UNIVERSIDADE FEDERAL DE SANTA CATARINA
CENTRO TECNOLÓGICO
DEPARTAMENTO DE ENGENHARIA DE PRODUÇÃO E SISTEMAS
CURSO ENGENHARIA DE PRODUÇÃO MECÂNICA

Leonardo José da Silva

Proposta de um painel gerencial baseado em mineração de dados para uma empresa do setor de saúde

Florianópolis
2022

Leonardo José da Silva

Proposta de um painel gerencial baseado em mineração de dados para uma empresa do setor de saúde

Trabalho Conclusão de Curso de Graduação em Engenharia de Produção Mecânica do Centro de Tecnológico da Universidade Federal de Santa Catarina como requisito para a obtenção do título de Bacharel em Engenharia de Produção Mecânica.
Orientador: Prof. Dr. Guilherme Ernani Vieira

Florianópolis
2022

Ficha de identificação da obra elaborada pelo autor,
através do Programa de Geração Automática da Biblioteca Universitária da UFSC.

Silva, Leonardo Jose
Proposta de um painel gerencial baseado em mineração de dados para uma empresa do setor de saúde / Leonardo Jose Silva ; orientador, Guilheme Ernani Vieira, 2022.
72 p.

Trabalho de Conclusão de Curso (graduação) - Universidade Federal de Santa Catarina, Centro Tecnológico, Graduação em Engenharia de Produção Mecânica, Florianópolis, 2022.

Inclui referências.

1. Engenharia de Produção Mecânica. 2. Painel gerencial. 3. Mineração de dados . 4. Business Intelligence. I. Vieira, Guilheme Ernani. II. Universidade Federal de Santa Catarina. Graduação em Engenharia de Produção Mecânica. III. Título.

Leonardo José da Silva

Proposta de um painel gerencial baseado em mineração de dados para uma empresa do setor de saúde

Este Trabalho Conclusão de Curso foi julgado adequado para obtenção do Título de Bacharel e aprovado em sua forma final pelo Curso Engenharia de Produção Mecânica

Florianópolis, 22 de julho de 2022.

Documento assinado digitalmente
Rogerio Feroldi Miorando
Data: 30/07/2022 20:42:48-0300
CPF: 598.652.290-72
Verifique as assinaturas em https://v.ufsc.br

Prof. Rogério Feroldi Miorando, Dr.
Coordenadora do Curso

Banca Examinadora:

Documento assinado digitalmente
Guilherme Ernani Vieira
Data: 29/07/2022 16:59:45-0300
CPF: 888.311.759-04
Verifique as assinaturas em https://v.ufsc.br

Prof. Guilherme Ernani Vieira, Dr.
Orientador
Universidade Federal de Santa Catarina

Documento assinado digitalmente
SERGIO FERNANDO MAYERLE
Data: 29/07/2022 20:14:23-0300
CPF: 344.463.119-72
Verifique as assinaturas em https://v.ufsc.br

Prof. Sérgio Fernando Mayerle, Dr.
Avaliador
Universidade Federal de Santa Catarina

Documento assinado digitalmente
Marco Antonio de Oliveira Vieira Goulart
Data: 29/07/2022 17:00:26-0300
CPF: 038.879.909-94
Verifique as assinaturas em https://v.ufsc.br

Prof. Marco Antonio de Oliveira Vieira Goulart, Dr.
Avaliador
Universidade Federal de Santa Catarina

Este trabalho é dedicado aos meus queridos pais e as grandes amizades que construí na Universidade.

AGRADECIMENTOS

Agradeço, antes de tudo, meus pais, Maria e José, por todos os ensinamentos, amor, apoio, incentivo e confiança. Que, em momentos de insegurança, trouxeram acalento e resguardo, mesmo que eu duvidasse de minhas capacidades.

Parte fundamental dos cinco anos que estive na Universidade, agradeço, em especial, a Yasmin, por ter sido minha dupla de curso e de vida, dividindo as lutas e celebrando as glórias. Ao João Henrique, por ter sido meu parceiro ao longo de toda a jornada. A Joana, uma das primeiras pessoas que me aproximei em Florianópolis, por ter sido uma grande amiga e me apoiado em diferentes momentos. A Esther, por todo o apoio e, principalmente, por ser estabilidade em momentos de tensão. A Ana, por, em pouquíssimo tempo, se tornar referência de apoio e amizade. Ao Felipe, Robson, Grappegia e Rafael, por se tornarem parceiros com os quais dividi incontáveis momentos desde o início da graduação.

Aos integrantes da Gestão Pluralidade, Marina, Letícia, Gabriel e Vitor, pela confiança e apoio para assumir o projeto que mais me orgulho, tornando-se amigos e desenvolvendo uma gestão pautada em diversidade.

Aos amigos do "Piscininha", Aline e Matheus, pelos bons momentos. Aos que me aproximei, ainda mais, durante o período pandêmico, Lucca, Leticia, Bernardo, Júlio, Letycia, Ana Garcia, Henrique, Gabriel Cidade, Gustavo e Antônio. Ao João Vitor Goedert, que, apesar das adversidades, viu em mim um amigo em quem poderia apostar e confiar, resultando em uma amizade que levo com muito carinho. Aos companheiros das mesas do CETEC, em especial, Osório, Hélio, Rafael e Matos, por momentos de descontração nas semanas conturbadas de graduação.

Ao Centro Acadêmico Livre de Engenharia de Produção (CALIPRO), por ter sido a minha primeira experiencia disruptiva, agregando valor a minha carreira profissional e me modelando como um cidadão que admiro. Ao PET Produção e ao Lempi por todas as capacitações e experiências enriquecedoras.

Por fim, ao professor Guilherme, por todo o suporte, auxílio e confiança, para a execução do trabalho, além de todos os professores dos cursos de Engenharia de Produção Mecânica, por todos os ensinamentos ao longo dos anos.

“A tarefa não é tanto ver aquilo que ninguém viu,
mas pensar o que ninguém ainda pensou
sobre aquilo que todo mundo vê.”
Arthur Schopenhauer

RESUMO

No contexto de uma empresa do setor de saúde, inserida em um ambiente altamente competitivo, identificar, apoiado em resultados visuais, seu desempenho individual e ante aos concorrentes, é parte essencial para sua consolidação no setor. A partir desta premissa, o objetivo deste trabalho é desenvolver, suportado nas etapas do processo CRISP-DM de mineração de dados, um painel gerencial capaz de solidificar as informações presentes no vasto banco de dados disponibilizado pela empresa e apresentá-las de maneira visual, a partir de um modelo de *Business Intelligence*. O modelo teste do painel foi validado com base nos demais relatórios da empresa e por cálculos desenvolvidos na própria base de dados. Toda a pesquisa buscou modelar indicadores capazes de expor, clara e visualmente, informações financeiras e comerciais para a empresa, além de desenvolver uma previsão quanto a seu faturamento. O método de pesquisa foi o de pesquisa-ação, adotando análises quantitativas e qualitativas. Como resultado, obteve-se um painel gerencial robusto que unifica indicadores desenvolvidos: Faturamento por canal, estados e ao longo dos meses, Projeção do faturamento, *Market Share* por marcas e Efeitos Preço, Volume e Mix. Percebeu-se a eficiência do modelo de mineração e a versatilidade no trato de dados pela ferramenta de *Business Intelligence.* Por fim, após as conclusões, o estudo apresenta algumas sugestões de trabalhos futuros.

Palavras-chave: Painel Gerencial. Mineração de Dados. *Busines Intelligence*.

ABSTRACT

In the context of a company in the health sector, inserted in a highly competitive environment, identifying, based on visual results, its individual performance compared to their competitors is an essential part of its consolidation in the sector. Based on this premise, the objective of this work is to develop, supported by the stages of the CRISP-DM data mining process, a management panel capable of solidifying the information present in the vast database provided by the company and presenting it in a visual manner, from a Business Intelligence model. The panel test model was validated based on other company reports and calculations developed in the database itself. The entire research sought to model indicators capable of displaying, clearly and visually, financial and commercial information for the company, in addition to developing a forecast regarding its revenue. The research method was action research, adopting quantitative and qualitative analyses. As a result, a robust management panel was obtained, unifying the following developed indicators: Revenue by channel, states and over the months, Revenue Projection, Market Share by brands and Price, Volume and Mix Effects. It was noticed the efficiency of the mining model and the versatility in the handling of data by the Business Intelligence tool. Finally, after the conclusions, the study presents some suggestions for future work.

Keywords: Management Panel. Data Mining. Business Intelligence.

LISTA DE FIGURAS

LISTA DE ABREVIATURAS E SIGLAS

ABEPRO	Associação Brasileira de Engenharia de Produção
BI	*Business Intelligence*
CRISP-DM	*Cross-Industry Standard Process for Data Mining*
CSV	*Comma Separated Values*
EAN	*European Article Number*
KDD	*Knowledge Discovery in Databases*
KPI	*Key Performance Indicator*
OLAP	*Online Analytical Processing*
PDF	*Portable Document Format*
SQL	*Standard Query Language*
TXT	*Text File Format*
UF	Unidade Federativa

SUMÁRIO

1 INTRODUÇÃO

O ambiente em que as organizações estão inseridas está cada vez mais complexo e turbulento (TURBAN *et al*., 2009). Com a vasta competitividade do mercado, implementar ferramentas que facilitem a tomada de decisões no cotidiano se faz cada vez mais necessário. Tratar de acessibilidade e identificar possíveis problemáticas de gestão são corroboradas com base em dados que mensuram os resultados, partindo de decisões subjetivas para decisões cada vez mais fundamentadas e baseadas em fatos. Atualmente, é muito importante que os gestores de uma empresa tenham uma visão sistêmica dos setores que compõe sua organização, tratando desde áreas administrativas até as ligadas a produção, por exemplo.

Segundo Goldschmidt e Passos (2005), em virtude das diversas tecnologias como internet, sistemas gerenciadores de banco de dados, leitores de códigos de barra e sistemas de informação em geral, são exemplos, dentre os mais variados, de recursos que têm viabilizado o aumento no volume de dados armazenados.

Segundo especialistas, estudiosos e até empreendedores, os dados podem ser considerados o novo petróleo da economia global, visto o seu potencial para a geração de valor, riqueza material e inovação (WEDEL; KANNAN, 2016). Ainda, de acordo com estimativas, o potencial de geração de riqueza, por meio da análise efetiva dos dados existentes supera o valor de 300 bilhões de dólares (MCKINSEY GLOBAL INSTITUTE, 2016), caracterizando-o como fator chave para a sustentabilidade dos negócios.

Deter informações, em virtude dos avanços tecnológicos, se tornou cada vez mais fácil. Em contrapartida, o direcionamento dessas informações é cada vez mais complexo, uma vez que com muitos dados, variadas são as interpretações possíveis. Ainda, reitera-se a problemática quanto a visualização de resultados, trazendo consigo a imprecisão de análises quando se tem, simplesmente, um banco de dados repleto de informações sem qualquer interpretação aparente.

Estudos comprovam que 83% das pessoas absorvem mais informações pela visão. Isso corrobora a hipótese de que apresentar dados de forma clara, transformando os indicadores quantitativos ou qualitativos em representações gráficas, podem ser fortes aliados no que tange a análise de desempenho ou para identificação de agentes detratores ou promotores quanto a algum indicador (KAIZEN INSTITUTE, 2016).

De acordo com Duarte (2012), um *Dashboard* pode ser definido como uma interface gráfica com capacidade de recolher, sumarizar e apresentar informações provenientes de múltiplas fontes, podendo ser uma saída interessante para que as empresas estejam cada vez mais próximas de uma boa articulação quanto a seus dados. As ferramentas de *Business Intelligence* (BI) permitem coletar, organizar, analisar e compartilhar o maior volume possível de informações sobre uma instituição (BIANCHI; SOUSA, PEREIRA, 2012q).

Diante disso, o presente trabalho tem como objetivo propor a implementação de um painel gerencial, capaz de apresentar de um modo gerencial, indicadores que facilitem a tomada de decisão pelos gestores e funcionários da empresa em análise. O modelo apresentado neste trabalho, faz uso de uma base de dados, que apresenta informações de uma multinacional do setor de saúde, em relação ao canal de farmacêutico de distribuição, correlacionando seus produtos com os respectivos concorrentes. Assim, o trabalho irá desenvolver um modelo de visualização que, além de apresentar os resultados, será capaz de apresentar previsões quanto ao desempenho nos próximos meses.

1.1 FORMULAÇÃO DA SITUAÇÃO PROBLEMA

De maneira geral, todas as empresas tem como objetivo ampliar cada vez mais sua lucratividade. Buscar estratégias que possibilitem essa ampliação é parte do desafio cotidiano de todas elas. Dessa maneira, é imprescindível entender alguns indicadores que possam contribuir para que os gestores tracem boas estratégias em relação aos objetivos da organização. Como atualmente o número de dados é cada vez maior, saber o que analisar e como apresentar isso de uma maneira clara e objetiva, é um grande desafio, uma vez que com grande volume de dados, muitas são as interpretações possíveis. Assim, haja vista o grande volume de dados que a empresa apresenta, surge a necessidade de um método centralizado para este tipo de análise, de maneira que os gestores consigam filtrar as informações rápida e intuitivamente, gerando analises concisas e eficientes, além de se desenvolver numa linguagem acessível, direta e visual, para que o público não habituado com ferramentas de *Business Intelligence*, consiga utilizar esse tipo de informação.

Por se tratar de uma grande multinacional do segmento de saúde, com vasto portifólio de produtos, a empresa está inserida em um ambiente empresarial altamente competitivo e detém a árdua missão de disputar espaço com diferentes concorrentes, sejam eles locais e regionais, ou internacionais e multinacionais. Entender onde está inserida e ter indicadores que mensurem seu desempenho, além de demonstrarem qual sua colocação ante as demais, são

essenciais para que a empresa se desenvolva e consiga atingir uma parcela ainda maior do mercado, possibilitando assim, a maximização de seus lucros. Com base nisso, reitera-se a necessidade de implementar uma ferramenta acessível e de rápida visualização para que os inúmeros dados sejam lidos de maneira precisa e fomentem o desenvolvimento de uma forte estratégia para a companhia.

Wilbur e Farris (2014) apontam que a análise da relação entre distribuição e *Market Share* é um tópico muito importante no contexto das atividades de *marketing* desempenhadas por empresas de consumo. Para Canuto *et al*. (2000), uma alteração nos preços relativos pode gerar distorções na sua quota de mercado. Em virtude disso, identifica-se a necessidade de um forte acompanhamento aos fatores que impactam diretamente a variação desse indicador de posicionamento no setor de atual.

É importante salientar que apesar de pesquisas anteriores reconhecerem a existência de diversos fatores que influenciam as vendas e participação de mercado de um empresa com seus produtos, como o nível de investimento em propaganda, nível de preço, promoção e estratégias de produto (ATAMAN *et al*., 2010; SRINIVASAN *et al*., 2000), a distribuição tem sido apontada como um dos principais fatores para explicar as variações no *Market Share* dos produtos das empresas em diversos mercados (ATAMAN *et al*., 2010; NIJS, MISTRA, ANDERSON, HANSEN, & KRISHNAMURTHI, 2010).

A partir das informações acima descritas e buscando estabelecer uma maneira concisa para analisar estrategicamente alguns dos principais aspectos que contribuem para o impacto de seus lucros, surge a necessidade de identificar mecanismos ou métodos capazes de minerar, ou seja, encontrar anomalias, padrões ou correlações em grandes conjuntos de dados e apresentar informações. A preocupação central desta pesquisa está, portanto, na exploração de um método baseado em técnicas de mineração de dados para identificar, projetar e apresentar os dados de maneira direcionada e interativa. Com isso, ao vislumbrar a relação entre estes fatores nos diferentes períodos de tempo, os gestores detêm maior embasamento na tomada de decisões para o gerenciamento de vendas, precificação, expansão ou retração da marca e outros.

