Com relação às vantagens competitivas, 20 pessoas responderam que as tecnologias ajudam a identificar e priorizar mudanças estratégicas, além

de facilitar o gerenciamento do desempenho da organização e consequentemente aumentam a competitividade do negócio (de Medeiros, Maçada e Hoppen, 2020).

3 METODOLOGIA

A pesquisa foi conduzida por meio de uma abordagem bibliográfica exploratória com o objetivo de investigar o impacto das ferramentas de *data science* na tomada de decisões estratégicas em pequenas e médias empresas (PMEs), focando na melhoria dos processos empresariais e na superação das barreiras comuns enfrentadas por essas empresas. A metodologia escolhida permite o levantamento de material teórico e estudos de caso que ofereçam suporte aos objetivos específicos definidos, bem como uma compreensão abrangente do tema em questão.

Para o desenvolvimento da pesquisa, foram utilizadas as seguintes fontes e materiais:

- Artigos e publicações científicas: foram consultadas bases de dados acadêmicas como Google Acadêmico, Elsevier, IEEE Xplore e SpringerLink. A busca foi direcionada para identificar trabalhos que discutam tanto os fundamentos teóricos quanto as aplicações práticas das técnicas de *data science* em ambientes empresariais;
- Relatórios e dados estatísticos: na contextualização do problema foram utilizados dados econômicos e setoriais obtidos de fontes como o Instituto Brasileiro de Geografia e Estatística (IBGE), Associação Brasileira de Empresas de Software (ABES), e publicações do Serviço Brasileiro de Apoio às Micro e Pequenas Empresas (SEBRAE). Esses documentos forneceram informações sobre a situação atual das PMEs no Brasil e foi possível inferir as tendências de crescimento do setor de TI;
- Livros e referências clássicas: na elaboração da fundamentação teórica, foram consultados livros e artigos clássicos sobre KDD, como o trabalho seminal de Fayyad *et al.* (1996) e estudos mais recentes como os de Coenen (2011) e Martínez-Plumed *et al.* (2019), que discutem a evolução da ciência de dados e os desafios atuais. Além disso, autores brasileiros contemporâneos que tratam da aplicação prática de técnicas de KDD utilizando ferramentas *open source* serão incluídos na revisão.

As principais palavras-chave que orientaram a busca incluem:

- *Data science* para negócios (*data science for business);*
- *Data science* para pequenas e médias empresas;
- Descoberta de conhecimento em banco de dados (KDD);
- Mineração de dados (*data mining);*
- Ciclo CRISP-DM;
- Barreiras na adoção de *data science.*

As buscas foram realizadas de março até agosto de 2024, utilizando sempre as fontes mais recentes disponíveis.

A partir da pesquisa foi desenvolvida a fundamentação teórica, buscando elucidar os conceitos fundamentais de *data science, KDD,* e *data mining*, além de uma análise detalhada sobre a relação entre a adoção dessas tecnologias e a melhoria dos processos de negócios em PMEs.

A pesquisa foi predominantemente qualitativa, uma vez que se destina a compreender os aspectos subjetivos e contextuais das práticas de *data science* em PMEs. Prodanov e Freitas (2013) destacam que a pesquisa qualitativa não se baseia na aplicação de técnicas estatísticas rigorosas, mas sim em uma análise aprofundada de dados textuais e contextuais. A pesquisa exploratória, conforme descrito por Gil (2008) e Selltiz *et al.* (1974), foi utilizada para desenvolver e esclarecer conceitos, fornecer uma visão ampla do tema e auxiliar na identificação de prioridades para futuras investigações.

Os resultados obtidos por meio da revisão bibliográfica e análise de trabalhos acadêmicos foram discutidos em comparação com o referencial teórico estabelecido. A metodologia qualitativa permitiu que as discussões sejam aprofundadas, examinando como as PMEs podem superar as barreiras identificadas e se beneficiar das tecnologias de *data science* para melhorar seus processos de negócio.

4 ANÁLISE E DISCUSSÕES

O processo de KDD, conforme descrito por Fayyad *et al.* (1996), é um dos pilares para a compreensão sobre como as informações podem ser extraídas de grandes volumes de dados. O KDD é um processo interativo e iterativo que envolve várias etapas, desde a seleção de dados até a interpretação dos resultados. Ele é amplamente confundido com o *data mining*, que é apenas uma das etapas do KDD. Coenen (2011) reforça essa diferenciação, apontando que o *data mining* é responsável pela identificação de padrões nos dados,

enquanto o KDD abrange um escopo mais amplo, incluindo a preparação dos dados e a avaliação dos resultados.

