Ciência de Dados para Micro e Pequenas Empresas

Sarah Keith Pereira de Oliveira, Wesley Duarte Pereira Bertipaglia, José Alexandre Ducatti*
* Orientador jose.ducati@fatec.sp.gov.br

Faculdade de Tecnologia, FATEC de S. J. do Rio Preto/SP

RESUMO

No cenário empresarial em constante evolução, as micro e pequenas empresas (MPEs) enfrentam desafios significativos para se manterem competitivas. No entanto, a ciência de dados emergiu como uma ferramenta poderosa para auxiliar essas empresas a enfrentar esses obstáculos e prosperar em um ambiente cada vez mais complexo. Este artigo explora como a ciência de dados pode ser aplicada de maneira estratégica pelas MPEs, capacitando-as a tomar decisões embasadas em dados, identificar oportunidades de mercado, aprimorar seus processos internos, otimizar recursos e impulsionar o crescimento sustentável e a eficiência operacional.

Palavras-chave: micro e pequenas empresas, ciência de dados, eficiência operacional.

ABSTRACT

In the constantly evolving business landscape, micro and small companies (MSE) face significant challenges to remain competitive. However, data science has emerged as a powerful tool to help these companies overcome these obstacles and thrive in an increasingly complex environment. This article explores how data science can be strategically applied by MSEs, enabling them to make data-driven decisions, identify market opportunities, improve their internal processes, optimize resources and drive sustainable growth and operational efficiency.

Keywords: micro and small businesses, data science, operational efficiency.

1. INTRODUÇÃO

A ciência de dados é uma ferramenta poderosa para a transformação empresarial em organizações de todos os tamanhos, impulsionando a inovação e o crescimento. Segundo a IBM (S/D) a ciência de dados combina matemática e estatística, programação especializada, análise avançada, inteligência artificial (IA) e *machine learning* com conhecimento em assuntos específicos para descobrir *insights* práticos, ocultos nos dados de uma organização.

Esses insights podem oferecer às MPEs uma vantagem competitiva significativa, pois permite que elas tomem decisões informadas e estratégicas. Isso pode ajudá-las a otimizar processos, identificar oportunidades de crescimento, compreender melhor seus clientes e concorrentes, e superar os obstáculos que podem surgir em seu caminho para o sucesso.

No entanto, as MPEs enfrentam desafios únicos ao incorporar a ciência de dados em suas operações, incluindo limitações orçamentárias e a falta de capacitação em gestão e tecnologia. Superar essas barreiras é crucial para que as MPEs alcancem todo o potencial da ciência de dados e se destaquem em seus mercados competitivos.

Neste artigo, exploraremos como a ciência de dados pode se tornar uma importante ferramenta para as MPEs, capacitando-as a superar os desafios que muitas vezes as impedem de atingir seu pleno potencial. Analisaremos casos de uso específicos, destacando como a

coleta, análise e interpretação de dados podem se traduzir em vantagens competitivas tangíveis. Além disso, discutiremos as ferramentas e estratégias-chave que as MPEs podem adotar para iniciar ou aprimorar suas iniciativas de ciência de dados.

2. OBJETIVO

O objetivo deste artigo é proporcionar uma visão abrangente de como a ciência de dados pode ser aplicada com sucesso nas Micro e Pequenas Empresas (MPEs) para superar seus desafios. Serão explorados como a ciência de dados pode ajudar as MPEs a tomar decisões mais informadas, melhorar a eficiência operacional, aprimorar a compreensão do mercado e do cliente, e impulsionar o crescimento e a competitividade.

3. METODOLOGIA

A metodologia empregada neste artigo baseia-se em uma abordagem de pesquisa qualitativa e quantitativa, que incluiu a revisão de literatura e exemplos ilustrativos baseados em uma base de dados simbólica para demonstrar como as técnicas de ciência de dados podem ser aplicadas em cenários empresariais específicos.

4. TRABALHOS SIMILARES

Na quarta edição do livro "Business Intelligence, Analytics, and Data Science" de Sharda et al (2015), os autores exploram o campo da inteligência empresarial e ciência de dados como ferramentas estratégicas cruciais para organizações. Outro trabalho que compartilha semelhanças em seus objetivos e objeto de estudo é o trabalho de Brietzig et al. (2022), os autores discutem a análise de dados como uma ferramenta para tomada de decisões que pode ser usada pelas MPEs.

5. FUNDAMENTAÇÃO TEÓRICA

Antes de explorar os benefícios da ciência de dados para micro e pequenas empresas, é crucial estabelecer algumas fundamentações teóricas que tangenciam os princípios e técnicas subjacentes à sua aplicação estratégica no contexto empresarial.

5.1 MICRO E PEQUENAS EMPRESAS (MPEs)

De acordo com Dornelas (2005) as Micros e Pequenas Empresas (MPEs) são empresas que possuem características específicas, tais como: menor número de funcionários, menor faturamento, menor poder de barganha e menor acesso a recursos financeiros em relação às grandes empresas.

Dornelas destaca que essas características podem ser transformadas em vantagens competitivas, uma vez que as MPEs são mais flexíveis e têm maior capacidade de inovação em comparação às grandes empresas.

Para o Serviço Brasileiro de Apoio às Micro e Pequenas Empresas (SEBRAE) (2013), às micro e pequenas empresas podem ser divididas em 3 categorias, seguindo os critérios da Lei Complementar 123/2006, também chamada de Lei Geral das Micro e Pequenas Empresas. Resumidamente, as MPEs são divididas da seguinte maneira:

- Microempreendedor Individual - Faturamento anual até R$ 81 mil;
- Microempresa - Faturamento anual até R$ 360 mil;
- Empresa de Pequeno Porte - Faturamento anual entre R$ 360 mil e R$ 4,8 milhões;

Os principais desafios para micro e pequenas empresas se manterem no mercado estão associados à falta de planejamento e pouca capacitação em gestão, ao excesso de burocracia para a obtenção de crédito e à alta carga tributária (Portal da Indústria, S/D).

Ainda segundo o Portal da Indústria (S/D), com a economia desaquecida e o custo elevado do crédito, as micro e pequenas empresas estão enfrentando maiores desafios e precisam de um tratamento especial para sobreviverem no País.

5.2 CIÊNCIA DE DADOS

Segundo a IBM (S/D) o volume crescente de fontes de dados e, subsequentemente, dos dados tornou a ciência de dados um dos campos que mais crescem em todos os setores. As organizações dependem cada vez mais dos cientistas de dados para interpretar dados e fornecer recomendações acionáveis para melhorar os resultados de negócios.

As informações obtidas com o data science, na maioria dos casos, são utilizadas para a tomada de decisões importantes, como a criação de novos produtos ou serviços, atualização de produtos, mudanças nos negócios e, até mesmo, qual será o futuro de uma organização (Oliveira, 2023).