Por meio dessas análises, a companhia pode se debruçar sobre seu histórico e entender como vem se desenvolvendo o comportamento de sua marca ante aos concorrentes, varejistas e público. O método almeja ser útil em estratégias de curto e longo prazo, uma vez que consegue mostrar o desempenho passado, trazer insumos quanto a predição de resultados financeiros e, ainda, analisar questões de mercado.

Portanto, resumidamente, o projeto consiste em implementar, com base no modelo CRISP-DM (Processo Padrão Interindústrias para Mineração de Dados), uma maneira de minerar e relacionar dados disponibilizados. Ainda, em conjunto a metodologia escolhida, busca-se, com base em uma técnica de previsão, apresentar um modelo preditivo, no qual os gestores consigam analisar o possível comportamento futuro de suas categorias, possibilitando evoluções quanto as estratégias. Por fim, o trabalho busca apresentar o modelo por meio de um painel gerencial interativo, que agrupe as informações e as apresente de maneira visual e direcionada.

1.2 JUSTIFICATIVA

Já não é segredo que os dados vêm crescendo de maneira exponencial com o advento da tecnologia, automação, internet etc. De acordo com o Harvard Business Review, a partir de 2012, cerca de 2,5 *exabytes* de dados passaram a ser criados a cada dia — número que está dobrando a cada 40 meses ou mais. Atualmente, mais dados cruzam a internet a cada segundo do que eram armazenados em toda a internet há 20 anos (EXAME, 2021). Com esse nível imenso de informações, as empresas detêm a oportunidade de trabalhar com informações cada vez mais precisas e em uma granularidade cada vez maior, podendo identificar tendências ou características mais detalhadas. Com base nessa gigantesca quantidade de dados, se gerenciados de forma apropriada, pode-se fornecer às empresas vantagens competitivas sustentáveis e, ainda, a geração de valor para os stakeholders (JANSSEN; VOORT; WAHYUDI, 2016).

Segundo Dino (2018), 44% das empresas fizeram investimentos em sistemas de gestão empresarial, seja por nova compra ou melhorias. No Brasil, a previsão do investimento de US$ 965 milhões para 2018, apenas com *Business Intelligence, Analytics e Big Data*. A reportagem relata que, em 2018, a alta foi de 4,5% sobre o ano de 2017, tendo as companhias globais investindo US$ 3,7 trilhões em Tecnologia da Informação. Isso corrobora a necessidade de as empresas estarem cada vez mais antenadas quanto a coleta e análise de dados, a fim de garantir forte competitividade ante as demais.

O principal benefício do BI é sua capacidade de fornecer informações precisas quando necessário, incluindo uma visão do desempenho corporativo geral e de suas partes individuais (TURBAN *et al*., 2009), provendo aos gerentes e analistas a capacidade de realizar análises apropriadas e executar ações (WIXOM *et al.*, 2011).

Dado o contexto no qual a empresa em análise está inserida e a forte evolução das empresas em relação a análise de dados, o presente trabalho tem como objetivo responder o

seguinte problema de pesquisa: “Como extrair, de uma vasta base de dados, análises que contribuam para um bom direcionamento estratégico dos gestores da empresa?”

Assim, o trabalho busca apresentar relatórios e indicadores capazes de apresentar de maneira clara o desempenho da companhia, além de desenvolver previsões com base em seu histórico. Entende-se que algumas pessoas da companhia ainda não têm total familiaridade com ferramentas mais robustas de análise de dados, trazendo, dessa forma, a oportunidade de desenvolver o estudo pautado na acessibilidade, garantindo que os resultados possam ser lidos pelo maior número de pessoas, sem prejuízos quanto a sua interpretação. Reitera-se que em relação a viabilidade técnica e econômica, o projeto se baseou em utilizar insumos e ferramentas já utilizados dentro da companhia, usando o que já possui e gerando valor sobre isso.

Por fim, ressalta-se a importância do estudo para a empresa, visto que apesar de possuir algumas ferramentas capazes de apresentar seu desempenho, não possui um instrumento que possa prever resultados e assim facilitar algumas análises.

1.3 OBJETIVOS

A fim de buscar maneiras de extrair, a partir de uma vasta base de dados, análises que contribuam para um bom direcionamento estratégico dos gestores da empresa, o capítulo apresenta o objetivo geral e os objetivos específicos.

1.3.1 Objetivo geral

Estabelecer um painel gerencial que unifique os principais indicadores financeiros e comerciais para as vinte principais categorias de produtos de uma multinacional no setor da saúde.

1.3.2 Objetivos específicos

- Desenvolver, com base no modelo CRISP-DM, todo o processo de mineração de dados de vendas da empresa no canal farmacêutico.
- Definir os atributos de previsão para o modelo definido.
- Desenvolver os cálculos e modelagens para os indicadores financeiros e comerciais.

- Desenvolver um painel gerencial para melhor visualização dos resultados.
- Analisar e validar os resultados obtidos.

1.4 ADERÊNCIA DO TRABALHO A ENGENHARIA DE PRODUÇÃO

O presente trabalho apresenta uma metodologia capaz de transformar os dados da empresa em conhecimento e informações de gerenciamento, por meio de técnicas de indução para propor hipóteses e solucionar questões empresariais. A metodologia tem como objetivo propor, a partir de uma sequência de etapas, um modelo consistente que apresente os resultados baseados em dados.

A precisão, confiabilidade e análise de dados para fundamentar decisões, carece de ferramentas e métodos eficazes, dado seu fator fundamental para que as informações sejam atribuídas de maneira correta, mitigando possíveis riscos quanto a falhas após a tomada de decisões.

A interpretação de dados, sejam quantitativos ou qualitativos, fomentam o bom direcionamento de qualquer setor de uma instituição, eliminando desperdícios, auxiliando no desdobramento da estratégia, direcionando esforços para o que realmente trará resultados e entre tantos outros pontos que contribuem para o processo de melhoria contínua.

Como consequência disso, é notável a correlação que a análise de dados tem com as mais variadas áreas de atuação. Reitera-se, dessa maneira, que para o desenvolvimento desse trabalho, múltiplas áreas da Engenharia de Produção estiveram envolvidas, com destaque para Gestão do Desempenho Organizacional, Gestão da Informação, Gestão Estratégica e Organizacional e Gestão do Conhecimento. Nota-se uma abordagem mais evidenciada quanto à Gestão do Desempenho Organizacional e a Gestão da Informação, uma vez que focaliza seus esforços na implementação de um modelo que apresente indicadores robustos, baseados em dados, quanto ao desempenho, buscando fomentar maior eficiência na tomada de decisões que impactam diretamente na receita da empresa. A Gestão do Desempenho Organizacional e a Gestão da Informação compõe subáreas da Engenharia Organizacional, proposta pela ABEPRO, tópico no qual o trabalho apresentará suas principais contribuições.

1.5 DELIMITAÇÃO

O presente trabalho representa uma situação específica de implementação de um modelo de análise e previsão de faturamento para uma empresa do ramo de saúde, e não se

almeja padronizar o processo para outras empresas. A partir disso, ressalta-se que o modelo foi criado para atender as necessidades específicas da empresa em estudo.

Além disso, a pesquisa limita-se em estruturar um modelo que atenda as expectativas da gestão de uma das áreas da empresa e em apresentar o piloto do que poderia ser o modelo a ser implementado, não chegando à implementação de fato.

1.6 ESTRUTURA DO TRABALHO

O presente trabalho está estruturado em cinco capítulos, que buscam explicitar, de maneira detalhada, o desenvolvimento de todo o projeto. Uma vez apresentado o primeiro capítulo, de introdução, os demais seguem a seguinte estrutura:

1. Capítulo 2 – Fundamentação Teórica: discorre sobre os conceitos de Dados, Mineração de Dados, Processo CRISP-DM e *Business Analytics* encontrados na literatura.
2. Capítulo 3 – Metodologia: apresenta-se o enquadramento metodológico do presente estudo, bem como a descrição da metodologia e ferramentas utilizadas na execução do projeto.
3. Capítulo 4 – Desenvolvimento: estrutura-se o processo para implementação de um modelo piloto, além de analisa os resultados, a fim de validar os resultados e propor melhorias.
4. Capítulo 5 – Conclusão: finalmente, são apresentadas as considerações finais do estudo, descrevendo as dificuldades e desafios encontrados e recomendações para trabalhos futuros.

2 FUNDAMENTAÇÃO TEÓRICA

Este capítulo aborda o conteúdo base para elucidação e elaboração do presente projeto. Dividido em três macro segmentações, este se apoia em diferentes referências para seu respectivo setor. A primeira macro área aborda os conceitos chave que nortearam o projeto e que delimitaram o objeto de estudo, abordando os temas priorizados para análise dos impactos na receita. A segunda, fundamenta o processo de mineração de dados, apresentando algumas das técnicas já existentes, bem como o método de análise envolvido em diversos projetos de mineração de dados. Ademais, a terceira, edifica a base quanto a técnica de redes neurais artificiais e sua aplicação em diferentes tipos de análises.

2.1 DADOS

Para que seja estabelecido um método coerente a ser adotado, é imprescindível compreender os tipos de dados a serem analisados. Dito isso, os dados podem ser categorizados de duas maneiras: qualitativos (tem por objetivo descrever e não apenas medir, permitem compreender a complexidade e detalhes das informações), e quantitativos (tem por objetivo coletar fatos concretos, apresentam os números que comprovam os objetivos gerais).

Em linhas gerais, os dados podem ser coletados interna ou externamente à organização. Contudo reitera-se a necessidade de avaliação quanto a sua confiabilidade, uma vez que dados externos não são inicialmente tratados e nem auditados pela organização, além de acarretarem em um custo de obtenção muitas vezes maior que os dados internos. A partir da coleta, seja ela externa ou interna, são estabelecidos grandes conglomerados de dados, denominados base de dados.

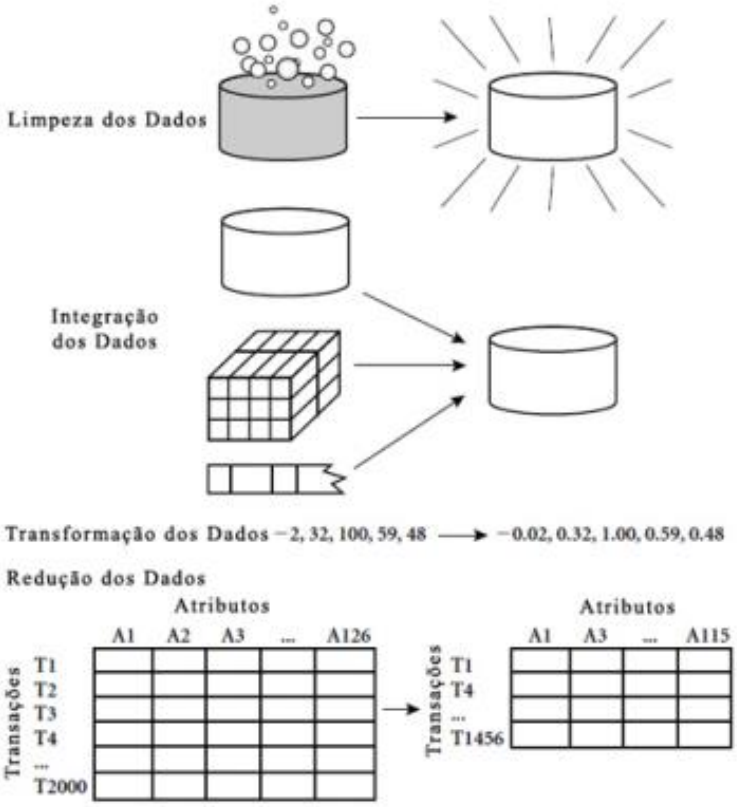
Usualmente, os dados são coletados com o propósito de suportar os processos de negócios operacionais, não tendo, em muitos casos, destinação específica para análises de mineração e exploratórias. Consequentemente, existe a possibilidade de que os dados coletados detenham baixa qualidade ou sejam enviesados, por exemplo, reduzindo sua aplicabilidade. Com base nisso, a utilização de técnicas avançadas de mineração tem por objetivo desenvolver um modelo válido mesmo que existam dados considerados “sujos” (HAND, 1998).

Segundo Adriaans e Zantinge (1996), a preparação de dados é uma etapa que representa cerca de 60% do esforço aplicado em um projeto de mineração de dados. A fase tem por objetivo preparar os dados disponibilizados, uma vez que podem ter inconformidades, não

estar dispostos no formato adequado para aplicação dos algoritmos de descoberta ou apresente algo que possa comprometer a qualidade dos resultados.

O processo que vai eliminando essas inconformidades, representado na Figura 1, é denominado preparação dos dados que, segundo Han *et al*. (2011), consiste principalmente em:

Figura 1 - Etapas de preparação de dados



Fonte: Han et al. (2011)

2.1.1 Limpeza dos dados

Bancos de dados reais detém uma grande possibilidade de apresentarem inconsistências. Com base nisso, as etapas de limpeza de dados objetivam eliminar incongruências como registros incompletos, valores errados ou dados inconsistentes, tratar dados ruidosos e encontrar valores discrepantes (HAN; KAMBER; PEI, 2011).

Para o completo desenvolvimento desta etapa, as técnicas permeiam entre a remoção de registros com problemas, atribuição de valores padronizados e aplicação de técnicas de agrupamento para fomentar a descoberta dos melhores valores. Ainda, Han *et al*. (2011) propõe

a utilização de um processo específico para a limpeza de dados, dada a magnitude do esforço empregado na etapa.

Para Han, Kamber e Pei (2011), dentre os métodos mais comuns para preencher dados ausentes são:

- Ignorar o registro;
- Imputar manualmente um valor de maneira empírica, respeitando o domínio de cada atributo;
- Usar uma constante global para preenchimento de valores ausentes, respeitando o domínio de cada atributo;
- Encontrar um registro com valor observado mais similar em relação aos demais atributos, inserindo o valor correspondente;
- Ordenando a base conforme um ou mais atributos e adotando o valor do registro imediatamente anterior para inserir no lugar do dado faltante;
- Usar a média para dados quantitativos e a moda para dados qualitativos.

2.1.2 Integração dos dados

Os dados podem ter fontes diversas (bancos de dados, arquivos, planilhas, entre outras), resultando na necessidade de estabelecer-se a integração dos dados de modo que haja um único repositório capaz de comporta-los. Com base nisso, é importante que haja uma análise de modo que possíveis duplicidades, redundâncias e demais fatores conflitantes sejam unificados ou excluídos.

Ferrari e Castro (2017) ressaltam três pontos a serem considerados em casos de união de bases:

- Duplicidade: ocasião na qual entidades muito semelhantes estão definidas em bases distintas com nomes e atributos diferentes, necessitando relacioná-las;
- Conflitos: quando valores de dados diferentes são apresentados em bases distintas como, por exemplo, unidades de medida diferentes para um mesmo dado.
- Redundância: quando um objeto ou atributo derivar de um ou mais objetos ou atributos da base.

2.1.3 Transformação dos dados

Dentro da mineração de dados, existem algoritmos que trabalham com diferentes especificidades, podendo trabalhar apenas com dados numéricos ou apenas com dados categóricos, sendo diversas técnicas de mineração limitadas a trabalhar com um ou outro. Ao passo que isso aconteça, faz-se necessária a transformação dos valores numéricos em categóricos ou vice-versa (CAMILO; SILVA, 2009).