Na seção 2.2 foram apresentadas as técnicas mais comuns para mineração de dados, sendo a sua compreensão importante de acordo com o objetivo da trajetória de *data science* adotada dentro de um negócio. Fawcett e Provost (2016) demonstram exemplos da utilização de classificação ou regressão para realizar previsões com base em dados existentes, sendo uma das tarefas mais elementares e que podem beneficiar tomadas de decisão e a competitividade de empresas. Outra tarefa muito relevante é compreender se existem grupos naturais nos conjuntos de dados, que auxilia na estratégia de negócio para cada agrupamento, contribuindo também para a eficiência dos processos.

A relevância dessas abordagens para o cenário dos negócios está no fato de que, à medida que as empresas acumulam grandes quantidades de dados, a capacidade de extrair informações úteis se torna mais difícil e crucial. O uso do *data mining* possibilita identificar padrões ocultos e tendências que podem apoiar decisões estratégicas. Para as PMEs, essa capacidade pode significar a diferença entre competir de maneira eficiente em um mercado dinâmico ou ficar para trás. No entanto, como apontado por Saeed (2020), as PMEs enfrentam barreiras significativas, como restrições financeiras e falta de conhecimento especializado, o que dificulta a adoção dessas tecnologias, especialmente para empresas mais tradicionais ou que não são de setores competitivos ou tecnológicos como manufatura/produção, software, logística e marketing, conforme descreve Bhardwaj (2022).

Na seção 2.1 o conceito de d*ata science* foi apresentado como uma evolução das técnicas de KDD e *data mining*, com uma abordagem mais ampla e orientada a dados. Martínez-Plumed *et al.* (2019) discutem como a natureza exploratória do *data science* permite que as empresas aproveitem ao máximo seus dados, não apenas seguindo processos estabelecidos com objetivos específicos, mas também questionando o que pode ser descoberto por meio da análise de dados, ou mesmo, trazendo ideias e aplicações onde o produto em si são os dados. Para as PMEs, essa visão mais flexível pode ser um diferencial, pois elas podem adotar soluções menos prescritivas ou modelos rígidos, fazendo com que as análises e benefícios possam ser melhor ajustadas às suas realidades.

No entanto, um ponto crítico destacado é a necessidade de infraestrutura e competências para que as empresas possam implementar soluções de *data science* de forma eficiente. Como descrito por Rautenbach *et al.* (2022), as PMEs em economias emergentes, como o Brasil, enfrentam desafios adicionais, incluindo qualidade insuficiente dos dados,

falta de infraestrutura e principalmente escassez de profissionais qualificados. Esses problemas agravam-se quando se considera a alta taxa de mortalidade empresarial em setores onde a análise de dados poderia contribuir significativamente para a sobrevivência, como no caso das micro e pequenas empresas (SEBRAE, 2023).

O estudo de Willetts *et al.* (2020) corrobora essa ideia, pois identifica uma série de barreiras organizacionais, tecnológicas e financeiras para a adoção de processos de *data science* por PMEs. Essas barreiras incluem, principalmente, a falta de recursos para investimentos em tecnologia e a inexperiência com análise de dados. Para superar esses desafios, é essencial que as empresas menores consigam visualizar o retorno sobre o investimento em *data science*, algo que é possível notar pela revisão sistemática elaborada por Bhardwaj (2022). O autor relata que aplicações em *data analytics* podem proporcionar inovação, produtividade e maximização de receita conforme os trabalhos de Saleem *et al.* (2020) e Rajani & Kumar (2017). Outros três trabalhos relatam os impactos no desempenho financeiro das PMEs que se beneficiam do *data science* em Maroufkhani (2020), Henage (2020) e Ferarris (2019).

A aplicabilidade de *data science* em PMEs é diretamente relacionada à sua capacidade de melhorar a eficiência operacional e a tomada de decisões estratégicas. A análise de grandes volumes de dados oferece a possibilidade de uma visão mais clara sobre as tendências de mercado, comportamento dos consumidores e até mesmo a identificação de novas oportunidades de negócios. No entanto, como salientado por Saeed (2020), o sucesso na implementação dessas tecnologias depende de recursos especializados, os quais muitas PMEs não possuem.

Um dos principais recursos são os cientistas de dados, que são escassos e têm custo elevado. Willetts *et al.* (2020) afirma que geralmente as PMEs não contratam analistas de dados, ou ainda, não tem profissionais com as habilidades e percepção necessárias para implementar processos de *data science.* Outro ponto importante, é que faltam serviços de consultoria, pois geralmente empresas que prestam esse tipo de serviço se envolvem em projetos de dados complexos e que consomem muito tempo para outras empresas de maior porte.