A ciência de dados é impulsionada por um conjunto de tecnologias essenciais, incluindo linguagens de programação, bancos de dados, ferramentas de visualização e bibliotecas de machine learning. A escolha das tecnologias depende das necessidades do projeto.

5.3 PYTHON E BIBLIOTECAS

Python é uma linguagem de programação interpretada, orientada a objetos e de alto nível com semântica dinâmica. Suas estruturas de dados integradas de alto nível, combinadas com digitação dinâmica e ligação dinâmica, tornam-no muito atraente para o desenvolvimento rápido de aplicativos, bem como para uso como script ou linguagem adesiva para conectar componentes existentes (Python, S/D). Além do python foram usadas algumas bibliotecas como o Pandas para análise e manipulação de dados, o Matplotlib para criar visualizações e o Scikit-learn para análise preditiva e prescritiva de dados.

5.4 SQLite

SQLite é uma biblioteca em linguagem C que implementa um mecanismo de banco de dados SQL pequeno, rápido, independente, de alta confiabilidade e completo. SQLite é o mecanismo de banco de dados mais usado no mundo. O SQLite está integrado em todos os telefones celulares e na maioria dos computadores e vem integrado em inúmeros outros aplicativos que as pessoas usam todos os dias (SQLite, S/D).

6. CIÊNCIA DE DADOS PARA MPEs

O ambiente de negócios está em constante mudança e está se tornando cada vez mais complexo. As organizações, tanto privadas como públicas, são conservadoras na forma como operam. Tais atividades exigem que as organizações sejam ágeis e tomem decisões estratégicas, táticas e operacionais frequentes e rápidas, algumas das quais muito complexas. Tomar tais decisões pode exigir quantidades consideráveis de dados, informações e conhecimentos relevantes. O processamento destes, no âmbito das decisões necessárias, deve ser feito rapidamente, frequentemente em tempo real, e geralmente requer algum suporte informatizado (Sharda et al, 2015).

Neste âmbito, a ciência de dados representa uma importante ferramenta para impulsionar o crescimento e a eficiência dos negócios, em especial para as MPEs, ao explorar

dados internos e externos, as MPEs podem obter insights valiosos para aprimorar a tomada de decisões, identificar oportunidades de mercado e otimizar processos internos. Além disso, ela pode ajudar as MPEs a reconhecer padrões e lacunas em seus negócios, removendo gargalos no fluxo dos processos e simplificando o gerenciamento com uma visão clara e precisa.

6.1 COLETA DE DADOS

A coleta e análise de dados constituem os pilares da transformação de informações em insights valiosos para as Micro e Pequenas Empresas (MPEs). Os dados são os registros de informações que são produzidas a todo momento, seja por pessoas ou por máquinas e que são dotados de potencial para gerar insights de solução de problemas e inovação (Oliveira, 2023).

Os gestores devem direcionar seus esforços para coletar dados em áreas onde os insights podem ter o maior impacto positivo, isso pode ser feito a partir da identificação dos indicadores-chave de desempenho ou KPIs (*Key Performance Indicators*).

Uma vez que os KPIs foram estabelecidos, as MPEs podem então começar a coletar dados relevantes de fontes apropriadas. Isso pode envolver o uso de planilhas, a implementação de sistemas de registros, bancos de dados, a integração de plataformas de gestão e CRM (*Customer Relationship Management*), a configuração de ferramentas de análise de site para monitorar o tráfego da web e muito mais.

É importante garantir que a coleta de dados seja consistente, confiável e alinhada com os KPIs definidos, pois isso garantirá que as informações coletadas sejam valiosas e utilizáveis na análise subsequente, além disso é importante garantir a disponibilidade dos dados com armazenamento e *backup* feitos em armazenamentos confiáveis, em uma máquina local ou em serviços em nuvem.

6.2 PROCESSAMENTO DE DADOS

Segundo Kotsiantis et al, (2006) os dados em sua forma original geralmente não estão prontos para serem usados em tarefas analíticas. Muitas vezes são sujos, desalinhados, excessivamente complexos e imprecisos. O processamento é necessário para converter os dados brutos em um formato refinado para algoritmos analíticos. Este processo é fundamental para garantir a qualidade e a utilidade dos dados para a construção e avaliação de modelos analíticos.

6.2.1 TRANSFORMAÇÃO DOS DADOS

A transformação de dados na ciência de dados refere-se ao processo de conversão de dados brutos em um formato adequado para análise e modelagem. O objetivo da transformação de dados é prepará-los para mineração de dados, para que possam ser usados para extrair insights e conhecimentos úteis (Geeks for Geeks, 2023).

Segundo Melo (2022), normalmente é necessário transformar os dados em algum ponto para facilitar a análise, o que geralmente envolve: adicionar, copiar, replicar, limpar e padronizar os dados, além de unificar conjuntos de dados e exportar os arquivos em diferentes formatos.

No exemplo a seguir na tabela 1, utilizamos a biblioteca pandas para converter as abas de um arquivo Excel em arquivos CSV individuais. Cada aba é transformada em um *data frame* e salva como um arquivo CSV com o nome correspondente à aba. O código exibe uma mensagem de conclusão bem-sucedida.

```
import pandas as pd
```

```
base = pd.ExcelFile('database/originals/sorveteria-doce-gosto.xlsx')
abas = base.sheet_names

for aba in abas:
    df = base.parse(aba)
    nome_arquivo_csv = f'{aba}.csv'
    df.to_csv(nome_arquivo_csv, index=False)

print("Conversão concluída. Arquivos CSV gerados com sucesso.")
```

Tabela 1: Transformação de Base de Dados em Excel para Arquivos CSV
Fonte: Elaborada pelos autores

6.2.2 LIMPEZA DOS DADOS

Os dados na sua forma original/bruta/do mundo real são geralmente sujos (Hernández & Stolfo, 1998; Kim et al., 2003). Portanto, nesta etapa os dados são filtrados e limpos.

A limpeza de dados envolve a correção de valores ausentes, erros de digitação, outliers e duplicatas, além de garantir a consistência de formatos, a validade e a anonimização de informações sensíveis. Essa etapa é essencial para assegurar a confiabilidade e a qualidade dos dados usados em análises subsequentes, evitando erros e interpretações inadequadas.