A partir disso, surge a necessidade de se estabelecer uma unicidade quanto a formatação dos dados, de modo que todos sejam devidamente processados. Para que essa unicidade se estabeleça, existem variados critérios para transformação de dados, dentre elas: normalização, generalização, agrupamento, suavização e a criação de novos atributos.

2.1.4 Redução de dados

Usualmente, a quantidade de dados para mineração é alta. Em alguns casos é tão alta que necessitaria de um tempo de processamento muito alto, *softwares* extremamente potentes ou, ainda, podem tornam-se impraticáveis. Dessa maneira, adotando técnicas de redução de dados, um grande volume de dados pode ser reduzido. Contudo, a integridade dos dados deve ser mantida, de modo que a mineração com os dados reduzidos possa, ainda assim, ser mais eficiente e capaz de reduzir os mesmos resultados analíticos. (HAN; KAMBER; PEI, 2011).

Ferrari e Castro (2017), destacam alguns dos métodos de redução, sendo estes:

- Seleção de atributos: remoção de atributos desnecessários;
- Compreensão de atributos: diminuição da dimensionalidade através da transformação de dados ou codificação;
- Redução do número de dados: dados são removidos, substituídos ou estimados por uma representação simples;
- Discretização: aumento do intervalo entre os dados, objetivando diminuir sua quantidade.

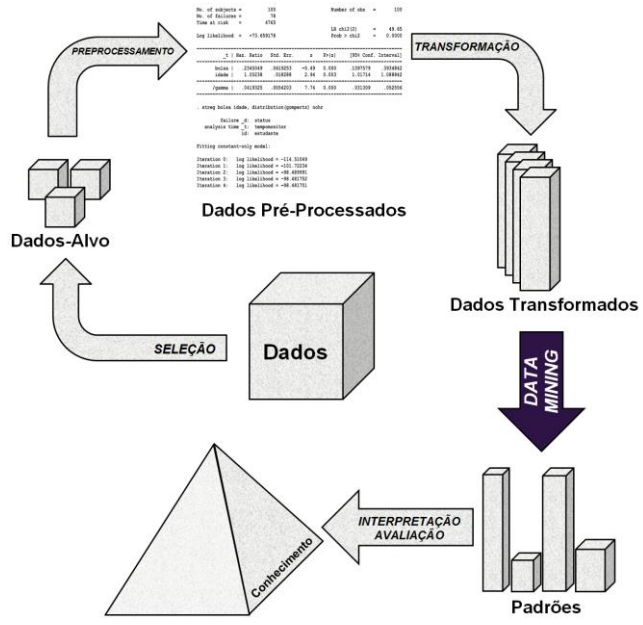
2.2 MINERAÇÃO DE DADOS

A mineração de dados tem suas raízes no método criado por J. W. Tukey nos anos 70 e 80, referenciada como Análise Exploratória de Dados (HOAGLIN *et al*., 1982). A partir disto, com os avanços tecnológicos, no armazenamento de dados, redução em custos de comunicação eletrônica, tecnologias de processamento cada vez mais desenvolvidas, avanços em técnicas de analise, melhorias na arquitetura cliente-servidor e com o advento dos repositórios de dados, bem como pelo aumento nas necessidades de análises rápidas em ambientes cada vez mais competitivos são, segundo Peacock (1998a), algumas das forças que pressionaram o avanço desta tecnologia.

Quanto a sua definição, existem variadas interpretações. Não existe um consenso quando a definição dos termos KDD (*Knowledge Discovery in Databases* ou Descoberta de Conhecimento nas Bases de Dados) e *Data Mining.* Alguns autores como Rezende (2005) e Han *et al*. (2011), consideram como sinônimos. Já Cios *et al*. (2007) e Fayyad *et al*. (1996), interpretam o KDD como todo o processo de descoberta de conhecimento, e veem mineração de dados parte do processo. Segundo Fayyad *et al* (1996), o KDD trata-se de uma tentativa de solucionar o problema causado pela chamada "era da informação": a sobrecarga de dados. Contudo, é de consenso que o processo de mineração deve ser iterativo, interativo e dividido em fases.

A Figura 2 apresenta a representação dos estágios do KDD.

Figura 2 - Estágios do Processo KDD e *Data Mining*



Fonte: Fayyad, Piatetsky-Shapiro e Smyth (1996)

A Mineração de Dados, ou *Data Mining* em inglês, fundamenta-se na produção de conhecimento a partir de dados acumulados. O processo permite estruturar o conhecimento presente em grandes volumes de dados corporativos, adotando tecnologias de banco de dados, reconhecimento de padrões, aprendizado automático, estatística e outras (HAND, 1998). Dentre as variadas técnicas de mineração de dados, destacam-se: redes neurais, árvores de decisão, métodos de indução de regras, métodos para análise de cesta de compras, regras de associação, técnicas de segmentação e dedução baseada em memória.

Ainda, por ser considerada multidisciplinar, suas definições tem variações significativas quando comparadas as interpretações de autores com campos de atuação distintos. Consideradas como fortes expressões no que diz respeito a mineração de dados, destacam-se as áreas da estatística, aprendizado de máquina e banco de dados. Zhou apresenta uma análise comparativa entre as três principais vertentes.

• Perspectiva estatística: "Mineração de Dados é a análise de grandes conjuntos de dados a fim de encontrar relacionamentos inesperados e de resumir os dados de uma forma que eles sejam tanto úteis quanto compreensíveis ao dono dos dados" (HAND *et al*., 2001).

• Perspectiva aprendizado de máquina: "Mineração de Dados é um passo no processo de Descoberta de Conhecimento que consiste na realização da análise dos dados e na aplicação de algoritmos de descoberta que, sob certas limitações computacionais, produzem um conjunto de padrões de certos dados" (FAYYAD *et al*., 1996).

• Perspectiva banco de dados: "Mineração de Dados é um campo interdisciplinar que junta técnicas de máquinas de conhecimentos, reconhecimento de padrões, estatísticas, banco de dados e visualização, para conseguir extrair informações de grandes bases de dados" (CABENA *et al*., 1998).

Assim sendo, no que tange os processos que definem e padronizam as fases e atividades da mineração de dados, apesar de diversos e com diferentes particularidades, todos, de maneira geral, apresentam uma estrutura semelhante.

2.2.1 Tarefas da Mineração de Dados

Em virtude de sua adaptabilidade e variadas possibilidades, a Mineração de Dados permite ao usuário uma grande gama de possibilidades de análise, dentre as mais comuns, segundo Larose (2005), estão:

- Descrição

- Classificação
- Predição
- Agrupamento
- Associação

2.2.1.1 Descrição

Para Larose (2005), a tarefa é utilizada para descrever os padrões e tendências dos dados, a descrição oferece possíveis interpretações dos resultados obtidos a partir da mineração. A tarefa é comumente utilizada em conjunto com técnicas de análise de dados para corroborar hipóteses como a influência de determinadas variáveis no resultado obtido.

Ainda, de acordo com Larose (2005), os modelos de previsão devem ser os mais claros e transparentes possíveis, apresentando resultados que detenham interpretações e explicações. Com a descrição as análises se tornam mais completas, garantindo o bom entendimento dos resultados. Dessa maneira, A descrição de alta qualidade, pode ser realizada com dados exploratórios, apresentando um método gráfico de explorar os dados em busca de padrões e tendências, por exemplo.

2.2.1.2 Classificação

A classificação é a tarefa utilizada para a categorização dos dados, com base em um conjunto de classes previamente definidas. Aqui, analisa-se um conjunto de registros, já categorizados de acordo com alguma especificação, visando aprender a como classificar novos registros (LAROSE, 2005).

Para Mattar (1998), a análise permite comparar dois ou mais grupos, a fim de determinar se diferem uns dos outros, bem como a natureza da diferença. A partir disto, e com base em um conjunto de variáveis independentes, é possível classificar os objetos em análise em duas ou mais categorias mutuamente exclusivas.

Em outras palavras, a classificação é uma tarefa que, por meio do aprendizado supervisionado, baseia-se em dados históricos para desenvolver a classificação de novos dados em agrupamentos específicos.

2.2.1.3 Predição

Similar as tarefas de classificação e estimação, a tarefa de predição busca descobrir o valor futuro de um determinado atributo da base de dados. Como exemplo:

• Predizer o faturamento de uma empresa três meses adiante;

• Predizer o percentual que será aumentado de trânsito na cidade caso ocorra uma venda expressiva de carros;

• Predizer a quantidade de produtos vendidos com base no histórico da empresa.

A partir disto, com base nas devidas considerações, alguns métodos de classificação e regressão podem ser adotados para predição.

As predições numéricas são métodos buscam descobrir um possível valor futuro para uma variável contínua. Em relação a predição de variáveis discretas, as técnicas de classificação podem ser aplicadas. Os métodos mais conhecidos para predição numérica são as regressões. Segundo Maroco (2007), as análises de regressão caracterizam-se pela definição de um vasto conjunto de técnicas estatísticas adotadas para modelar relações entre as variáveis e predizer o valor de uma ou mais variáveis dependentes (denominadas resposta) a partir de um conjunto de variáveis independentes (denominadas preditoras), sendo as preditoras os atributos dos registros e as de resposta o que se busca predizer.

2.2.1.4 Agrupamento

O agrupamento, também conhecido como clusterização, é utilizado para categorizar os dados com base em classes (ou clusters), baseando-se nas similaridades destes. Essa tarefa permite identificar similaridades entre os dados, gerando informações mais cada vez mais robustas. Salienta-se que, quando comparados entre os dados do mesmo grupo, os dados possuem similaridades, mas quando comparados a outros clusters individualmente, apresentam diferenças significativas (LAROSE, 2005).

Segundo Pereira (1999), o procedimento de clusterização inicia com o cálculo das distâncias entre os objetos em análise, dentro de um ambiente multiplano, tendo como eixos, todas as medidas realizadas. Por fim, são efetuados os agrupamentos por proximidade geométrica, permitindo a identificação dos grupos dentro do universo dos dados analisados.

Diferente da classificação, o agrupamento não tem como objetivo classificar, estimar ou predizer dados, seu único objetivo é apenas identificar e agrupar os dados de acordo com

determinadas especificações, não necessitando, dessa maneira, que os registros estejam previamente categorizados.

Exemplo disso, é a origem dos dados, como: dados enviados de farmácias e dados enviados de supermercados, que seriam agrupados como Canal Farmácia e Canal Alimentar.

2.2.1.5 Associação

As regras de associação são usadas para identificar as dependências e a correlação entre os atributos dentro de um grande conjunto de dados. Estas regras provaram ser muito úteis no campo de marketing e varejo, bem como muitos outros diversos campos. Geralmente uma regra de associação caracteriza-se como "Se X antecede, então Y é consequência", tendo uma medida do suporte e confiança associada à esta. (OUYANG, 2012). O suporte é a fração das transações totais que contém os itens analisados, ou seja, caso sejam analisados três elementos, o suporte será a porcentagem de transações que incluam tanto A, B e C (LAROSE, 2005).

Para Han e Kamber (2011), na etapa inicial, encontram-se os conjuntos de itens que ocorrem com frequência e, na próxima etapa, cria-se regras de associação para esses itens. Esta tarefa encontra relações de associação ou correlações interessantes entre um grande conjunto de dados de itens.

Tem sido reconhecido que os algoritmos convencionais para regras de associação de mineração só podem minerar os conjuntos de dados que possuem atributos binários. Dessa maneira, atributos quantitativos devem ser tratados corretamente, da mesma forma que os atributos booleanos, em o campo de mineração de dados (ITAN; YENTY, 2008).

2.2.2 Principais Técnicas da Mineração de Dados

Na literatura existem diversas formas de classificar as técnicas de previsão. Para o presente trabalho utiliza-se a classificação adota por Han *et al*. (2006), na qual são agrupados de acordo com a tarefa executada.

2.2.2.1 K-means (Agrupamento)

A técnica visa examinar os elementos totais de uma série de dados e associá-los a um centroide, selecionando aleatoriamente k registros, em que cada um representa um agrupamento. Para os registros remanescentes, utiliza-se de uma função capaz de medir a

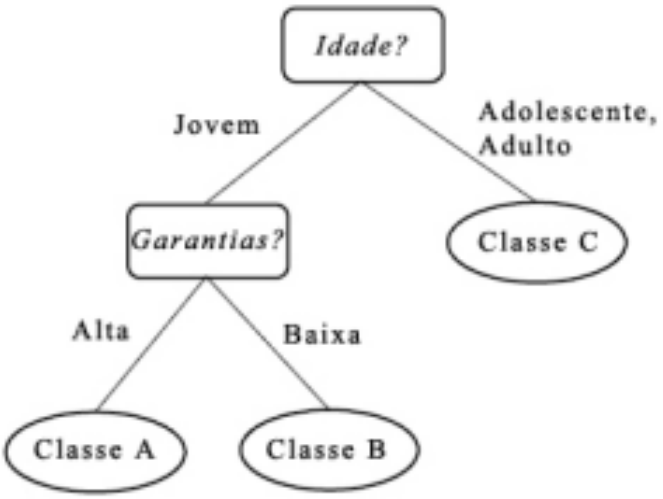
distância entre os centroides, a fim de identificar a similaridade entre eles. Ressalta-se que quanto maior a similaridade, menor é a distância entre os pontos e ainda, que o centro do *cluster* é recalculado a cada novo objeto inserido (HAN; KAMBER, 2011).

2.2.2.2 Árvores de decisão (Classificação)

A técnica de classificação da Árvore de Decisão pode ser feita em etapas seriais ou paralelas de acordo com a quantidade de dados, eficiência do algoritmo e memória disponível. Uma árvore serial é um modelo lógico, como uma árvore binária construída usando um conjunto de dados de treinamento. Ajuda a prever o valor de uma variável de destino, fazendo uso de variáveis de previsão (RAHPEYMAI, 2002).

Consiste em conjuntos de regras organizados hierarquicamente. É uma estrutura recursiva simples para representar um procedimento de decisão no qual uma instância futura é classificada em classes predefinidas presentes e tenta dividir observações em subgrupos mutuamente exclusivos. Cada parte de uma árvore corresponde a um ou mais registros do conjunto de dados original (RAHPEYMAI, 2002). Os nós superiores são nomeados como o nó raiz (sem *link* de entrada) e representam todas as linhas no conjunto de dados fornecido. Os outros nós são nomeados como nós internos ou de decisão (apenas um *link* de entrada) usado para teste em um atributo. Os nós mais inativos são nomeados como nós terminais (sem *link* de saída) e denotam uma classe de decisão, como mostrado na Figura 3.

Figura 3 - Árvore de decisão



Fonte: Han; Kamber (2006)

2.2.2.3 Regressões (Predição)

De acordo com Draper e Smith (1998), os modelos de regressão são classificados como lineares, linearizáveis e não lineares. Sendo as regressões lineares, aquelas que são lineares em relação a algum parâmetro, de modo que a derivada do modelo em relação aos parâmetros não dependa dos parâmetros em si, ou seja, a relação entre as variáveis preditoras e a resposta segue um comportamento linear.

$$\frac{\partial}{\partial \theta_j} f_i(X, \theta) = g(X)$$

Os modelos linearizáveis que, como base em transformações, tornam-se lineares.

$$Y = \theta^x \varepsilon$$

Por fim, os modelos não lineares são os que pelo menos uma das derivadas parciais dependa de algum parâmetro do modelo, ou seja, a relação entre as variáveis preditoras e a resposta não segue um comportamento linear.

$$Y = \theta_0 + \theta_1^x + \varepsilon$$

Ressalta-se que, para Corrêa *et al*. (2010), a regressão linear e não linear tem sido o mecanismo que possibilita aos pesquisadores a verificação de problemas em diferentes áreas. Destacando a utilização destas modelagens estatísticas na aplicação e utilização das novas tecnologias computacionais.