Outra questão convergente entre os autores foi o desafio organizacional quando se trata da implementação de novas tecnologias de *data science* nos negócios. Rautenbach *et al.* (2022) relata que a resistência a mudança e dificuldade de adaptação é um grande impeditivo para muitas PMEs, sendo de grande importância o apoio gerencial. A falta de

profissionais com conhecimento em análise de dados em um nível executivo maior, leva à falta de liderança estratégica, sendo uma das causas apontadas para a baixa adoção em países em desenvolvimento. Neste mesmo sentido Willetts *et al.* (2020) afirma que as PMEs raramente se interessam por tendências de gerenciamento, podendo encarar a tecnologia como um modismo passageiro e muitas ainda tem dificuldade em se afastar da tomada de decisão com base em palpites ou instinto.

5 CONSIDERAÇÕES FINAIS

A integração de *data science* com processos de negócios é uma oportunidade inexplorada para muitas PMEs. Apesar das barreiras, a adoção dessas tecnologias pode fornecer vantagens competitivas, como a redução de custos e o aumento da flexibilidade nas operações empresariais. Essas melhorias podem ser vistas na automação de processos, na previsão de demanda e até mesmo na melhoria da satisfação do cliente.

A abordagem moderna do *data science* é ampla, sendo que frações dos processos podem ser implementadas nas empresas de forma que gerem benefício em suas operações e resultados. As tarefas clássicas tem foco em mapear tendências, classificar ou prever comportamentos dos clientes, porém, algumas trajetórias trazem o dado como um produto por si só, como é possível notar em um aplicativo de mapas e geolocalização.

Essa discussão foi explorada conforme a proposta inicial de trazer os conceitos dos processos relacionado a dados, principalmente no que diz respeito à KDD, *data mining* e *data science*, com a abordagem das definições contidas nos trabalhos de referência que foram precursores na área. Além disso, também foram descritos os ciclos de aplicação mais conhecidos, sendo um deles o CRISP-DM e outro mais recente o DST correlacionando as atividades de *data science* com os objetivos e o ambiente empresarial nas seções 2.3 e 2.1 respectivamente.

Pelo desenvolvimento deste trabalho foi possível notar que existem barreiras significativas para a adoção das práticas de análise e exploração de dados pelas PMEs. Trabalhos relacionados a este tema foram citados, classificando ou categorizando restrições e dificuldades das empresas em se beneficiar das tecnologias conforme descrito na seção 2.4.1. Algumas barreiras mais frequentes nos trabalhos são a falta de recursos financeiros, falta de habilidades e conhecimento dos profissionais das empresas, falta de estratégia da liderança para implantação dos processos e dificuldade com a cultura organizacional.

Na seção 2.4.2 foram apresentados trabalhos que indicam uma correlação entre a adoção de tecnologias para tomada de decisão e um aumento do desempenho das empresas, em indicadores diferentes como participação de mercado ou faturamento.

Toda a fundamentação foi importante para gerar a discussão proposta, resolvendo os objetivos de elucidar os conceitos, apresentar as barreiras para adoção do *data science* por PMEs, além de relatar os resultados documentados da vantagem que as empresas obtêm ao adotar tais práticas. Desta forma, considera-se cumprido o objetivo geral do trabalho, que se propõe a investigar o impacto e os desafios na adoção do *data science* em pequenas e médias empresas.

Para trabalhos futuros sugere-se o tema do impacto da automação inteligente de processos empresariais, dado que é uma tecnologia em ascensão, apontada como uma tendência de mercado pela pesquisa da ABES (2023), demonstrando que as empresas estão dispostas a cada vez mais delegar atividades para os algoritmos de inteligência artificial.

REFERÊNCIAS

ASSOCIAÇÃO BRASILEIRA DAS EMPRESAS DE SOFTWARE. **Mercado Brasileiro de Software:** panorama e tendências. 1ª. ed. São Paulo: ABES, 2023. Disponivel em: <https://abes.com.br/dados-do-setor/>.

BHARDWAJ, S. Data analytics in Small and Medium Enterprises (SME): A systematic review and future research directions. **Information resources management journal**, Sambalpur, 2022.

BOSCARIOLI, C.; DA SILVA, L. A.; PERES, S. M. **Inrodução à mineração de dados:** com aplicações em R. 1ª. ed. Rio de Janeiro: Elsevier, 2016.

COENEN, F. Data mining: Past, present and future. **The Knowledge Engineering Review**, Liverpool, março 2011.

COLEMAN, S. Y. Data-Mining Opportunities for Small and Medium Enterprises with Official Statistics in the UK. **Journal of Official Statistics**, v. 32, n. No. 4, p. 849-865, 2016. ISSN https://doi.org/10.1515/jos-2016-0044.