Na tabela 2 preparamos um exemplo usando uma base de vendas, onde as colunas nulas foram removidas, os nomes das colunas foram ajustados, a formatação da coluna 'total' é corrigida, e a coluna de datas é convertida para o formato datetime. O conjunto de dados tratado é então salvo como 'database_sells.csv'.

```
import pandas as pd

vendas = pd.read_csv('database/originals/Vendas.csv')

vendas = vendas.dropna(axis=1)
vendas.columns = ['id', 'data_venda', 'id_cliente', 'id_produto', 'quantidade', 'total']

vendas['total'] = vendas['total'].str.replace('R$', '').str.replace(',', '.', regex=True).astype(float)
vendas['data_venda'] = pd.to_datetime(vendas['data_venda'])

vendas.to_csv('database_sells.csv', index=False)
```

Tabela 2: Limpeza de Base de Dados e Exportação
Fonte: Elaborada pelos autores

6.2.3 ARMAZENAMENTO DOS DADOS PROCESSADOS

Na etapa de armazenamento de dados, os dados coletados são armazenados. Os dados precisam ser armazenados de acordo com a legislação de proteção de dados. Os dados também precisam ser armazenados para que aqueles que precisam utilizá-los possam acessá-los de forma rápida e fácil (University of Technology Sydney, 2022).

No exemplo da tabela 3, uma conexão SQLite é estabelecida ('sorveteria.db'), e um data frame do Pandas é criado a partir de um arquivo CSV ('database_stock.csv'). O conteúdo do data frame é então inserido na tabela 'stock' no banco de dados SQLite.

```
import sqlite3
import pandas as pd

conn = sqlite3.connect('sorveteria.db')
stock = 'database/transformed/database_stock.csv'
df_stock = pd.read_csv(stock)

df_stock.to_sql('stock', conn, if_exists='replace')
```

Tabela 3: Armazenamento dos Dados em Banco de Dados SQL
Fonte: Elaborada pelos autores

6.2.4 INTEGRAÇÃO DOS DADOS

Integração de dados é o processo de reunir dados de diferentes origens para uma visualização unificada e mais prática, para que sua empresa possa tomar decisões melhores e mais rápidas. A integração de dados pode consolidar todos os tipos de dados, estruturados, não estruturados, em lote e streaming, para fazer tudo, desde consultas básicas de bancos de dados de inventário a análises preditivas complexas (Google Cloud, S/D).

No exemplo na tabela 4, dados são extraídos de tabelas em um banco de dados SQLite ('sorveteria.db'). As tabelas 'sells' e 'clients' são mescladas com base nos IDs de clientes, e a tabela resultante é mesclada com a tabela 'stock' usando IDs de produtos. As colunas de IDs são removidas, e o resultado é salvo como 'view_sells_unified.csv'.

```
import pandas as pd
import sqlite3

conn = sqlite3.connect("sorveteria.db")

stock = pd.read_sql("SELECT * FROM stock", conn)
sells = pd.read_sql("SELECT * FROM sells", conn)
clients = pd.read_sql("SELECT * FROM clients", conn)

view_sells = pd.merge(sells, clients[['id_cliente', 'nome_cliente']], on='id_cliente', how='left')
view_sells = pd.merge(view_sells, stock[['id_produto', 'nome_produto']], on='id_produto',
how='left')
view_sells = view_sells.drop(['id_cliente', 'id_produto'], axis=1)

view_sells.to_csv('view_sells_unified.csv', index=False)
```

Tabela 4: Integração entre Bases de dados
Fonte: Elaborada pelos autores

6.3 ANÁLISE DE DADOS

A análise de dados é um compilado de informações sobre determinada ação ou estratégia de uma empresa que dão um feedback preciso sobre todos os aspectos da organização e auxiliam na prospecção de iniciativas de melhorias (Serasa Experian, S/D).

As técnicas de análise de dados podem variar desde análises estatísticas até algoritmos de aprendizado de máquina, dependendo da complexidade dos dados e dos objetivos comerciais.

6.3.1 ANÁLISE DESCRITIVA

Geralmente a análises descritivas são as primeiras manipulações realizadas em um estudo quantitativo e tem como principal objetivo resumir, sumarizar e explorar o comportamento dos dados. Isso pode ser feito através de tabelas de frequências, gráficos e medidas de resumo numérico (Previdelli, S/D).

Ela é útil para entender os dados, identificar padrões e tendências, e responder a perguntas básicas sobre os dados. As técnicas estatísticas mais comuns usadas na análise descritiva são as medidas de tendência central (média, mediana e moda), as medidas de dispersão (desvio padrão, variância e intervalo interquartil) e a frequência. As representações gráficas mais comuns usadas na análise descritiva são o histograma, o diagrama de dispersão e o *boxplot*. Cada tipo de análise possui um propósito específico e deve ser escolhida com base na natureza dos dados e nas perguntas que buscamos responder.

Tendência Central

Uma medida de tendência central é uma medida sumária que tenta descrever todo um conjunto de dados com um único valor que representa o meio ou centro da sua distribuição (Australian Bureau of Statistics, S/D).

Muitos ERPs disponibilizam este tipo de análise por padrão, mas ela também pode ser feita com o auxílio de diversas outras ferramentas, tais como planilhas (Google Planilhas, LibreOffice e Excel), e algumas linguagens de programação, como Python e R. No exemplo da tabela 5 usamos a biblioteca Pandas do Python, onde basta invocar o método *`.describe()`* do data frame para ter um resumo descritivo do data frame.

count	3.0
mean	2.0
std	1.0
min	1.0
25%	1.5
50%	2.0
75%	2.5
max	3.0

Tabela 5: Resultado da Função Describe do Pandas
Fonte: Elaborada pelos autores

Dispersão

Segundo o Statistique Canada (2021) para descrever melhor os dados, também é bom ter uma medida da dispersão dos dados em torno do centro da distribuição. Esta medida é chamada de medida de dispersão. As medidas de dispersão mais comumente usadas são: intervalo, variância e desvio padrão.

No exemplo da tabela 6, um arquivo CSV de vendas ('sheet_sales.csv') é lido usando pandas. Em seguida, são calculadas estatísticas de dispersão, incluindo variância, desvio padrão e amplitude das vendas diárias. Além disso, são determinados os quartis (Q1 e Q3) e o

intervalo interquartil (IQR) da distribuição dos valores na coluna 'total'. Os resultados são impressos para análise estatística das vendas.

```
import pandas as pd

vendas = "sheets/transformed/sheet_sales.csv"
dados = pd.read_csv(vendas)

# estatísticas de dispersão
variancia = round(dados['total'].var(), 2)
desvio_padrao = round(dados['total'].std(), 2)
amplitude = round(dados['total'].max() - dados['total'].min(), 2)

# quartis
q1 = round(dados['total'].quantile(0.25), 2)
q3 = round(dados['total'].quantile(0.75), 2)
iqr = round(q3 - q1, 2)

print(f"Variância das vendas diárias: {variancia}")
print(f"Desvio Padrão das vendas diárias: {desvio_padrao}")
print(f"Amplitude das vendas diárias: {amplitude}")
print(f"Primeiro Quartil (Q1): {q1}")
print(f"Terceiro Quartil (Q3): {q3}")
print(f"Intervalo Interquartil (IQR): {iqr}")
```

Tabela 6: Análise de Dispersão Estatística Sobre Vendas
Fonte: Elaborada pelos autores

Frequência

Frequência e frequência relativa são dois conceitos fundamentais em estatística. Eles descrevem a frequência com que valores ou categorias aparecem em um conjunto de dados e que proporção do conjunto de dados eles representam.