2.2.2.4 Mineração de Itens Frequentes (Associação)

A Mineração de Itens Frequentes é a técnica de extrair qualquer conjunto de itens frequente existente (com uma frequência de ocorrência não inferior a algum limite) em dados. Essa técnica foi proposta no início da década de 1990 para descobrir itens que ocorrem frequentemente na análise de cestas de mercado (AGRAWAL *et al*., 1993), sendo inicialmente chamada de mineração de grandes conjuntos de itens.

Considerada a raiz do campo de mineração de padrões, que engloba múltiplas técnicas que visam extrair conjuntos de itens em vários formulários e para vários propósitos (AGGARWAL; HAN, 2014). A variação mais simples da técnica é a extração de conjuntos de itens que não aparecem com frequência nos dados ou aqueles que foram descartados pela técnica. Isso deu origem à definição de mineração de padrões como o problema de mineração de conjuntos de itens que, frequentemente (ou raramente), aparecem nos dados (KOH; RAVANA, 2016).

Para Agrawal *et al*. (1993), a técnica pode ser dividida em duas etapas:

- Cria-se um conjunto de itens frequentes, respeitando um valor mínimo para tal;
- Geram-se as regras de associação do conjunto.

A fim de garantir que os resultados sejam válidos, introduz-se os conceitos de suporte e confiança. A medida de suporte indica o percentual de registros (com base em todo o conjunto de dados) que se encaixam na regra definida, enquanto a confiança mede o percentual de registros que atendem especificamente a regra.

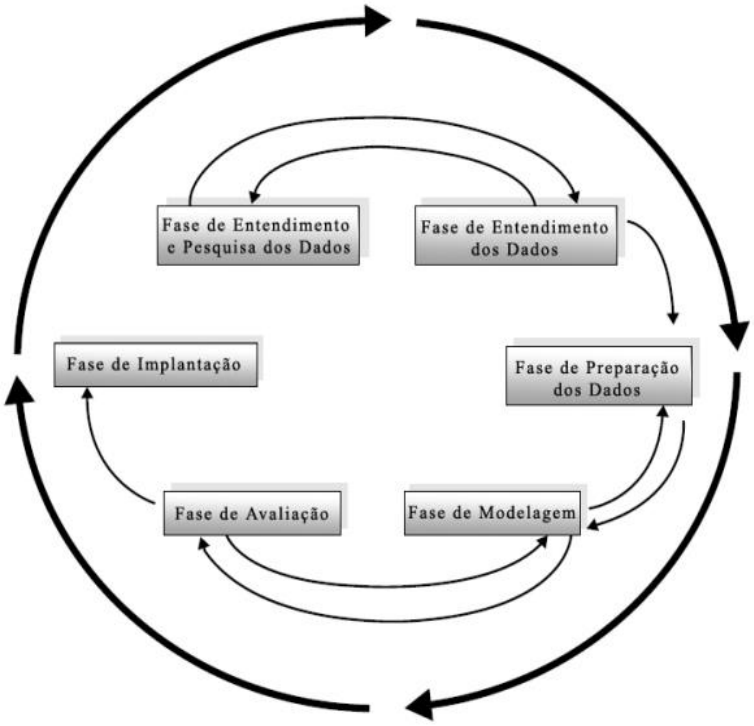
2.3 PROCESSO CRISP-DM

De acordo com Uber (2004), o CRISP-DM (*Cross-Industry Standard Process for Data Mining*) teve sua concepção em 1996, visando possibilitar a uniformização de técnicas e conceitos para auxiliar a busca por conhecimentos específicos, facilitando a tomada de decisão.

Baseada no processo de KDD, a metodologia surge com o intuito de criar processos que padronizassem o desenvolvimento de projetos de mineração de dados. Chapman *et al*. (2000) define que o processo de mineração é cíclico e este ciclo está dividido em seis fases: Entendimento do Negócio, Entendimento dos Dados, Preparação dos Dados, Modelagem, Avaliação e Disponibilização. Ressalta-se que suas fases não são necessariamente executadas em sequência, dada a possibilidade de durante o processo ocorrerem alterações ou modificação quanto ao sequenciamento. Segundo Chapman *et al*. (2000), o próprio resultado de cada fase poderá determinar a próxima fase a ser executada.

O esquema apresentado na Figura 4 ilustra o sequenciamento das fases, indicando suas respectivas dependências mais comuns e importantes, indicando, também, as possíveis alterações do sequenciamento. Ainda, a seta circular externa indica o ciclo natural do processo de mineração de dados.

Figura 4 - Ciclo das fases da metodologia CRISP-DM



Fonte: Larose (2005)

2.3.1 Entendimento do Negócio ou Domínio

A primeira etapa de CRISP-DM é, possivelmente, a etapa mais importante, haja vista que ela delimitará todo o processo. A etapa caracteriza-se por definir todo o objetivo do projeto, por identificar as necessidades da empresa e, ainda, por traçar o plano inicial que norteará o projeto como um todo. Para Azevedo e Santos (2008), a fase dedica-se à compreensão dos requisitos do projeto, sob uma ótica de negócios que, a partir do aprendizado, possibilita definir um problema de mineração. Dessa maneira, caso essa etapa não seja realizada da maneira correta, o projeto pode ser completamente invalidado.

2.3.2 Entendimento dos dados

Inicializada pela coleta de dados, essa etapa consiste em identificar os dados a serem utilizados, bem como, descrever quais as fontes de dados, a maneira como foram coletados, identificação de grupos peculiares e, ainda, identificar alguma dificuldade para suas atualizações, a fim de desenvolver hipóteses sobre informações ocultas (Chapman et al., 2000). Ainda, de acordo com Shearer (2000), esta etapa é essencial para evitar problemas inesperados durante a fase seguinte (preparação de dados), que é, usualmente, a mais longa de um projeto.

2.3.3 Preparação dos dados

Bem como apresentada de maneira detalhada na sessão anterior, a preparação de dados consiste em atividades ligadas ao tratamento dos dados, podendo incluir a seleção do que será de fato usado como *dataset*, limpeza e transformação caso seja necessário (Chapman *et al*., 2000).

A preparação de dados é uma das etapas de maior importância que, salvo algumas exceções, exige a maior parte do tempo da mineração de dados, estimando-se cerca de 60% do tempo e esforço de um projeto. Com isso, dedicar esforços nas etapas iniciais (Entendimento do Negócio ou Domínio e Entendimento dos Dados), pode reduzir a sobrecarga a ela relacionada, mas ainda se fará necessário grande esforço para a preparação dos dados a serem utilizados (IBM, 2016).

2.3.4 Modelagem

A modelagem é etapa na qual o modelo começa a ganhar forma e os primeiros resultados podem ser explorados. É definida de acordo com as necessidades do negócio e com a tipologia das variáveis. Com isso, a fase pode ser dividida pelos marcos: seleção dos algoritmos, geração do projeto teste, aplicação dos algoritmos e avaliação do modelo gerado. De acordo com a necessidade, variadas são as possibilidades quanto as técnicas e ferramentas que poder ser utilizadas. A partir de diversas iterações, o modelo final é definido, uma vez que as configurações podem mudar, bem como uma série de ajustes pode se fazer necessária. Ainda, é comum que seja necessário retornar para a etapa de preparação com a necessidade de manipulações nos dados para atender o modelo proposto (IBM, 2016).

2.3.5 Avaliação

Segundo Azevedo e Santos (2008), na fase de avaliação os resultados obtidos são analisados com mais detalhes, quando especialistas do negócio são necessários. Assim, tendo o modelo, é possível avaliar se os resultados correspondem as expectativas do projeto. Esse processo se dá por três tarefas: avaliação dos resultados, revisão do projeto e determinação de próximos passos. Caso em alguma dessas etapas ocorra alguma inconformidade ou exista espaço para melhorias, os esforços devem ser direcionados para as mudanças necessárias.

2.3.6 Implementação

Após obtenção de êxito no desenvolvimento da avaliação, a implementação é a etapa final. Nessa etapa, o modelo é colocado em produção, agregando valor para os usuários, de modo a ser factível, caracterizado como um modelo para obtenção de conhecimento preciso, em que além de ser aderente às necessidades da organização, seja interpretável e com capacidade operacional. De maneira resumida, o conhecimento deverá ser apresentado para as partes interessadas de maneira que elas compreendam as informações (Chapman *et al*, 2000).

2.4 BUSINESS INTELLIGENCE

Para Wixom e Watson (2010), o termo *"Business Intelligence"* pode ser entendido como uma ampla categoria de tecnologias, aplicativos e processos, para coletar, armazenar, acessar e analisar dados, visando contribuir para uma melhor tomada de decisão por seus usuários. Silva (2011), acrescenta que BI consiste em transformar metódica e consciente dos dados das mais variadas fontes de dados, sejam estruturados e não estruturados, em novas formas de proporcionar informação e conhecimento dirigidos aos negócios. Ainda, afirma que as informações disponibilizadas são orientadas aos resultados.

De acordo com Reginato e Nascimento (2007), a ferramenta de gestão, BI, consiste em: armazenamento de dados (*Data Marts* e *Data Warehouse*), na análise de informações (*Online Analytical Processing* – OLAP) e na mineração de dados (*Data Mining*).

Segundo Nagar *et al*. (2016), as ferramentas de BI são um mecanismo ou meio de implementar a ideia de inteligência de negócios em um determinado conjunto de dados, de modo que sejam apresentados os resultados visualmente.

Atualmente, diversas organizações que fazem uso das ferramentas de BI, vêm ganhando mais valor pela grande variedade de informações em todos os níveis, maximizando o uso de seus ativos, ao passo que conseguem identificar rápida e diretamente suas informações (TURBAN *et al*., 2009). Ainda, Thompson (2004) descreveu, com base em pesquisa, que os principais benefícios do BI são a geração de relatórios rápidos e de precisão, a melhoria na tomada de decisões, no serviço prestado ao cliente e em uma maior receita.

2.4.1 Power BI

Power BI é uma ferramenta de *front-end* da Microsoft, que detém diversos componentes, como: *Power Query, Power View, Power Map, Power Pivot, Power Q&A* e *Power BI Desktop*, permitindo o desenvolvimento de modelos e solução que interliguem (ou não) todas essas tecnologias (MICROSOFT, 2022).

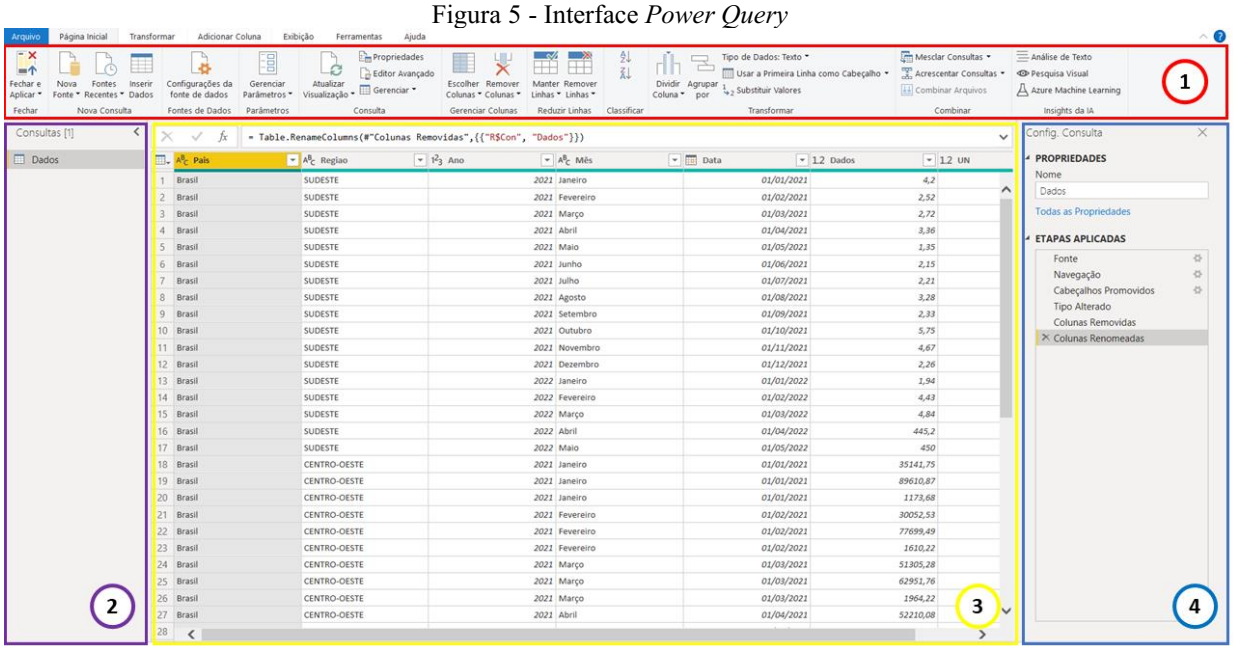
Para Sá et al. (2020), o *Power BI* é o *software* de BI *self service* que entrega todas as funcionalidades previstas com uma plataforma classificada líder, pela Gartner. A ferramenta se destaca, ainda, pela capacidade de tratamento dos dados, visto que nem todos os concorrentes de mercado tem essa mesma aptidão (LAGO; ALVES, 2018).

Haja vista seus principais complementos, a ferramenta é dividida em três partes principais, Dados (*Power Query Editor*), Modelo e Relatório. O *Power Query Editor*, de acordo com o portal oficial da Microsoft (2022), é um mecanismo para transformar e preparar os dados. Possui uma interface gráfica para obter dados de diferentes fontes e um editor do *Power Query* para aplicar transformações. A tecnologia é capaz de tratar os dados vindos de diversas fontes (como arquivos em nuvem, dados na internet, documento em TXT/CSV, PDF, SQL, Access, Excel, MailChimp etc.). Nele também é possível fazer interação com outros relatórios de outros sistemas (*DirectQuery*).

A interface do *Power Query* é dividida em quatro partes, como indicado na Figura 5.

1. Menu de opções: fornece diversas funcionalidades para adicionar transformações e selecionar opções de consulta;
2. Painel de consulta: lista de todos as consultas (bases importadas) disponíveis;
3. Modo de exibição: exibição da base a ser trabalhada;
4. Configurações de consulta: exibição da consulta selecionada e demais etapas de modificação ordenadas.

Figura 5 - Interface *Power Query*



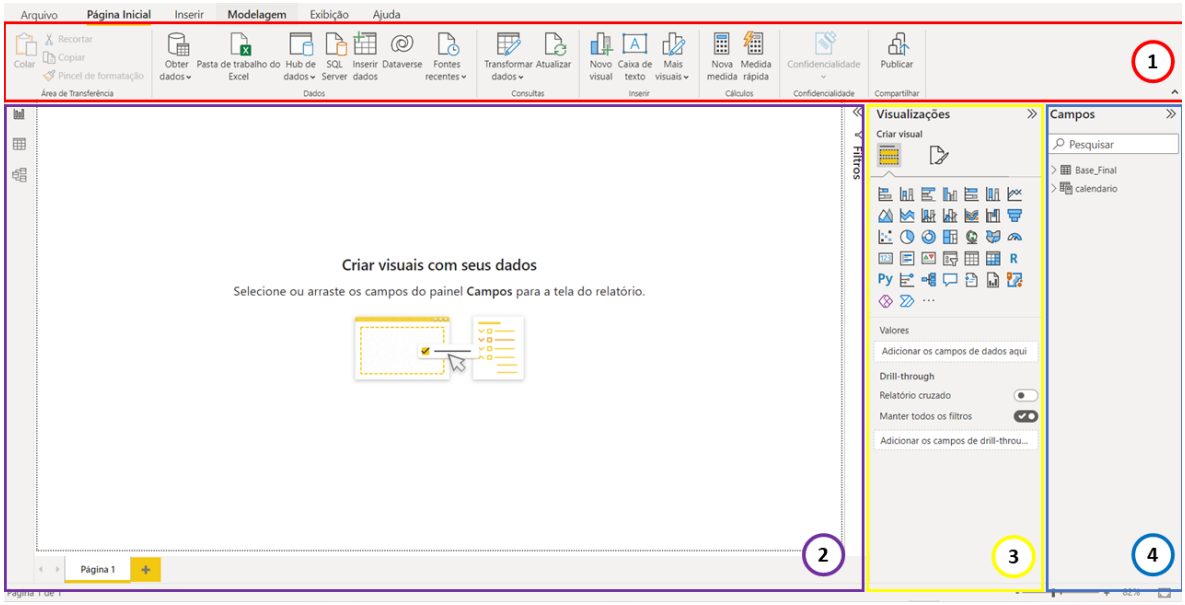
Fonte: Autor (2022)

De acordo com a Microsoft (2022), o *Power View* é uma tecnologia de visualização de dados que permite criar gráficos interativos, mapas e outros elementos visuais que fazem seus dados sobressaírem.