DE MEDEIROS, M. M.; MAÇADA, A. C. G.; HOPPEN, N. Data science for business: benefits, challenges and opportunities. **The bottom line**, Porto Alegre, 2020.

FAWCETT, T.; PROVOST, F. **Data Science para Negócios:** O que você precisa saber sobre mineração de dados e pensamento analítico de dados. 1ª. ed. Rio de Janeiro: Alta Books, 2016.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; PADHRAIC, S. From data mining to knowledge discovery in databases. **AI Magazine**, v. 17, n. 3, 1996. ISSN DOI: 10.1609/aimag.v17i3.1230.

GIL, A. C. **Métodos e técnicas de pesqusa social**. 6. ed. São Paulo: Atlas, 2008.

GOLDSCHMIDT, R.; PASSOS, E. **Data Mining:** um guia prático. 1ª. ed. Rio de Janeiro: Elsevier, 2005.

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA. **Indicadores IBGE:** Contas Nacionais Trimestrais. 1ª. ed. Rio de Janeiro: IBGE, 2023. Disponivel em: <https://biblioteca.ibge.gov.br/visualizacao/periodicos/2121/cnt_2022_4tri.pdf>.

LIMA, A.; COLAÇO JUNIOR, M.; NASCIMENTO, A. V. R. P. **Um survey com empresas brasileiras acerca da utilização de business intelligence (BI) e um diagnóstico sobe a infraestrutura e metodologias associadas**. Conferência Ibero-Americana de Engenharia de Software - Trilha de Engenharia de Software Experimental. Buenos Aires: [s.n.]. 2017. p. 15.

MARQUESONE, R. **Big Data:** técnicas e tecnologias para extração de valor dos dados. 1ª. ed. São Paulo: Casa do Código, 2018.

MARTÍNEZ-PLUMED, F. et al. CRISP-DM Twenty Years Later: From. **IEEE Transactions on Knowledge and**, Valência, 2019. 3048-3061.

MINISTÉRIO DA ECONOMIA. Brasil registrou abertura de 3.838.063 novas empresas em 2022. **gov.br**, 2023. Disponivel em: <https://www.gov.br/economia/pt-br/assuntos/noticias/2023/janeiro/brasil-registrou-abertura-de-3-838-063-novas-empresas-em-2022>. Acesso em: 31 março 2024.

PRODANOV, C. C.; FREITAS, E. C. **Métodos e técnicas da pesquisa e do trabalho acadêmico**. 2. ed. Rio Grande do Sul: Universidade Feevale, 2013.

RAUTENBACH, S.; DE KOCK, I.; GROBLER, J. Data science for small and medium-sized enterprises: a structured literature review. **The South African Journal of Industrial Engineering**, Stellenbosch, novembro 2022.

SAEED, T. Data Mining for Small and Medium Enterprises: A Conceptual Model for Adaptation. **Intelligent Information Management**, Madinah, vol. 12, n. No. 5, setembro 2020. 183-197. Disponivel em: <https://www.scirp.org/journal/paperinformation?paperid=102830>.

SCHROER, C.; KRUSE, F.; GÓMEZ, J. M. A Systematic Literature Reviw on Applying CRISP-DM Process Model. **Procedia Computer Science**, Oldenburg, 2021. 526-534.

SEBRAE. A taxa de sobrevivência das empresas no Brasil. **SEBRAE**, 2023. Disponivel em: <https://sebrae.com.br/sites/PortalSebrae/artigos/a-taxa-de-sobrevivencia-das-empresas-no-brasil,d5147a3a415f5810VgnVCM1000001b00320aRCRD>. Acesso em: 31 março 2024.

SEBRAE. Pequenos negócios: a base da economia do nosso país. **SEBRAE**, 2023. Disponivel em: <https://sebrae.com.br/sites/PortalSebrae/artigos/pequenos-negocios-a-base-da-economia-do-nosso-pais,85e97325a3937810VgnVCM1000001b00320aRCRD>. Acesso em: 31 março 2024.

SELLTIZ, C.; WRIGHTSMAN, L. S.; COOK, S. W. **Métodos de pesquisa nas relações sociais**. São Paulo: Editora Pedagogica e Universitária, 1974.

TAWIL, A.-R. H. et al. **Trends and challenges towards an effective data-driven decision making in UK SMEs: case studies and lessons learnt from the analysis of 85 SMEs**. Birmingham Cìty University. Birmingham. 2023. (https://doi.org/10.48550/arXiv.2305.15454).

WILLETTS, M.; ATKINS, A. S.; STANIER, C. Barriers to SMEs adoption of bid data analytics for competitive advantage. **2020 Fourth International Conference On Intelligent Computing in Data Sciences (ICDS)**, Fez, 30 novembro 2020.