Segundo Geeks for Geeks (2023) a frequência é o número de vezes que um valor ou categoria específica aparece em um conjunto de dados, e a frequência relativa é a proporção ou porcentagem de vezes que um valor ou categoria específica aparece em um conjunto de dados.

No exemplo da tabela 7, um arquivo CSV de vendas ('sheet_sales.csv') é lido usando pandas, e em seguida, a frequência dos valores na coluna 'total' é calculada usando value_counts(). O resultado é exibido como a frequência dos diferentes valores de venda.

```
import pandas as pd

vendas = "sheets/transformed/sheet_sales.csv"
dados = pd.read_csv(vendas)

# frequencia
frequencia = dados['total'].value_counts()

print("Frequencia dos valores de venda:")
print(frequencia)
```

Tabela 7: Frequência Dos Totais de Venda
Fonte: Elaborada pelos autores

Histograma

Um histograma é uma espécie de gráfico de barras que demonstra uma distribuição de frequências. No histograma, a base de cada uma das barras representa uma classe e a altura representa a quantidade ou frequência absoluta com que o valor de cada classe ocorre (Siqueira, 2023).

Na figura 1 exemplificamos um histograma da relação entre número de clientes e idade foi gerado usando a biblioteca Matplot do Python.

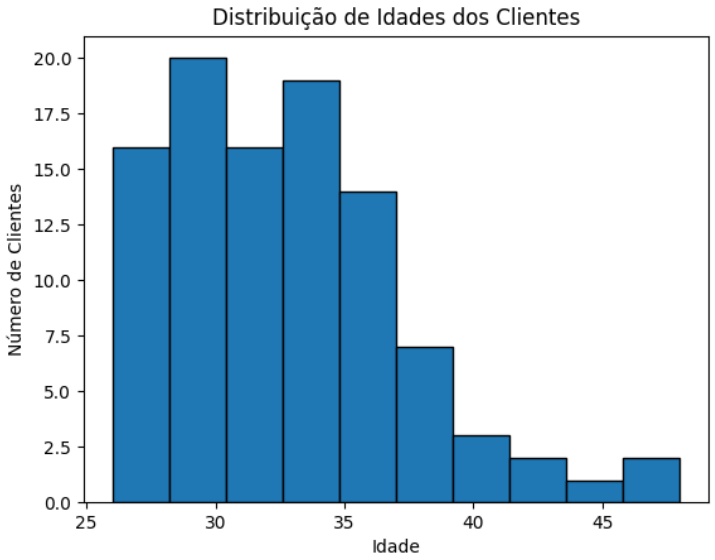

Figura 1: Histograma de Distribuição de Idades dos Clientes
Fonte: Elaborada pelos autores

Diagrama de Dispersão

O diagrama de dispersão representa graficamente pares de dados numéricos, com uma variável em cada eixo, para procurar uma relação entre eles. Se as variáveis estiverem correlacionadas, os pontos cairão ao longo de uma linha ou curva. Quanto melhor a correlação, mais próximos os pontos se aproximam da linha (American Society of Quality, S/D).

No exemplo a seguir na figura 2, um diagrama de dispersão dos totais de vendas mensais durante o ano de 2022 foi gerado usando a biblioteca Matplot do Python.

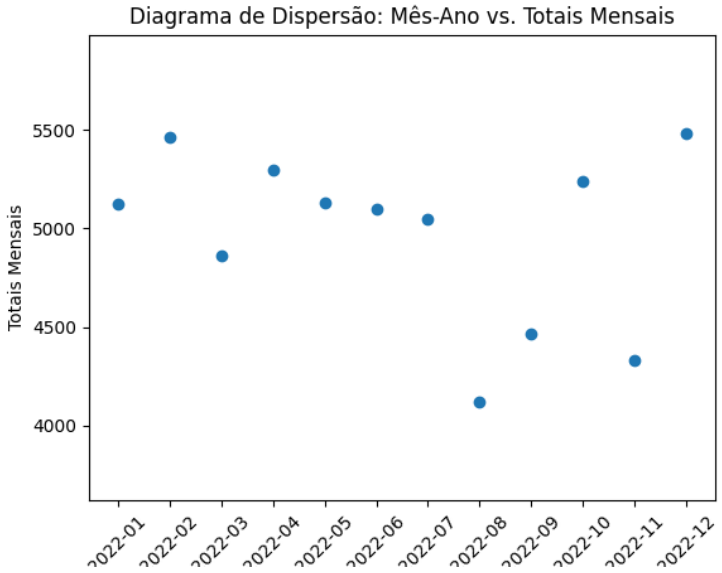

Figura 2: Diagrama de Dispersão das Vendas Mensais durante o ano de 2022
Fonte: Elaborada pelos autores

Boxplot

Um boxplot ou diagrama de caixa é um gráfico que resume um conjunto de dados. A forma do boxplot mostra como os dados são distribuídos e também mostra quaisquer valores discrepantes. É uma maneira útil de comparar diferentes conjuntos de dados, pois você pode desenhar mais de um boxplot por gráfico. Eles podem ser exibidos ao lado de uma linha numérica, horizontal ou verticalmente (Newcastle University, S/D).

Na figura 3 fizemos um exemplo de análise dos gastos mensais de três perfis de clientes em uma loja online. O ‘perfil1’ representa clientes com gastos mais baixos, o ‘perfil2’ representa clientes com gastos médios, e o ‘perfil3’ representa clientes com gastos mais altos. O boxplot permite comparar visualmente a distribuição e a variação dos gastos entre os diferentes perfis de clientes.