A interface é dividida em quatro partes, como apresentado na Figura 6:

1. Menu de opções: opções para interagir e manipular as visualizações criadas;
2. Painel de Dados: local onde as informações são apresentadas;
3. Visualizações: opções de gráficos, tabelas e matrizes para melhorar a visualização dos dados;
4. Campos: listagem com todas as consultas (dados) presentes no modelo.

Figura 6 - Interface *Power View*



Fonte: Autor (2022)

2.5 KPIS

Os KPIs (Key Performance Indicator) para uma empresa são instrumentos de avaliação e mensuração que permitem comprovar com objetividade, a partir da experiência e observação, a progressão quantitativa de um ou múltiplos processos de uma empresa (TERRIBILI FILHO, 2010).

Com base nisso, a empresa consegue analisar e identificar como está cada um dos números e traçar metas plausíveis e realizáveis. No que tange a visualização dos indicadores, tal qual a visualização dos dados apresentados nos capítulos anteriores, deve ser clara e direcionada, a fim de que a maior parte dos colaboradores detenham acesso e sejam capazes de compreender. Dessa maneira, decidir quais parâmetros avaliar é crítico para a efetividade do *dashboard*. Em retrospecto, muitos projetos não obtêm resultados satisfatórios pois as informações referentes aos KPIs não estavam claras, tornando-os confusos e ineficientes na tarefa de comunicar (KERZNER, 2013).

Ressalta-se que KPIs são aqueles que demonstram, em termos gerais, qual caminho a empresa está traçando de acordo com suas metas a níveis estratégico, tático e operacional (PARMENTER, 2015).

Dessa maneira, é imprescindível que os indicadores sejam claros e englobem os principais pontos de atenção para as empresas.

3 PROCEDIMENTOS METODOLÓGICOS

Este capítulo objetiva classificar a pesquisa quanto a seu enquadramento metodológico, em que apresenta a abordagem, natureza, propósito e procedimento técnico. Ainda, detalha-se quais os procedimentos adotados ao longo de toda execução do trabalho.

3.1 ENQUADRAMENTO METODOLÓGICO

Este estudo está inserido dentro da área Engenharia Organizacional, nas subáreas de Gestão do Desempenho Organizacional e Gestão da Informação, de acordo com a classificação da Associação Brasileira de Engenharia de Produção (ABEPRO).

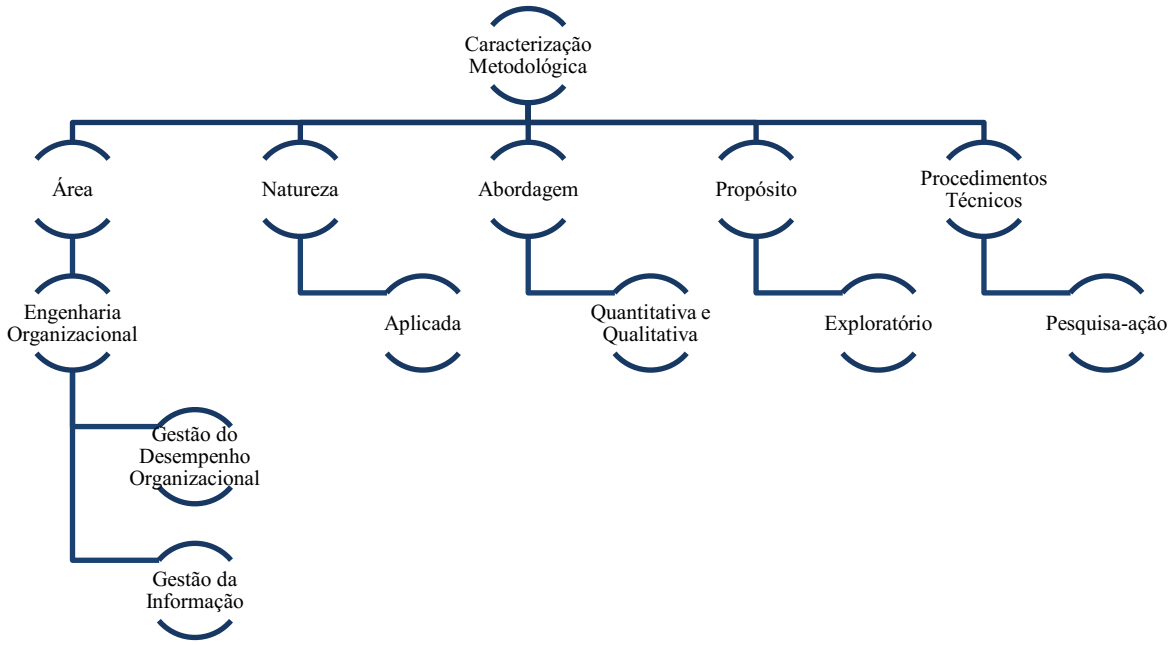
A pesquisa permeia sob diferentes pontos de vista. Em relação à forma de abordagem, pode ser caracterizada tanto qualitativa quanto quantitativa, já que objetiva desenvolver análises traduzindo números em informações e, ainda, analisa dados indutivamente (GIL, 2008).

Em relação à natureza, é classificada como aplicada, dado que visa gerar conhecimento para aplicações práticas na empresa objeto de estudo (SILVA; MENEZES, 2001). Referente ao propósito, a pesquisa é classificada como exploratória, o qual se justifica para temas ainda pouco explorados na literatura, que, portanto, anseiam por aprofundamento do conhecimento geral no assunto (GIL, 2008). Ainda, objetiva familiarizar um problema, envolvendo levantamento bibliográfico, análise de exemplos e entrevistas com aqueles que possuem relação com problema pesquisado (GIL, 2008).

Do ponto de vista de procedimento técnico, a pesquisa se desenvolve como uma pesquisa-ação, uma vez que se caracteriza pela coleta de informações, análises críticas em relação ao projeto e implementação de soluções. A metodologia da pesquisa-ação envolve participantes conduzindo inquéritos sistemáticos com a finalidade de ajudá-los a melhorar as suas próprias práticas, que por sua vez, podem também melhorar o seu ambiente de trabalho e os ambientes de trabalho das pessoas que fazem parte dela (KOSHY; KOSHY; WATERMAN, 2010). A “pesquisa-ação é uma forma de investigação-ação que utiliza técnicas de pesquisa consagradas para informar à ação que se decide tomar para melhorar a prática” (TRIPP, 2005). Portanto, um trabalho é classificado como pesquisa-ação quando é realizado em conjunto com ações ou resoluções de um problema coletivo, envolvendo os pesquisadores e outros participantes de modo cooperativo, e em que o pesquisador participa ativamente em sua concepção (GIL, 1991).

A Figura 7 apresenta, de maneira esquemática, a caracterização da presente pesquisa:

Figura 7 - Enquadramento Metodológico



Fonte: Autor (2022)

3.2 ETAPAS DA PESQUISA

Ao almejar cumprir o objetivo de construir um painel gerencial que unifique alguns indicadores financeiros e comerciais, apresentar resultados financeiros com base em previsões e, ainda, fornecer resultados eficientes quanto às suas variações, apresentando-as de maneira visual, o presente estudo segue os passos do modelo CRISP-DM para o desenvolvimento de um projeto de Mineração de Dados. A escolha do modelo se dá por apresentar uma vasta literatura disponível e por ser considerado o padrão de maior aceitação (LAROSE, 2005). Propõe-se, dessa forma, indicar e relacionar os dados de vinte diferentes categorias da companhia às etapas do modelo. A partir desta premissa, após a aplicação do modelo e dos cálculos de previsão, inicia-se a criação de um painel (*dashboard*) direcionado para a apresentação dos resultados, a fim de conceber uma ferramenta de fácil acessibilidade e com teor gerencial. Assim, o estudo pode ser resumido em três etapas: Processo CRISP-DM, Cálculos de Previsão e Criação do Painel Gerencial.

Apesar de serem etapas distintas, suas atividades se interligam e podem ser agrupadas no sequenciamento do Processo CRISP-DM, tornando-se parte do fluxo e garantindo sua execução com base em um único sequenciamento de atividades. Dessa maneira, o procedimento

metodológico limita a necessidades de retrabalho, uma vez que unifica atividades com processos semelhantes.

Nessa sessão, explica-se o fluxo de atividades do modelo, bem como sua adaptação para atender as necessidades do projeto em questão. Apresentando, inicialmente, as macros etapas separadamente e o fluxo final, após a adaptação.

3.2.1 Mapeamento CRISP-DM

O Mapeamento do modelo CRISP-DM utilizado no estudo é caracterizado pelas seis etapas recapituladas de forma resumida na Figura 8.

Figura 8 - Resumo CRISP-DM

Fonte: Autor (2022)

Todas as etapas descritas na Figura 8 foram seguidas à risca, demonstrando assim, a eficiência de um modelo adotado em diversos estudos atuais de mineração de dados, como é o caso de:

- Uma revisão sistemática da literatura sobre a aplicação do modelo de processo CRISP-DM (SCHRÖER; KRUSE; GÓMEZ, 2021).
- Criação de projeto de ciência de dados utilizando a metodologia CRISP-DM em conformidade com a LGPD (LIMA, 2021).
- Previsão de tempos de internamento num hospital português: aplicação da metodologia CRISP-DM (LAUREANO; CAETANO; CORTEZ, 2014).
- Desenvolvimento de solução CRISP-DM para classificação do evento prisão de coluna no processo de perfuração de poços offshore (PINTO, 2018).

.

3.2.1.1 Entendimento do Negócio ou Domínio

Nesta etapa, o estudo busca entender os problemas a serem sanados com o modelo proposto. Essencial para o desenvolvimento, a etapa é responsável por identificar, a partir das necessidades da empresa do setor de saúde, os principais indicadores capazes apresentar seu estado atual e direcioná-la para a ampliação de sua lucratividade, além dos indicadores que à auxiliem a compreender sua desenvoltura de vendas e das concorrentes no mercado do setor de saúde. Por fim, delimita-se os requisitos do estudo, como os atributos base para os cálculos dos indicadores e a ferramenta adotada para a modelagem.

3.2.1.2 Entendimento dos dados

A etapa visa entender as informações relacionadas aos dados, como sua atualização, fonte, limitações, tratamento etc. Analisa-os quanto a sua disponibilização, se podem ser apresentados no estudo ou se devem ser mascarados e se são capazes de suprir as necessidades definidas na etapa anterior. Dessa maneira, a etapa está relacionada a compreensão total dos dados utilizados, desde sua geração, passando por formatação e entendimento do nível de informações que podem ser extraídas, até em relação a como seriam processados e atualizados.

3.2.1.3 Preparação dos dados

O desenvolvimento da preparação de dados consolida o bom andamento posterior, de modelagem, uma vez que é nesta etapa que os dados são tratados, transformados e formatados. A base disponibilizada pela empresa é avaliada, a fim de encontrar possíveis problemas como: dados sem leitura, formatos diferentes do que a ferramenta de modelagem aceita ou, ainda, eliminar os dados que não sejam essências para solucionar problema identificado. Dessa maneira, a etapa é responsável por suprimir todas as inconformidades encontradas.

3.2.1.4 Modelagem

Na modelagem os cálculos são estabelecidos, bem como a criação das interações necessárias para atender os objetivos propostos para o protótipo. Nesta etapa, desenvolve-se a criação de indicadores financeiros, incluindo a projeção do faturamento, e comerciais da empresa e de seus concorrentes no setor farmacêutico. A partir dos cálculos e interações, ainda nesta etapa, desenvolve-se a criação de um *dashboard* que agrupa todas as informações e as apresenta de maneira clara e visual.

3.2.1.5 Avaliação

A avaliação consiste na análise do modelo, buscando garantir a conformidade dos resultados. Nesta etapa, o modelo é testado e comparado com outros relatórios que apresentem informação semelhantes, além de calcular manualmente, na própria base de dados, os indicadores desenvolvidos.

3.2.1.6 Implementação

A etapa de implementação consiste na aplicação do modelo dentro da empresa, garantindo a boa gestão de informações e a facilidade de acesso a indicadores importantes para a gestão da companhia.

4 DESENVOLVIMENTO

4.1 MAPEAMENTO CRISP-DM

A sessão apresenta as etapas do Modelo CRISP-DM de maneira detalhada, elencando os pontos de atenção e o que foi definido e desenvolvido em cada uma delas.

4.1.1 Entendimento do Negócio ou Domínio

Para a execução desta etapa, algumas perguntas devem ser respondidas, como: Qual o objetivo a ser alcançado? Quais os requerimentos? Quais os riscos? Os benefícios? Quais as ferramentas a serem utilizadas? (CHAPMAN *et al*., 2000).

Dentre os objetivos da análise, pode-se definir que o principal está relacionado a apresentar, de maneira clara e direta, os resultados das principais marcas da empresa, gerando *insights*. Dessa forma, é imprescindível delimitar alguns instrumentos de análise, como faturamento, perspectiva quanto a unidades vendidas, quais as regiões mais rentáveis etc.

Inicialmente, para que fossem entendidas as necessidades da empresa para o presente estudo, identificou-se que o principal canal no qual ela detinha maior relevância relacionada a sua lucratividade, era o canal farmacêutico. De acordo com a empresa, o canal corresponde a cerca de 65% de todo seu faturamento e é alvo do foco estratégico das áreas comerciais e de vendas. Dessa maneira, a consolidação de informações para o canal se mostrou uma preocupação mais urgente em detrimento às demais.

Em virtude do elevado número de marcas que a empresa detém, foi definido um escopo realista capaz de apresentar resultados quanto às 30 principais. Por serem caracterizadas como os principais agentes promotores do nome da empresa e, ainda, suas maiores alavancas de faturamento e desenvolvimento, as marcas adotadas no presente estudo, são suficientes para que seus objetivos sejam alcançados. Ainda, a empresa apresenta seu portifólio de produtos divido em 37 categorias, nas quais estão incluídas as marcas objeto de análise, as pouco promotoras e, ainda, as que estão sendo retiradas do portfólio da empresa. Em virtude disso, delimita-se como escopo inicial as 20 categorias que agrupam as 30 principais marcas da empresa, possibilitando boas análises e direcionamentos concisos.

Com esta delimitação, ressalta-se a possibilidade de exclusão de categorias que estariam crescendo, mas que ainda não representavam impactos significativos de faturamento.

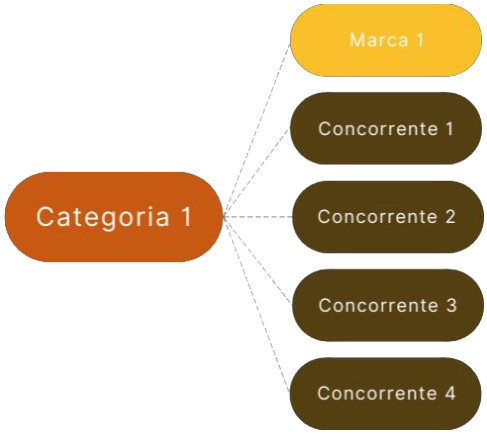
Todavia, entende-se que apesar de possíveis, esses impactos são pontuais, garantindo uma boa continuidade das análises e demonstrando que o estudo apresenta baixos riscos.

Por se tratar de uma multinacional, a empresa possui diferentes produtos em diferentes regionalidades. Com isso, foi necessário, inicialmente, definir a regionalização do estudo, delimitando-o apenas no território nacional. Antes mesmo de analisar os dados disponibilizados, era essencial que isso fosse definido a fim de traçar estudos direcionados para cada estado do país. Esta delimitação contribui para facilitar as próximas etapas, uma vez que em diferentes países, existiriam fontes de dados diferentes, gerando classificações diferentes e uma maior complexidade em seu tratamento.

Em relação ao mercado de atuação, entende-se que a empresa apresenta para cada uma de suas marcas, diversos concorrentes, de diferentes tamanhos e diferentes impactos no mercado. Tendo em vista que estudo é direcionado para as principais marcas de uma empresa já consolidada no mercado, entende-se que, apesar de impactarem na sua penetração, concorrentes com baixa relevância, não seriam apresentados nas visões do *Dashboard*, contribuindo para uma visualização mais direcionada e clara.

Dessa maneira, a divisão de como os estudos se sucedem pode ser apresentada, de maneira simplificada, na Figura 9.

Figura 9 - Representação da subdivisão de categorias



Fonte: Autor (2022)

Ainda, buscando desenvolver um modelo acessível e com baixa complexidade de interpretação, definiu-se como objetivo criar um painel gerencial que, de forma eficiente e eficaz, permita monitorar a aplicação face ao registro do desempenho dessas marcas e, assim, prestar um melhor serviço aos consumidores. De maneira resumida, o estudo se dará nas 30 principais marcas, agrupadas em 20 categorias, focalizadas no território nacional e com o

objetivo de construir um ambiente de visualização robusto e direcionado, capaz de apresentar os resultados e as previsões de maneira clara.

4.1.2 Entendimento dos dados

A criação de um modelo, seja ele para acompanhamento de resultados ou de previsão de vendas quantitativo, se deu a partir da coleta de dados, agrupados com informações que identificavam seus resultados e propiciavam explicações quanto às vendas.

Os dados adotados no estudo são atualizados, configurados e disponibilizados mensalmente por uma fornecedora de dados para as áreas de interesse por meio de um banco de dados relacional, cuja linguagem de comunicação é a SQL (*Standard Query Language)*. Com base nisso, sua atualização se faz de maneira automatizada para a empresa, não necessitando realizar importações de dados volumosas, levando grande período de tempo. Dentre os dados disponibilizados, identifica-se que os períodos que sucedem o ano de 2020, apresentam muitos produtos que já saíram de linha, produtos sem leitura ou alguns dados que não correspondem às especificações do estudo. Com isso, foram utilizados dados a partir do ano 2020, garantindo uma base de dados concisa e que gerará menos inconformidades a serem corrigidas na etapa posterior.

A base de dados, apesar de robusta, refere-se apenas aos clientes do canal farmacêutico da companhia, não possuindo dados quanto a sua distribuição no setor alimentar, como supermercados e atacados ou de perfumaria.

Nota-se, ainda, que os dados não estão todos padronizados para que sejam diretamente incluídos na ferramenta de modelagem. Dessa maneira, precisarão ser tratados na etapa subsequente.

É imprescindível ressaltar que os dados apresentados neste estudo foram todos mascarados e suas respectivas titulações não serão apresentadas, de modo a preservar o sigilo de dados da empresa, a fim de garantir uma boa gestão estratégica ante a suas concorrentes.

A apresentação dos resultados, ou seja, no Painel Gerencial, deve contemplar as informações com base na lista de indicadores estabelecida na revisão da literatura e, ainda, na previsão calculada. Tendo como base a extração de dados disponibilizada para o período que inicia no ano de 2020 e vai até o último mês atualizado até o presente estudo (maio de 2022), entende-se que a base tem a volumetria necessária para os devidos cálculos necessários e não apresenta riscos muito relevantes no que tange sua atualização.

Com base nisso, reitera-se a finalidade de trazer informações relevantes quanto a exploração de uma base de dados robusta, garantindo a boa execução de estratégias das marcas e ampliação no que tange a penetração e o crescimento no mercado.

4.1.3 Preparação dos dados

A etapa que consiste na manipulação de dados se desenvolveu com base no tratamento dos dados e na seleção do que, efetivamente, iria compor o *dataset* da modelagem.

Inicialmente, a base extraída do banco de dados da empresa, foi convertida para formato CSV e conectada com o Power BI através do Power Query. A base possui registros do ano de 2020 até maio 2022, com mais de um milhão de linhas se exportados em uma única tabela de dados. Com isso, a base foi separada em duas tabelas distintas, a fim de não haver problemas quanto a sua execução, a primeira contendo informações gerais do fabricante, marca, EAN (*European Article Number*), categoria, unidades vendidas, faturamento e, ainda, dados mais gerais, como por exemplo, o período no qual se desenvolveu a arrecadação.

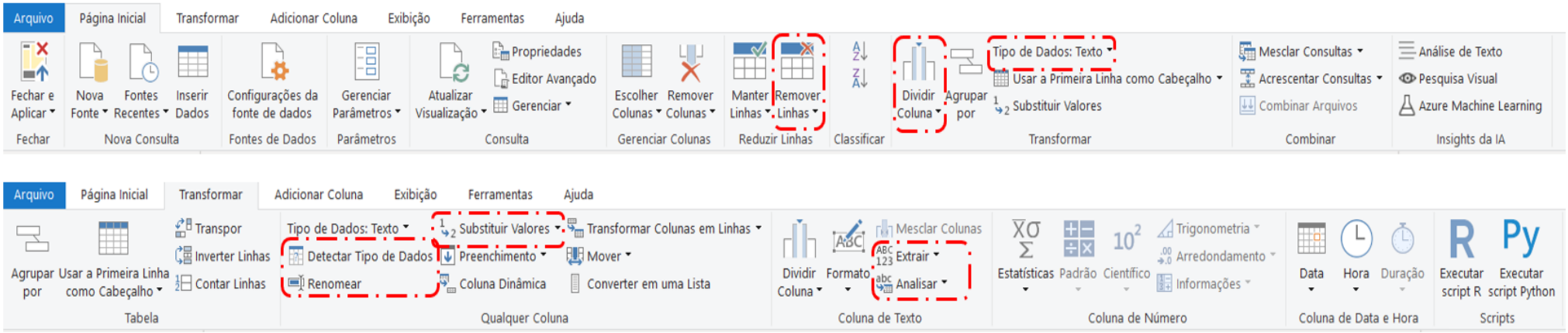
Já a segunda era um pouco mais sucinta, contendo informações quanto a UF, unidades vendidas, EAN, faturamento e o período no qual os dados estão relacionados. Dessa maneira, reitera-se que apesar de o modelo apresentar mais de uma tabela de dados, incluindo as auxiliares, todos os dados utilizados no relatório originaram-se da mesma fonte de dados da empresa.

A ferramenta classifica os dados em, basicamente, três tipos: número, texto e data. Com isso, se faz necessário compreender as necessidades inerentes de cada uma das colunas, a fim de estabelecer classificações condizentes com o que é esperado nos resultados.

Após sua classificação, por ser uma base com alta gama de dados, o tratamento detalhado é de extrema importância para que os cálculos sejam precisos e apresentem resultados confiáveis.

Utilizando as funcionalidades do Power Query, a etapa iniciou com a exploração dos dados, buscando compreender, para cada uma das informações presentes no banco de dados, os pontos de atenção a serem corrigidos. As funcionalidades utilizadas estão representadas na Figura 10.

Figura 10 - Funcionalidade utilizadas *Power Query*

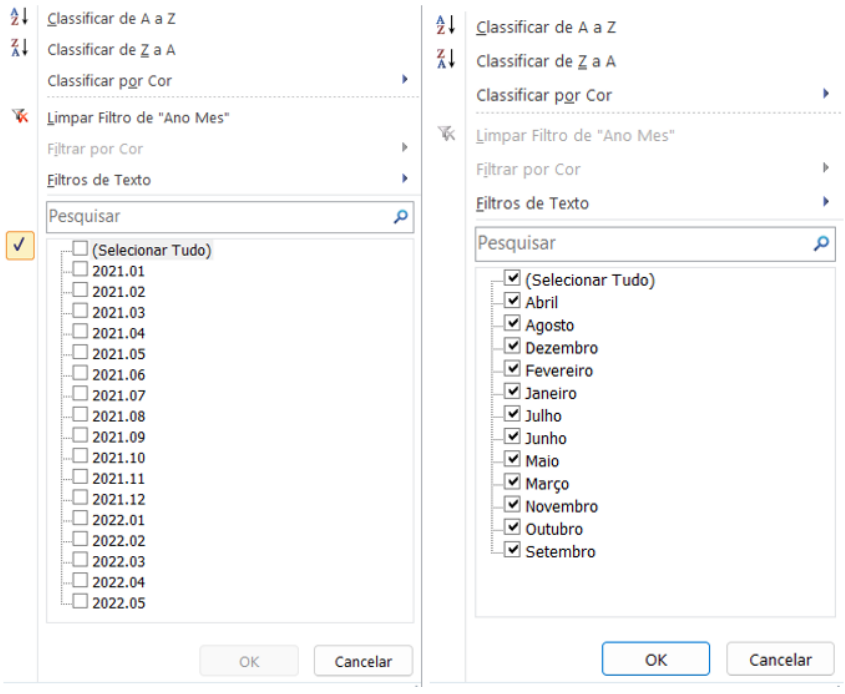


Fonte: Autor (2022)

Para cada uma das colunas das tabelas de dados, foi aplicada a funcionalidade de detecção de dados e, em casos de necessidade, a de mudança seu tipo, para texto, número decimal, número inteiro ou data. Os atributos que se referiam ao nome da Marca, Canal, Fabricante, Categoria, Estado e Região, foram classificados como tipo texto. Ainda, o atributo EAN, apesar de ser um código, também foi tratado como texto, para que não houvesse possibilidade de o sistema interpretá-lo como parte de algum cálculo. Em relação aos atributos de Unidades e Faturamento, os tipos atribuídos foram número inteiro e decimal fixo, respectivamente. Desse modo, após todos os dados estarem tratados quanto a seu tipo, a etapa seguiu para a formatação dos dados.

Os dados relacionados aos meses de estudos, apresentavam a formatação de "ANO.MÊS", como apresentado no lado esquerdo da Figura 11. Com base nisso, o sistema não era capaz de agrupar os dados em relação ao mês, nem ao ano. Utilizando a ferramenta de divisão de colunas, foi possível ter uma coluna com as informações do número do mês, bem como, seu respectivo ano.

Figura 11 - Formato inicial e atualizado das datas



Fonte: Autor (2022)

Ainda, houve a necessidade de converter o número do mês para seu respectivo nome, utilizando a função de substituição de valores. Após sua transformação, os dados ficaram formatados de modo que os agrupamentos mensais fossem padronizados, independentes do ano de origem, como apresentado no lado direito da Figura 11.

No que tange o tratamento quanto ao número de dados analisados, fez-se uma limpeza em relação à quantidade de categorias a serem estudas. Em virtude de a base apresentar, inicialmente, dados históricos a partir de 2017, diversas categorias listadas não faziam mais parte do portfólio da empresa, mas ainda apresentavam leituras quanto a seus concorrentes ou alguns dados considerados como "lixo da base". Dessa maneira, para que a base fosse reduzida, gerando menor tempo de processamento, e assertividade dos resultados, os dados de dezessete das trinta e sete categorias existentes, foram excluídos.

Ainda, analisando os dados remanescentes na base, identificaram-se leituras cujos valores estavam zerados ou apresentavam alguma divergência, com base na utilização da funcionalidade de análise dos dados. Estes, foram mapeados e apresentados à fornecedora dos dados, que por sua vez, identificou erros nos cálculos internos e devolveu uma base com as leituras acertadas. Ainda quanto aos dados zerados, alguns dos EANs não apresentavam leituras quanto ao faturamento, nem quanto às unidades vendidas, gerando resultados menores para os agrupamentos nos cálculos. A partir disso, foi necessária a exclusão de linhas que apresentavam erros, com a utilização da função de exclusão de linhas.

No mesmo sentido, foram identificados erros na disponibilização dos dados referentes a unidades vendidas, que apresentavam unidades com valores decimais. Essas informações sofrem cálculos errôneos quanto à conversão de unidades dos produtos da categoria de analgésicos, que, para algumas marcas, o cálculo foi desenvolvido em relação aos *blisters* (cartela de comprimidos) e em outras, às caixas vendidas. Após a apresentação dos problemas identificados, a fornecedora reprocessou os dados padronizando-os em relação a um único tipo de conversão. Ressalta-se a importância dessa descoberta não só para o estudo, mas para a companhia como um todo, que usava a mesma base para diversos estudos.

Após o tratamento de todas as pendências existentes nos dados, foi extraída uma base final em formato CSV. Esta base foi chamada de "base_empresa", sendo utilizada para toda a Análise e Mineração de Dados realizada neste trabalho. Contudo, como relatado anteriormente, em virtude de sua dimensão, a base foi dívida em duas tabelas distintas, doravante denominadas, "Categoria&Marca" e "Região". Após todo o tratamento de dados na ferramenta *Power Query* e aplicação das demais etapas do CRISP-DM, caso houvesse a necessidade de tratar algum dado, eram nessas tabelas que que se desenvolviam as adaptações. Então ocorria o tratamento ou criação de um novo atributo, conforme a necessidade, gerando novamente uma segunda base de dados atualizada.

4.1.4 Modelagem

A etapa de modelagem consistiu na aplicação de todos os cálculos necessários para o desenvolvimento do painel gerencial. Nessa etapa, diversos indicadores de desempenho foram analisados e aplicados, a fim de se estabelecer informações factíveis e de fácil entendimento para seus visualizadores (gestores usuários).

Para que o painel apresentasse informações suficientes para atender as necessidades da empresa, buscou-se modelar o estudo de modo a apresentar informações financeiras e de mercado, agrupando em métricas que poderiam ser usadas nas modelagens de ambos.

4.1.4.1 Modelagem Financeira

A modelagem financeira consistiu nos cálculos exclusivos para à companhia. Os indicadores aqui projetados, referem-se exclusivamente ao desempenho de faturamento e unidades vendidas da empresa.

Para elaboração da modelagem financeira, as métricas foram criadas com base na tabela “Região”, que apresentava informações quanto à distribuição por estados, categorias, unidades vendidas, faturamento e datas. Em virtude da não correlação com os dados de mercado, como o faturamento das concorrentes, por exemplo, a escolha da tabela se deu pelo fato de apresentar a distribuição regional dos dados.

A primeira modelagem foi desenvolvida de maneira simples, ligando os dados de faturamento da empresa com ao seu respectivo estado, de modo a apresentar sua distribuição regionalizada.