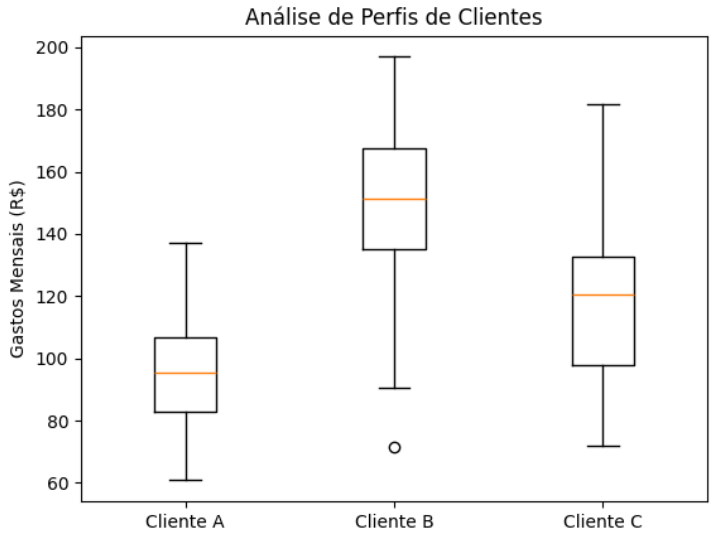

Figura 3: Análise de Perfis de Clientes com Boxplot
Fonte: Elaborada pelos autores

6.3.2 ANÁLISE PREDITIVA

A análise preditiva é o processo de usar dados para prever resultados futuros. O processo usa análise de dados, machine learning, inteligência artificial e modelos estatísticos para encontrar padrões que possam prever comportamentos futuros. As organizações podem usar dados históricos e atuais para prever tendências e comportamentos com segundos, dias ou anos de antecedência, com muita precisão (Google Cloud, S/D).

Muitas empresas e organizações usam análise preditiva para orientar decisões futuras. Por exemplo, os analistas de marketing usam a análise preditiva para determinar as vendas futuras de seus produtos, as estações meteorológicas a usam para previsão do tempo e os corretores da bolsa a usam para maximizar os retornos de negociação (Amazon AWS, S/D).

Existem muitas técnicas e modelos para análises preditivas, cada um com um propósito específico, as técnicas estatísticas mais comuns usadas na análise preditiva são a regressão, classificação e séries temporais. As técnicas de aprendizado de máquina mais utilizadas são as árvores de decisões, redes neurais e algoritmos de reforço.

Regressão

A análise de regressão é usada para prever um valor numérico com base em variáveis independentes. Os modelos de regressão, como a regressão linear e a regressão logística, são comuns nessa área.

Segundo a Amazon AWS (S/D) a regressão linear é uma técnica de análise de dados que prevê o valor de dados desconhecidos usando outro valor de dados relacionado e conhecido. Ele modela matematicamente a variável desconhecida ou dependente e a variável conhecida ou independente como uma equação linear.

A regressão logística é uma técnica de análise de dados que usa matemática para encontrar as relações entre dois fatores de dados. Em seguida, essa relação é usada para prever o valor de um desses fatores com base no outro. A previsão geralmente tem um número finito de resultados, como sim ou não (Amazon AWS, S/D).

No exemplo a seguir, na tabela 8, os dados de uma tabela de marketing são utilizados para realizar uma análise de regressão linear. O investimento em publicidade ('custos') é relacionado às vendas ('alcance'), e o resultado é visualizado por meio de um gráfico de dispersão com a linha de regressão.

```
import pandas as pd
import matplotlib.pyplot as plt
from sklearn.linear_model import LinearRegression
import sqlite3

conn = sqlite3.connect("sorveteria.db")
marketing = pd.read_sql("SELECT * FROM marketing", conn)
df = pd.DataFrame(marketing)

x = df[['custos']]
y = df['alcance']

model = LinearRegression()
model.fit(x, y)

previsoes = model.predict(x)
plt.scatter(x, y, color='blue')
```

```
plt.plot(x, previsoes, color='red', linewidth=2)
plt.xlabel('Investimento em Publicidade (R$)')
plt.ylabel('Vendas (R$)')
plt.title('Regressão Linear: Investimento em Publicidade vs. Vendas')
plt.grid(True)

file_name = 'plt_analysis_regression.png'
diretorio_destino = 'plots/predictive/'
plt.savefig(f'{diretorio_destino}{file_name}')
```

Tabela 8: Análise de Regressão dos Investimentos em Publicidade
Fonte: Elaborada pelos autores

Classificação

Na ciência de dados e nas estatísticas, a classificação é definida como a identificação de quais categorias (às vezes chamadas de subpopulações) uma nova observação deve ser incluída, com base em um conjunto de treinamento de dados contendo observações (ou instâncias) cuja adesão à categoria foi validada (O'reilly, S/D).

No exemplo a seguir, na tabela 9, usamos uma base que continham dados sobre o cliente, páginas acessadas e se ele efetuou uma compra, a partir disso utilizamos um modelo de classificação do SkLearn para prever se um cliente específico fará uma compra com base em seu comportamento de navegação no site ou não.

```
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import accuracy_score, classification_report

dados = pd.read_csv('dados_do_cliente.csv')

X = dados.drop('Compra', axis=1)
y = dados['Compra']
X_treino, X_teste, y_treino, y_teste = train_test_split(X, y, test_size=0.2, random_state=42)

modelo = RandomForestClassifier()
modelo.fit(X_treino, y_treino)

previsoes = modelo.predict(X_teste)
acuracia = accuracy_score(y_teste, previsoes)
relatorio_classificacao = classification_report(y_teste, previsoes)

print(f'Acurácia do modelo: {acuracia}')
print(f'Relatório de Classificação:\n{relatorio_classificacao}')

novo_cliente = pd.DataFrame({'tempo_gasto': [10], 'paginas_visitadas': [5], 'produtos_visualizados':
[3]})
previsao_novo_cliente = modelo.predict(novo_cliente)

print(f'Previsão para o novo cliente: {previsao_novo_cliente}')
```

Tabela 9: Técnica de classificação para Análise Preditiva de Vendas Online
Fonte: Elaborada pelos autores

Séries Temporais

A análise de séries temporais refere-se a problemas nos quais as observações são coletadas regularmente. intervalos de tempo e há correlações entre observações sucessivas (Cambridge University, S/D).