Ressalta-se, nesse cálculo, a necessidade de voltar ao tratamento de dados para formatação do nome dos estados. Em virtude da acentuação nos nomes dos estados, a ferramenta não era capaz de realizar a identificação dos dados, deixando essas regiões sem leituras. Dessa maneira, foi necessário um novo tratamento nos dados, a fim de que os erros fossem corrigidos e as leituras computadas.

Em seguida, foram desenvolvidos os cálculos acumulados do estudo, apresentando o faturamento e a quantidade vendida acumulados até o período de análise estabelecido pelo usuário do painel. Para isso, bastou modelar os cálculos de modo que os dados de faturamento da empresa fossem somados até a data escolhida. Dessa forma, a modelagem definida esta apresentada na Figura 12.

Figura 12 - Modelagem faturamento acumulado

```
Faturamento Acumulado = CALCULATE(
[Faturamento Empresa],FILTER(ALL
(Calendario),Calendario[Date]<=MAX
(Calendario[Date])))
```

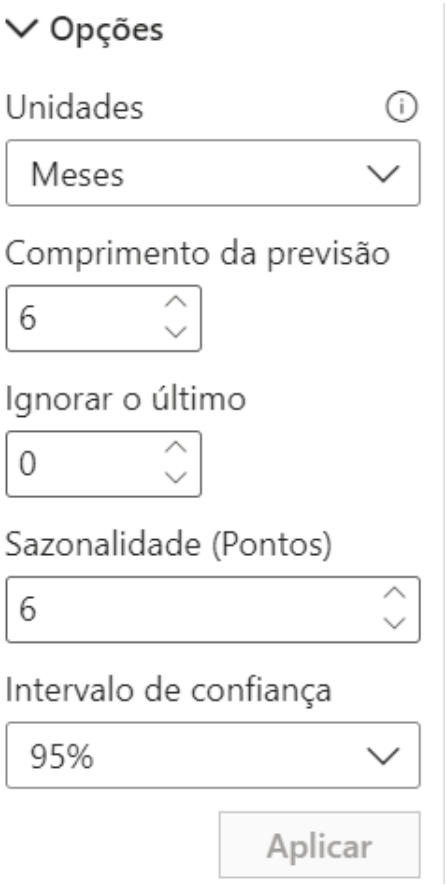
Fonte: Autor (2022)

O mesmo processo foi desenvolvido em relação ao cálculo acumulado de unidades vendidas, substituindo apenas o atributo "[Faturamento Empresa]" por "[Quant Vendida]".

Por fim, foi desenvolvida uma modelagem quanto à previsão de receita da empresa. Para essa modelagem, utilizou-se a tecnologia de visualização de dados e realização de análises estatísticas avançadas, *Power View,* presente nas ferramentas do *Power* BI. O *Power View* faz análises avançadas dos dados em gráficos de linha para gerar previsões que incorporam tendências e sazonalidade, com base no algoritmo de suavização exponencial Holt-Winters, possibilitando a seleção automática do modelo apropriado, com base nos dados históricos. Sua utilização se deu pela necessidade de atribuição de fatores de tendência e sazonalidade, uma vez que o algoritmo permite a correção de múltiplos parâmetros e da maior relevância a leituras mais recentes. Ressalte-se a relação de pesos atribuídos pela ferramenta, que atribui pesos menores a leituras mais antigas de maneira automática.

A visualização dos atributos de entrada para a previsão está descrita na Figura 13.

Figura 13 - Atributos de previsão



Fonte: Autor (2022)

O atributo “Comprimento da previsão” foi definido com base no período no qual são revisadas as estratégias da empresa. A cada seis meses são redesenhados os direcionamentos para que as metas sejam alcançadas. Dessa maneira, definiu-se que a previsão de seis meses a frente seria a ideal, uma vez que quanto mais elevado o comprimento de previsão, menor será sua assertividade.

Em relação ao atributo “Ignorar o último”, definiu-se por não ignorar nenhum mês de análise, a fim de apresentar o comportamento da previsão com o maior histórico possível.

O atributo “Sazonalidade (Pontos)” leva em consideração oscilações sazonais, ou seja, é com base no atributo que se pode mitigar erros quanto a uma projeção com base em dados sem uma constância considerável. Dessa maneira, levando em consideração as oscilações históricas apresentadas pela companhia, identificou-se que em intervalos de aproximadamente seis meses, ocorre um crescimento ou decaimento gradual do faturamento. Portanto, definiu-se o atributo com um valor igual a seis.

Por fim, o último atributo está relacionado à acuracidade da previsão. Trata-se de uma estimativa que apresenta o intervalo no qual o parâmetro (neste estudo, o faturamento) se

encontra com um determinado nível de probabilidade. Como citado em capítulos anteriores, a previsão é uma estimativa e não apresenta o valor real do parâmetro em todos os casos. Dessa forma, o intervalo de confiança é o que define um valor máximo e mínimo para o parâmetro em questão. É valido ressaltar que quanto maior o intervalo de confiança, maior será o número de possibilidades para a estimativa. Dessa forma, para o presente estudo, definiu-se um intervalo de confiança, com base em nível de confiança de 95%, que garante uma boa cobertura de assertividade e menos possibilidades de valores, como é o caso de 99%.

Dessa maneira, com base nos atributos e nos dados históricos disponibilizados na tabela "Região", o gráfico contendo os resultados do modelo é gerado.

4.1.4.2 Modelagem comercial

A modelagem de mercado, diferente da modelagem financeira, consistiu nos cálculos das relações entre os dados da empresa e de seus concorrentes. Os indicadores projetados nessa etapa, promovem um comparativo de informações, a fim de compreender a atuação da empresa ante seus concorrentes, como por exemplo, a parcela de mercado que ela detém.

Para elaboração da modelagem de mercado, as métricas foram criadas com base na tabela "Categoria&Marca", que apresentava informações quanto a distribuição de marcas, categorias, unidades vendidas, faturamento e datas. Com esses dados, foi possível criar indicadores comerciais não só para empresa, mas também para seus concorrentes, de modo que as estratégias pudessem ser traçadas analisando o resultado das concorrentes em comparação aos seus.

A primeira modelagem foi em relação à parcela de mercado, mais conhecida como *Market Share.* Essa métrica é uma porcentagem que corresponde à relevância da empresa diante dos competidores do segmento em que atua. O *Market Share* pode ser interpretado por diversas óticas, como penetração da marca, número de usuários etc. Para o presente estudo, foi utilizada a ótica de volume de vendas, uma vez que a base de dados apresentava as unidades vendidas de cada marca das empresas que compõe o mesmo setor de atuação da empresa em análise.

Dessa maneira, o *Market Share* pode ser representado da seguinte maneira:

$$Market\ Share = {}^{Vendas\ da\ empresa}/_{Vendas\ totais\ do\ segmento}$$

No desenvolvimento do modelo, que visava apresentar o *Market Share* por marcas dentro de uma mesma categoria, era imprescindível que os cálculos levassem em consideração o total da categoria e não o total geral da base de dados, a fim de que os resultados fossem referentes apenas a categoria em que está enquadrado.

$$Market\ Share = {}^{Vendas\ da\ marca\ na\ categoria}/_{Vendas\ totais\ das\ marcas\ dentro\ da\ categoria}$$

Em seguida, foi desenvolvida a modelagem quanto aos efeitos preço, volume e mix. A análise de preço, volume e mix consiste na apresentação dos impactos que cada um desses *drivers* (preço, volume e mix) tem na receita total da empresa, de modo que se possa ter visibilidade da situação de cada tipo de produto e, assim montar planos de ação para recuperação dos drivers detratores de maneira mais rápida e assertiva. Esse indicativo poderia fazer parte da sessão de modelagem financeira, apresentando resultados exclusivamente da empresa em análise. Contudo, optou-se por desenvolver um modelo capaz de filtrar as informações dos concorrentes, de modo que fosse possível analisar seus desempenhos e identificar pontos de atenção, para que, com base nos *drivers* detratores e promotores que as principais concorrentes adotem, a empresa possa se precaver.

Dentro da empresa, a gerência define os *drivers* de preço, volume e mix, em que eram calculados de maneira manual, gerando grande desperdício de tempo para a empresa. Com base nesses indicadores comerciais, é possível mensurar o desempenho das vendas em determinados períodos, contribuindo para a compreensão do que influenciou o ganho ou a perda de faturamento de um período para o outro. Dessa maneira, cada *driver* acarreta em uma interpretação para a empresa.

- Preço: utilizado para identificar o impacto de reajustes no preço e das políticas de desconto;
- Volume: retrata a variação das vendas e a evolução da carteira de clientes em termos absolutos;
- Mix: representa a variação no balanceamento entre os produtos mais caros e os mais baratos.

Assim, os cálculos são desenvolvidos da seguinte forma:

$$Efeito\ Preço = (Preço\ Médio_{Atual} - Preço\ Médio_{Ano\ Anterior}) * Unidades\ Vendidas_{Atual}$$

$$Efeito\ Volume = (Unidades\ Vendidas_{Atual} - Unidades\ Vendidas_{Ano\ Anterior}) * Preço\ Médio_{Ano\ Anterior} - Efeito\ Mix$$

$$Efeito\ Mix = (\%\ Unidades\ Vendidas_{Atual} - \%\ Unidades\ Vendidas_{Ano\ Anterior}) * (Preço\ Médio_{Atual} - Preço\ Médio_{Ano\ Anterior}) * Unidades\ Vendidas_{Atual}$$

A partir dos resultados encontrados, é importante que sejam correlacionados, a fim de definir os possíveis pontos de atenção na desenvoltura dos produtos. Dessa maneira, de acordo com o que é adotado na empresa, os resultados podem ser interpretados da seguinte maneira:

- Efeito Volume positivo, Efeito Mix negativo, Efeito Preço negativo = possivelmente o mix de produtos vendidos deteve pouca variedade e baixo valor agregado.
- Efeito Volume positivo, Efeito Mix positivo, Efeito Preço negativo = possivelmente o mix de produtos vendidos deteve muita variedade e baixo valor agregado.
- Efeito Volume positivo, Efeito Mix positivo, Efeito Preço positivo = possivelmente o mix de produtos vendidos deteve muita variedade e com alto valor agregado.
- Efeito Volume positivo, Efeito Mix negativo, Efeito Preço positivo = possivelmente o mix de produtos vendidos deteve pouca variedade e alto valor agregado.

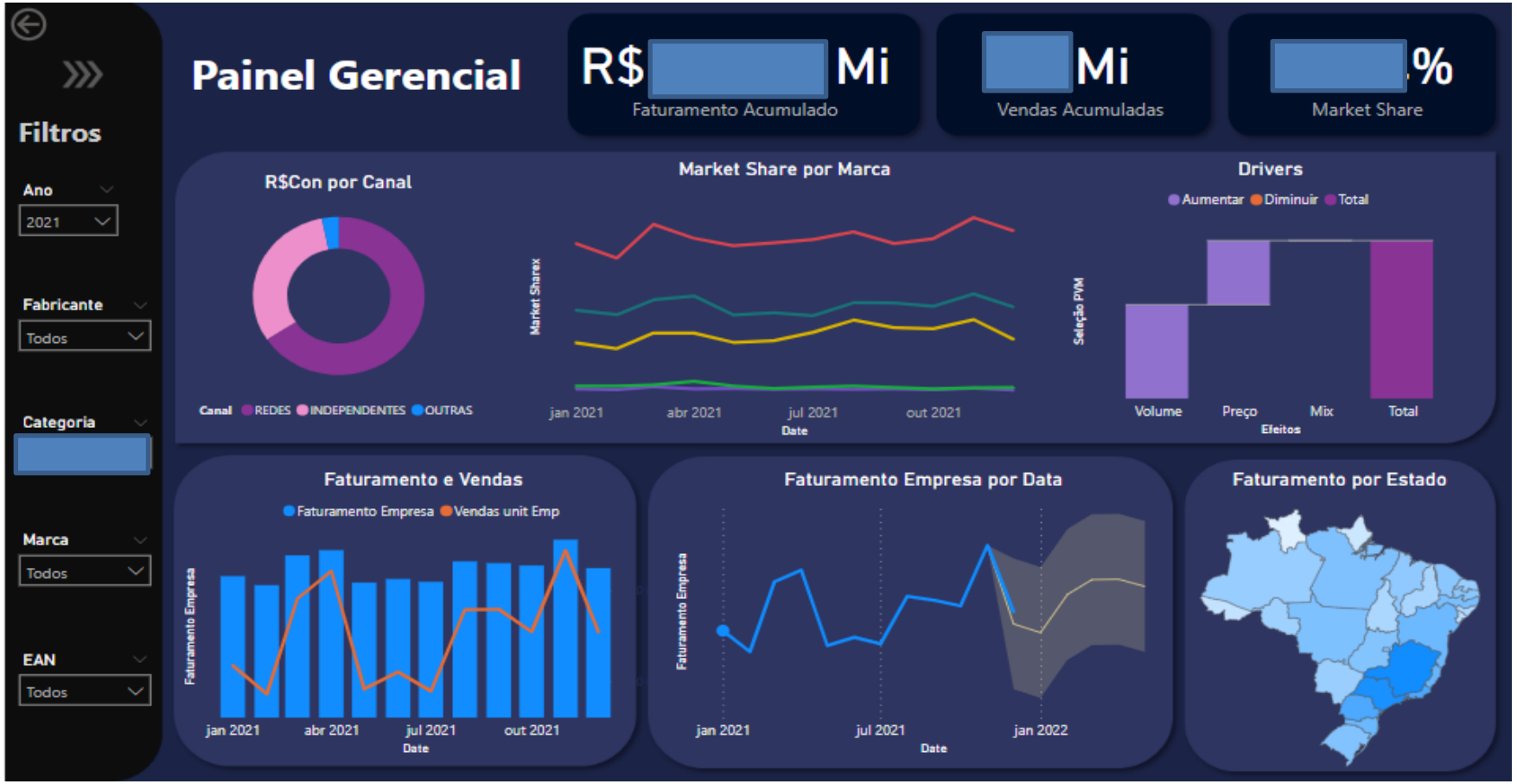
As outras quatro combinações não representavam interpretações muito relevantes para a companhia, dessa maneira não foram descritas.

Com base nas métricas apresentadas, os indicadores para o modelo geral do painel são concluídos, de modo que consiga atender as necessidades das gerências da empresa em relação às informações financeiras e de desempenho no setor farmacêutico.

4.1.4.3 Resultado das modelagens

Com base em toda a modelagem de cálculos e indicadores e após a adaptação das questões estéticas relacionadas à apresentação do painel gerencial completo, obteve-se como resultado o modelo apresentado na Figura 14. Os resultados numéricos, neste trabalho, foram ocultados de modo a preservar as informações confidenciais da empresa, mas sem prejuízo quanto aos resultados do trabalho em questão. O painel apresentado na Figura 14 demonstra, ainda, os agrupamentos de informações correlatas, dispondo, na sessão superior, os dados gerais da empresa, na sessão do meio os dados relacionados aos indicadores comerciais e na última sessão os dados financeiros. Ainda, podem ser identificados os filtros que podem ser aplicados no painel gerencial, localizados no canto esquerdo do painel.

Figura 14 - Painel gerencial desenvolvido



Fonte: Autor (2022)

4.1.5 Avaliação

Na etapa de avaliação, os resultados foram analisados por meio da interação com o painel gerencial e comparados tanto com cálculos desenvolvidos na própria fonte de dados, quanto com outros painéis gerenciais da empresa que apresentavam informações parecidas. Essa etapa de validação é corriqueira com todos os *Dashboards* publicados para todos os colaboradores da companhia e realizada por mais de uma pessoa, a fim de mitigar possíveis equívocos de cálculo. Ainda, para esse estudo, foi solicitada a interação com colaboradores com pouco contato com o tipo de ferramenta implementada, buscando encontrar pontos que não ficam claros no modelo e que possam ser aprimorados.