No exemplo da tabela 10, os dados da tabela ‘stock’ são usados para gerar uma série temporal do estoque, que é visualizada através de um gráfico, onde o eixo ‘x’ representa o ID do produto e o eixo ‘y’ representa a quantidade em estoque.

```
import pandas as pd
import matplotlib.pyplot as plt
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error
import sqlite3

conn = sqlite3.connect("sorveteria.db")
stock = pd.read_sql("SELECT * FROM stock", conn)
df = pd.DataFrame(stock)
df.set_index('id_produto', inplace=True)

X = df.index.values.reshape(-1, 1)
y = df['estoque'].values
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

model = LinearRegression()
model.fit(X_train, y_train)

y_pred = model.predict(X_test)

mse = mean_squared_error(y_test, y_pred)
print(f'Mean Squared Error: {mse}')

future_values = model.predict(X_test.reshape(-1, 1) + 1)

plt.figure(figsize=(12, 6))
plt.plot(df.index, df['estoque'], label='Histórico de Estoque')
plt.plot(X_test, y_pred, label='Previsões no Conjunto de Teste', linestyle='dashed')
plt.plot(X_test + 1, future_values, label='Previsões Futuras', linestyle='dashed', color='red')
plt.title('Análise Preditiva da Série Temporal do Estoque')
plt.xlabel('ID do Produto')
plt.ylabel('Estoque')
plt.legend()
plt.grid(True)

file_name = 'plt_predictive_analysis.png'
out_dir = 'plots/predictive/'
plt.savefig(f'{out_dir}{file_name}')
```

Tabela 10: Série Temporal da Quantidade em Estoque de Cada Produto
Fonte: Elaborada pelos autores

Árvores de decisão

A análise em árvore de decisão implica delinear visualmente os possíveis resultados, custos e consequências de uma decisão complexa. Este método é particularmente útil para a análise de dados quantitativos e a tomada de decisão baseada em números (Asana, 2023).

A árvore de decisão pode esclarecer para a gestão, como nenhuma outra ferramenta analítica que eu conheça, as escolhas, os riscos, os objetivos, os ganhos monetários e as necessidades de informação envolvidas num problema de investimento (Magee, 1964).

No exemplo da tabela 11, dados da tabela ‘dados_sorveteria’ são usados para treinar um modelo de classificação de árvore de decisão, onde o modelo é usado para prever a variável ‘Compra_Sorvete’ com base nas features 'Temperatura' e ‘Preco’. A precisão do modelo é avaliada no conjunto de teste, e a estrutura da árvore de decisão é exibida.

```
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.tree import DecisionTreeClassifier, export_text

conn = sqlite3.connect("sorveteria.db")
stock = pd.read_sql("SELECT * FROM dados_sorveteria", conn)

df = pd.DataFrame(dados_sorveteria)

# features (X) e target (y)
X = dados_sorveteria[['Temperatura', 'Preco']]
y = dados_sorveteria['Compra_Sorvete']

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

modelo_arvore = DecisionTreeClassifier(random_state=42)
modelo_arvore.fit(X_train, y_train) # Treinar o modelo

precisao = modelo_arvore.score(X_test, y_test)
print(f'A precisão do modelo é: {precisao:.2f}')

texto_arvore = export_text(modelo_arvore, feature_names=list(X.columns))
print("Árvore de Decisão:")
print(texto_arvore)
```

Tabela 11: Modelo de Árvore de Decisão de Compra Sobre a Relação Temperatura Ambiental e Preço do Sorvete
Fonte: Elaborada pelos autores

Redes neurais

As redes neurais, também conhecidas como redes neurais artificiais (ANNs) ou redes neurais simuladas (SNNs), são um subconjunto de machine learning e estão no cerne dos algoritmos de deep learning. Seu nome e estrutura são inspirados no cérebro humano, imitando a maneira como os neurônios biológicos enviam sinais uns para os outros (IBM, S/D).

As redes neurais contam com dados de treinamento para aprender e melhorar sua precisão ao longo do tempo. No entanto, uma vez que esses algoritmos de aprendizagem são ajustados para aumentar a precisão, eles se tornam ferramentas poderosas de ciência da computação e inteligência artificial, permitindo-nos classificar e agrupar dados a uma alta velocidade (IBM, S/D).

Em um exemplo prático, uma loja online de moda implementa uma rede neural para aprimorar a personalização das recomendações de produtos. Alimentando a rede neural com dados de preferências de compra passadas, comportamento de navegação e feedback dos clientes, o modelo aprende padrões complexos para sugerir itens de moda de forma mais precisa. A utilização da rede neural resulta em recomendações altamente personalizadas, impulsionando as taxas de conversão e elevando a satisfação do cliente.

Algoritmos de Reforço

Algoritmos de reforço ou boosting é um método usado em aprendizado de máquina para reduzir erros na análise preditiva de dados. Os cientistas de dados treinam software de aprendizado de máquina, chamados de modelos de aprendizado de máquina, em dados rotulados para fazer suposições sobre dados não rotulados (Amazon AWS, S/D).

Exemplo prático, uma pequena empresa de entrega utiliza algoritmos de reforço para otimizar suas rotas de entrega. O algoritmo aprende com a experiência, ajustando as rotas com base em fatores dinâmicos, como tráfego e demanda. À medida que os veículos entregam pedidos, o sistema reforça as ações que levam a entregas mais rápidas e eficientes. Isso resulta em economia de tempo, redução de custos operacionais e uma melhoria geral na eficiência do serviço de entrega.

6.3.3 ANÁLISE PRESCRITIVA

A análise prescritiva é o processo de uso de dados para determinar um curso de ação ideal. Ao considerar todos os fatores relevantes, este tipo de análise produz recomendações para os próximos passos. Por causa disso, a análise prescritiva é uma ferramenta valiosa para a tomada de decisões baseada em dados (Harvard Business School Online, 2021).

Essa forma de análise de negócios pode mostrar o melhor curso de ação para determinada situação. Enquanto a análise descritiva mostra o que já aconteceu e a análise preditiva tenta prever o que pode acontecer a seguir, a prescritiva usa essas informações para fornecer soluções possíveis, com base em situações semelhantes (Microsoft 365 Team, 2019). As técnicas mais comuns de análise prescritivas envolvem a otimização e aprendizado de máquina.

Otimização

Segundo a Washington University (S/D) a otimização é uma técnica matemática que tem o objetivo de maximizar ou minimizar alguma função em relação a algum conjunto, muitas vezes representando uma gama de escolhas disponíveis em uma determinada situação. A função permite a comparação das diferentes opções para determinar qual pode ser a “melhor”. Aplicações comuns: Custo mínimo, lucro máximo, erro mínimo, design ideal, gestão ótima, princípios variacionais.

Em um exemplo prático uma MPE pode adotar um sistema de otimização de estoque, onde através de algoritmos de previsão de demanda e machine learning, o sistema analisa padrões de vendas passadas, eventos sazonais e fatores externos para ajustar os níveis de estoque e evitar excessos ou escassez.

Machine Learning

O machine learning (ML) é o subconjunto da inteligência artificial (IA) que se concentra na construção de sistemas que aprendem, ou melhoram o desempenho, com base nos dados que consomem. A inteligência artificial é um termo amplo que se refere a sistemas ou máquinas que imitam a inteligência humana (Oracle, S/D).

O machine learning e a IA são frequentemente abordados juntos, e os termos às vezes são usados de forma intercambiável, mas não significam a mesma coisa. Uma distinção

importante é que, embora todo machine learning seja IA, nem toda IA é machine learning (Oracle, S/D).

Para ilustrar essa distinção, considere um e-commerce que utiliza machine learning para aprimorar as recomendações de produtos. Ao analisar dados de compras passadas e preferências, o sistema treina um modelo de recomendação personalizado, proporcionando sugestões em tempo real. O sucesso dessa implementação resulta em experiências de compra mais personalizadas, gerando um impulso nas vendas e na satisfação do cliente. A manutenção contínua do modelo garante sua adaptabilidade às mudanças nas preferências dos clientes.

6.4 INTERPRETANDO INFORMAÇÕES

A interpretação de informações é uma etapa crucial na análise de dados para micro e pequenas empresas (MPEs). Envolve a extração de insights significativos a partir dos resultados da análise. Esses insights são então relacionados aos objetivos de negócios das MPEs, avaliando como podem contribuir para metas específicas. É essencial contextualizar os insights no contexto da empresa, determinando ações práticas e estratégicas que podem moldar a trajetória de sucesso da MPE.

6.5 ESTRATÉGIAS DE IMPLEMENTAÇÃO

As micro e pequenas empresas (MPEs) podem implementar ciência de dados de maneira eficaz adotando várias estratégias, que podem incluir treinamento interno, terceirização de serviços, parcerias estratégicas, e consultorias. Além disso, a escolha dos softwares que auxiliam nos processos de ciência de dados também é muito importante, as micro e pequenas empresas (MPEs) podem aproveitar uma variedade de ferramentas e recursos de licença comercial livres para implementar a ciência de dados em áreas-chave de suas operações de forma gratuita. Na tabela 12 listamos algumas gratuitas que podem ser usadas para implementação da ciência de dados nas MPEs.

Tipo	Software
Planilhas	Google Planilhas, LibreOffice
ERPs	ERPNext, Dolibarr
Linguagens	Python, R, Kotlin
Bibliotecas	Pandas, Scikit, Matplot
Bancos de Dados	MariaDB, SQLite, Redis

Tabela 12: Ferramentas Gratuitas para Implementação da Ciência de Dados
Fonte: Elaborada pelos autores

O uso de soluções em nuvem acessíveis e ferramentas de análise de dados, a implementação gradual de projetos de menor escala e a promoção de uma cultura de dados, bem como o acompanhamento contínuo dos resultados para ajustar as estratégias à medida que a empresa evolui. A escolha da estratégia depende das características específicas da MPE, seus recursos e objetivos de negócios.

7. RESULTADOS E DISCUSSÕES

Considera-se que a adoção da ciência de dados nas MPEs pode ter um impacto positivo significativo, com a melhoria da capacidade de compreender as necessidades dos clientes, identificar tendências de mercado e personalizar suas ofertas. O que consequentemente resulta na melhora em seus índices de desempenho (KPIs), o que, por sua vez, impulsiona o desenvolvimento das MPEs.

No entanto, durante nossas discussões, também destacamos os desafios que as MPEs enfrentam ao implementar a ciência de dados, a falta de recursos financeiros e mão de obra especializada pode ser uma barreira significativa. Muitas MPEs ainda precisam superar essas limitações para aproveitar totalmente os benefícios da análise de dados.

A busca por profissionais capacitados é essencial para MPEs ao incorporar a ciência de dados. Procurar talentos em instituições de ensino e estabelecer parcerias com consultores e empresas especializadas oferece acesso à expertise, mesmo com recursos limitados. Investir no desenvolvimento da equipe por meio de treinamentos cria uma cultura orientada a dados, capacitando as MPEs a superar os desafios da implementação.

8. CONCLUSÃO

Este estudo destaca que a ciência de dados desempenha um papel crucial na transformação das MPEs em organizações ágeis e orientadas por dados, capazes de tomar decisões estratégicas, táticas e operacionais mais informadas e rápidas. Concluímos que a adoção da ciência de dados é uma estratégia essencial para o sucesso sustentável das MPEs em um ambiente de negócios cada vez mais complexo e dinâmico.

9. REFERÊNCIAS BIBLIOGRÁFICAS

AUSTRALIAN BUREAU OF STATISTICS. **Measures of central tendency**. S/D. Disponível em: <https://www.abs.gov.au/statistics/understanding-statistics/statistical-terms-and-concepts/measures-central-tendency#:~:text=Measures%20of%20shape-,Definition,or%20centre%20of%20its%20distribution>. Acesso em 28 de outubro de 2023.

AMAZON AWS. **O que é análise preditiva?**. S/D. Disponível em: <https://aws.amazon.com/pt/what-is/predictive-analytics/>. Acesso em 21 de outubro de 2023.

AMAZON AWS. **O que é regressão logística?.** S/D. Disponível em : <https://aws.amazon.com/pt/what-is/logistic-regression/#:~:text=A%20regress%C3%A3o%20log%C3%ADstica%20%C3%A9%20uma,resultados%2C%20como%20sim%20ou%20n%C3%A3o>. Acesso em: 29 de outubro de 2023.

AMAZON AWS. O **que é regressão linear?.** S/D. Disponível em : <https://aws.amazon.com/pt/what-is/linear-regression/#:~:text=A%20regress%C3%A3o%20linear%20%C3%A9%20uma,independente%20como%20uma%20equa%C3%A7%C3%A3o%20linear>. Acesso em: 29 de outubro de 2023.

AMAZON AWS. **What is boosting in machine learning?**. S/D. Disponível em: <https://aws.amazon.com/what-is/boosting/#:~:text=Boosting%20algorithms%20combine%20multiple%20weak,common%20in%20machine%20learning%20models>. Acesso em 28 de outubro de 2023.

AMERICAN SOCIETY OF QUALITY. **WHAT IS A SCATTER DIAGRAM?**. S/D. Disponível em: <https://asq.org/quality-resources/scatter-diagram#:~:text=The%20scatter%20diagram%20graphs%20pairs,points%20will%20hug%20the%20line>. Acesso em 28 de outubro de 2023.

ASANA. **O que é uma análise em árvore de decisão? Cinco passos para tomar melhores decisões.** 2023. Disponível em : <https://asana.com/pt/resources/decision-tree-analysis>. Acesso em: 29 de outubro de 2023.

BRIETZIG, N. G., MIRANDA, J. B. **Análise de dados como ferramenta de tomada de decisão para micro e pequenas empresas**. 2022. Disponível em : <https://repositorio.animaeducacao.com.br/handle/ANIMA/25674>. Acesso em 16 de setembro de 2023.

DORNELAS, J. C. A. (2005). **Gestão de pequenas empresas: estratégias e habilidades essenciais.** São Paulo: Elsevier.

GOOGLE CLOUD. **O que é análise preditiva?**. S/D. Disponível em: <https://cloud.google.com/learn/what-is-predictive-analytics>. Acesso em 21 de outubro de 2023.

GOOGLE CLOUD. **O que é integração de dados?**. S/D. Disponível em : <https://cloud.google.com/learn/what-is-data-integration?hl=pt-br#section-1>. Acesso em : 22 de outubro de 2023.

HARVARD BUSINESS SCHOOL ONLINE. **What is prescriptive analytics? 6 examples**. 2021. Disponível em: <https://online.hbs.edu/blog/post/prescriptive-analytics>. Acesso em 21 de outubro de 2023.

HERNÁNDEZ, M. A., & STOLFO, S. J. 1998. Real-world data is dirty: **Data cleansing and the merge/purge problem**. *Data MIning and Knowledge Discovery,* 2(1), 9-37.

IBM. **What is Data Science?.** S/D. Disponível em : <https://www.ibm.com/br-pt/topics/data-science>. Acesso em 16 de setembro de 2023.

IBM. **O que são redes neurais?.** S/D. Disponível em : <https://www.ibm.com/br-pt/topics/neural-networks>. Acesso em: 29 de outubro de 2023.

KOTSIANTIS, S. B., KANELLOPOULOS, D., PINTELAS, P., E. (2006). **Data preprocessing for supervised learning**. *International Journal of Computer Science*, 1(2), 111-117.

MAGEE, J. F. **Decision Trees for Decision-Making.** 1964. Disponível em : <https://hbr.org/1964/07/decision-trees-for-decision-making>. Acesso em: 29 de outubro de 2023.

MATPLOT. **Matplotlib**. S/D. Disponível em: <https://matplotlib.org/>. Acesso em: 22 de outubro de 2023.

MELO, M. **O que é transformação de dados?**. 2022. Disponível em : <https://www.dio.me/articles/o-que-e-transformacao-de-dados>. Acesso em : 22 de outubro de 2023.

MICROSOFT 365 TEAM. **Os benefícios da análise de negócios**. 2019. Disponível em: <https://www.microsoft.com/pt-br/microsoft-365/business-insights-ideas/resources/benefits-of-business-analytics>. Acesso em 21 de outubro de 2023.

OLIVEIRA, MÁRCIO. **O que é ciência de dados?.** 2023. Disponível em : <https://www.dio.me/articles/o-que-e-ciencia-de-dados-CVE416>. Acesso em 16 de setembro de 2023.

ORACLE. **O que é Machine Learning?.** S/D. Disponível em : <https://www.oracle.com/br/artificial-intelligence/machine-learning/what-is-machine-learning/>. Acesso em: 29 de outubro de 2023.

PANDAS. **Pandas.** S/D. Disponível em: <https://pandas.pydata.org/>. Acesso em: 22 de outubro de 2023.

PORTAL DA INDÚSTRIA. **Qual a definição de micro e pequena empresa?.** S/D. Disponível em: <https://www.portaldaindustria.com.br/industria-de-a-z/micro-e-pequena-empresa/>. Acesso em 13 de maio de 2023.

PREVIDELLI, I. **Análise Descritiva**. S/D. Disponível em: <https://biostatistics-uem.github.io/Bio/descritiva.html>. Acesso em 21 de outubro de 2023.

PYTHON. **Python.** S/D. Disponível em: <https://www.python.org/>. Acesso em: 22 de outubro de 2023.

SCIKIT-LEARN. **Scikit-learn.** S/D. Disponível em: <https://scikit-learn.org/stable/>. Acesso em: 22 de outubro de 2023.

SEBRAE. **Confira as diferenças entre micro empresa, pequena empresa e MEI.** 2013. Disponível em:
<https://sebrae.com.br/sites/PortalSebrae/artigos/entenda-as-diferencas-entre-microempresa-pequena-empresa-e-mei,03f5438af1c92410VgnVCM100000b272010aRCRD
>. Acesso em 13 de maio de 2023.

SERASA EXPERIAN. **Análise de dados: o que é, tipos e como implantar.** 2023. Disponível em : <https://www.serasaexperian.com.br/conteudos/marketing/analise-de-dados-o-que-e-tipos-e-como-implantar/>. Acesso em 16 de setembro de 2023.

SHARDA, R., DELEN, D., & TURBAN, E. (2015). **Business Intelligence, Analytics and Data Science: A Managerial Perspective.** Pearson.

SIQUEIRA, D. **Histograma: O que é, Exemplos, Gráficos e Tipos**. 2023. Disponível em: <https://www.alura.com.br/artigos/o-que-e-um-histograma#:~:text=O%20que%20%C3%A9%20histograma%3F,valor%20de%20cada%20classe%20ocorre>. Acesso em: 22 de outubro de 2023.

SQLITE. **SQLite.** S/D. Disponível em: <https://www.sqlite.org/index.html>. Acesso em: 22 de outubro de 2023.

STATISTIQUE CANADA. **Measures of dispersion**. 2021. Disponível em: <https://www150.statcan.gc.ca/n1/edu/power-pouvoir/ch12/5214876-eng.htm>. Acesso em 28 de outubro de 2023.

UNIVERSITY OF TECHNOLOGY OF SYDNEY. **What is data processing?**. 2022. Disponível em : <https://studyonline.uts.edu.au/blog/what-data-processing>. Acesso em : 22 de outubro de 2023.

GEEKS FOR GEEKS. **Data Transformation in Data Mining**. 2023. Disponível em: <https://www.geeksforgeeks.org/data-transformation-in-data-mining/>. Acesso em 28 de outubro de 2023.

GEEKS FOR GEEKS. **Difference Between Frequency and Relative Frequency**. 2023. Disponível em: <https://www.geeksforgeeks.org/difference-between-frequency-and-relative-frequency/>. Acesso em 28 de outubro de 2023.

NEWCASTLE UNIVERSITY. **Box and Whisker Plots**. S/D. Disponível em: <https://www.ncl.ac.uk/webtemplate/ask-assets/external/maths-resources/statistics/data-presentation/box-and-whisker-plots.html>. Acesso em 28 de outubro de 2023.

O'REILLY. **Data classification in statistics**. S/D. Disponível em: <https://www.oreilly.com/library/view/statistics-for-data/9781788290678/3c00d80d-39d6-4d35-953c-bfa0f0df0647.xhtml>. Acesso em 28 de outubro de 2023.

CAMBRIDGE UNIVERSITY. **TIME SERIES**. S/D. Disponível em: <https://www.statslab.cam.ac.uk/~rrw1/timeseries/t.pdf>. Acesso em 28 de outubro de 2023.

WASHINGTON UNIVERSITY. **WHAT IS OPTIMIZATION?** S/D. Disponível em: <https://sites.math.washington.edu/~burke/crs/515/notes/nt_1.pdf>. Acesso em 28 de outubro de 2023.