A partir das análises dos próprios colaboradores e do desenvolvedor do modelo (autor), não foram identificadas inconsistências quanto aos resultados, garantindo efetividade nas análises permitidas. Ressalta-se a descoberta e tratamento de inconformidades na própria base de dados, até então, não identificados pela empresa.

Ainda quanto ao retorno dos avaliadores, destacam-se comentários quanto a facilidade de acesso e navegação pelos resultados, resultando em uma interface completa e bem desenvolvida, que proporciona análises diretas e bem fundamentadas mesmo para aqueles com pouca familiaridade com a ferramenta.

Como resultado da etapa de análise, identifica-se o sucesso do modelo, que apresentou informações concisas e uma boa avaliação.

4.1.6 Implementação

Até a data do presente estudo, a etapa de implementação foi a única não concluída, em virtude dos processos de automatização da utilização de dados no modelo e da burocracia para publicação de novos painéis gerenciais.

Com a extração de dados e posterior utilização na ferramenta em que foi produzida, foi realizada de maneira manual, por se tratar de uma proposta de modelo, para que a implementação seja realizada, é necessária a conexão direta com o banco de dados para que não existam problemas quanto à atualização. Esta será uma etapa posterior ao presente estudo. Contudo, reitera-se que os dados disponibilizados no modelo propostos são passíveis de análises, uma vez que apresentam resultados corretos.

Ainda, em relação a burocracia de implementação de um novo painel, ressalta-se que sua utilização é possibilitada para um restrito grupo de pessoas, por, ainda, não se tratar de um painel oficial da empresa.

4.2 DISCUSSÃO DE RESULTADOS

Para discorrer sobre os resultados obtidos com o trabalho desenvolvido, optou-se por analisar a efetividade do modelo escolhido para desenvolvimento dos cálculos e indicadores e ainda a qualidade dos resultados apresentados no painel gerencial desenvolvido. Dessa maneira, para que o projeto obtivesse sucesso, todas as etapas do modelo CRISP-DM deveriam ser seguidas, objetivando ter uma base de dados configurada para as necessidades do projeto, além da apresentação de todos os indicadores necessários para gestores da empresa em análise.

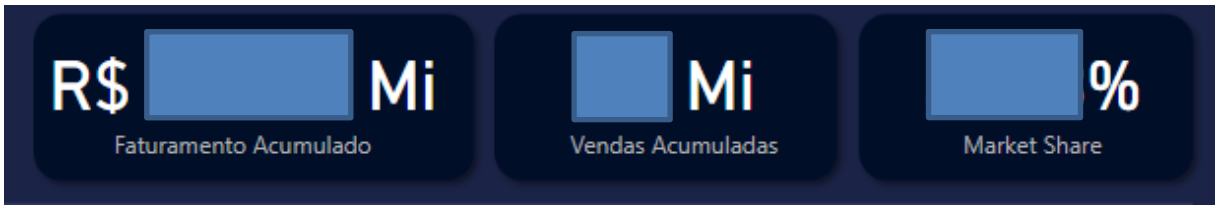
Como destacado ao longo da sessão anterior, todas as etapas do modelo CRISP-DM foram seguidas, desde o entendimento das necessidades para o projeto até sua possível implementação. Ressalta-se para os resultados quanto à utilização do modelo, a comprovação de caracterizar-se como forte aliado no tratamento e processamento de grandes volumes de dados, permitindo a criação de um painel robusto e com informações corretas e relevantes para as partes interessadas.

Com base na mineração dos dados utilizando o modelo CRISP-DM, foi possível identificar erros na base de dados, que anteriormente haviam passado despercebidos e acabavam impactando nos resultados dos demais relatórios da empresa. Dessa maneira, o projeto conseguiu ir além, contribuindo para a correção de um problema que afetava toda a empresa. Os resultados inerentes ao desenvolvimento do modelo mostraram a necessidade de aplicação de técnicas de mineração de dados para empresa, que além da fonte utilizada, detém diversas outras fontes de dados de igual ou maior magnitude.

No que tange o desenvolvimento do painel gerencial, objetivava-se desenvolver indicadores referentes aos dados de mercado e aos dados financeiros da empresa, de modo que fosse possível analisar resultados internos e seus reflexos no mercado como um todo.

Os dois primeiros indicadores representados na Figura 15, apresentam informações de faturamento e unidades vendidas acumuladas da empresa, definidos como indicadores financeiros. Em contrapartida, o terceiro indicador está relacionado ao desempenho da empresa ante as concorrentes, medido com a participação no mercado (*Market Share*), dentro da categoria em análise, definido como um indicador comercial.

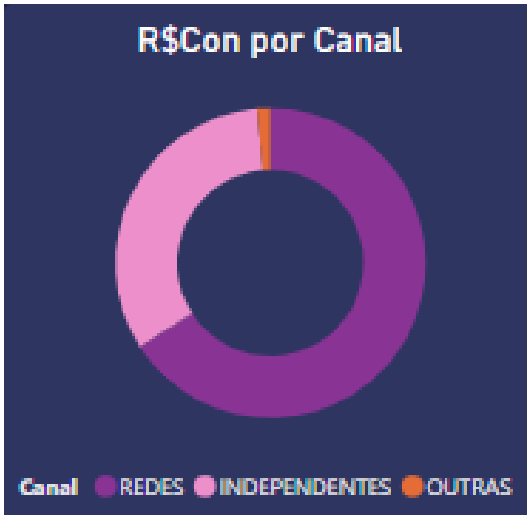
Figura 15 - Indicadores globais da empresa



Fonte: Autor (2022)

O segundo indicador presente no painel gerencial, representado na Figura 16, é relacionado aos canais de atendimento aos quais os dados estão relacionados. Esse indicador, também comercial, apresenta a porcentagem de quais canais estão tendo maior participação da empresa ou de suas concorrentes.

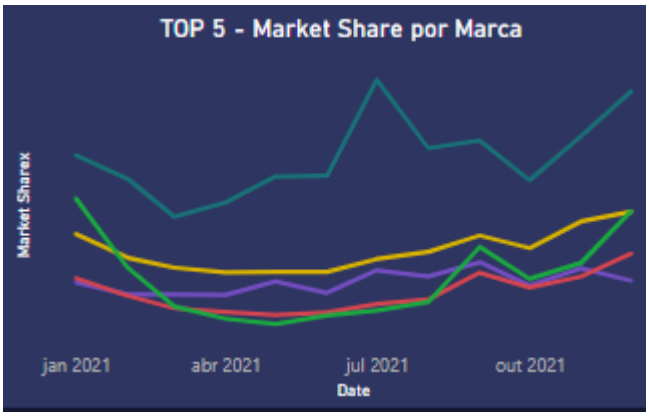
Figura 16 - Visão de faturamento por Canal



Fonte: Autor (2022)

O terceiro indicador, representado na Figura 17, é o principal dentre os indicadores comerciais. Ele representa as TOP 5 participações de *Market Share* de todos as marcas presentes no mercado, podendo ser analisada dentro dos agrupamentos de categoria. Dessa maneira, a empresa pode analisar o posicionamento de suas marcas ante às demais, se ganha ou perde parcela do mercado e, em caso de perda, para quem está perdendo. Ainda, o indicador apresenta as marcas com as quais a empresa deve se preocupar mais, possibilitando que novas estratégias de posicionamento possam ser traçadas.

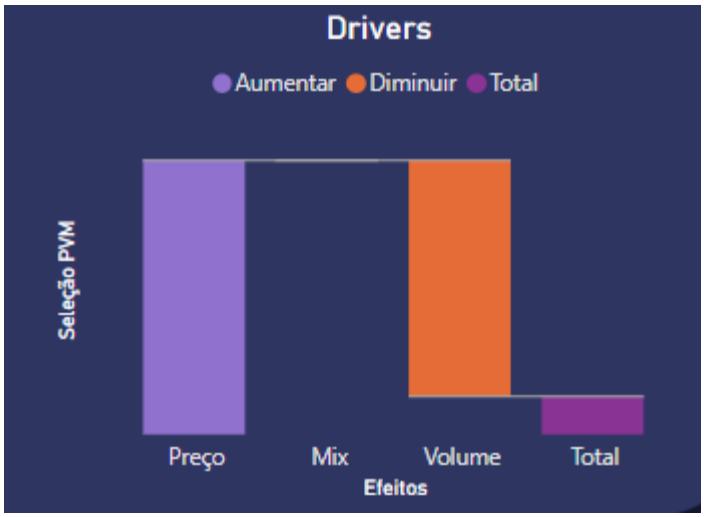
Figura 17 - Visão TOP 5 – *Market Share* por marca



Fonte: Autor (2022)

O quarto indicador, representado na Figura 18, é o último a compor os indicadores comerciais. Caracteriza-se pela apresentação dos drivers que afetam o faturamento da empresa. Como relatado anteriormente, o indicador é definido a fim de representar se o que mais impactou o desempenho de faturamento da empresa foi o preço, volume ou mix, de modo que seja possível aferir novas estratégias para o determinado produto. Reitera-se que, por ser considerado um indicador comercial, os gráficos trazem as mesmas informações quanto aos produtos das concorrentes, de modo que a empresa possa estar a par de como as outras marcas vêm se desenvolvendo no mercado e o que mais às afeta.

Figura 18 - Visão Drivers

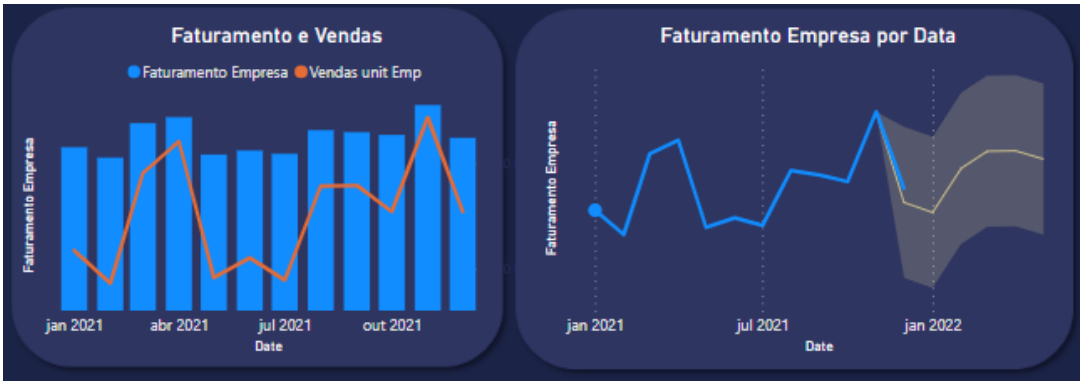


Fonte: Autor (2022)

Por fim, os demais indicadores financeiros são apresentados nas Figuras 19 e 20. Na Figura 19, são apresentados gráficos quanto a distribuição de faturamento e unidades vendidas e também a previsão de receita para os próximos seis meses. Com estas representações, a empresa pode analisar seu desenvolvimento ao longo dos meses e, ainda, identificar a projeção de seus resultados para o futuro.

O gráfico destacado no lado direito da Figura 19 representa a simulação do comportamento futuro do faturamento da empresa, com base na técnica de previsão previamente apresentada. Ressalta-se para essas representações, que a empresa não possuía, até então, estudos relacionados a projeções de resultados. Apesar de adotar um método relativamente simples, a previsão é capaz de apresentar padrões futuros e embasar as decisões de maneira mais assertiva.

Figura 19 - Visão de distribuição de faturamento e vendas e Visão contendo previsão



Fonte: Autor (2022)

Por fim, na Figura 20, o modelo apresenta a distribuição de faturamento da empresa para cada uma das regiões em que vende seu portifólio de produtos, adaptando os resultados de acordo com os filtros anteriormente apresentados, para cada uma das categorias. Dessa maneira, a empresa pode compreender quais os estados promotores e quais os detratores, possibilitando que sejam traçadas estratégias de recuperação ou potencialização da distribuição dos resultados. Com isso, é possível, ainda, analisar se as estratégias das equipes de venda de cada estado estão surtindo efeito para a potencialização da lucratividade da companhia.

Figura 20 - Visão Faturamento por estado

Fonte: Autor (2022)

A parir dos resultados demonstrados e, com base na validação dos resultados, percebe-se que o trabalho conseguiu atingir todos os objetivos apresentadas, de modo que a visualização dos resultados se desenvolveu de maneira clara e objetiva, como definido no início do estudo. Assim, conclui-se a efetividade dos processos e bom direcionamento quanto às suas aplicações.

5 CONCLUSÃO

Dado o grande volume de dados disponibilizados atualmente, mecanismos de análise robustos, unificados e visuais, se tornaram parte imprescindível para uma boa gestão de informações. A mineração de dados é primordial para o tratamento e solidificação concisa de bancos de dados extensos. Ainda, ressalta-se a importância de *Business Intelligence*, como aliado na criação de relatórios precisos para uma visualização completa e direcionada.

O presente trabalho teve como objetivo geral estabelecer, com base no modelo CRISP-DM de mineração de dados e *Business Intelligence*, um painel gerencial que unifique os principais indicadores financeiros e comerciais para as vinte principais categorias de produtos de uma multinacional no setor da saúde. O projeto desenvolveu as etapas do modelo de mineração de dados, CRISP-DM e concebeu um painel gerencial agrupando os indicadores financeiros e comerciais para empresa analisada.

Os objetivos específicos estabelecidos também foram atendidos, em que, o primeiro visava desenvolver, com base no modelo CRISP-DM, todo o processo de mineração de dados de vendas da empresa no canal farmacêutico. Todas as etapas do modelo foram desenvolvidas, levando em consideração o bom entendimento do que era necessário extrair como informação. Concomitante a este, para o segundo objetivo do trabalho, foram caracterizados os atributos de entrada necessários para que a previsão de faturamento da empresa fosse estabelecida, levando em consideração, principalmente, os quesitos de sazonalidade e intervalo de confiança.

O terceiro objetivo era o desenvolvimento dos cálculos e modelagens para os indicadores financeiros e comerciais. Para cada um dos indicadores presentes no painel gerencial apresentado, foram desenvolvidos cálculos e modelagens específicas, que pudessem alcançar resultados satisfatórios. Os indicadores financeiros levaram em consideração dados de faturamento, volume das vendas, categorias, estados e os períodos em que foram realizadas as leituras. Já os indicadores comerciais, tiveram como base, os dados das marcas da empresa analisada e de suas concorrentes, as categorias nas quais eram enquadradas, faturamento e volume de vendas para cada uma e os períodos em que foram realizadas as leituras.

O quarto objetivo, caracterizou-se pelo agrupamento dos resultados gerados na modelagem, unificando em um único painel, todos os indicadores desenvolvidos, além de sua concretização estética, para que as informações pudessem ser lidas clara e visualmente. Por fim, para o quinto e último objetivo, todos os resultados foram analisados e validados, com base em diferentes relatórios oficias da empresa e cálculos na própria base de dados, a fim de garantir a efetividade do modelo proposto.

Todos os objetivos estabelecidos para o projeto são capazes de agregar algum valor para a empresa, contudo, seus benefícios não se limitam, apenas, às entregas e análises apresentadas. O projeto, por apresentar diversos indicadores chave, possibilita que a tomada de decisão seja cada vez mais ágil, reduzindo tempo de resposta, além de potencializar o alinhamento estratégico, ampliando seus direcionamentos. A transparência de informações é ainda maior, permitindo às equipes da empresa um entendimento amplo dos resultados e, ainda, o reflexo das estratégias definidas. Por fim, ressalta-se a integração entre os departamentos, por se caracterizar com uma ferramenta acessível a todos, além de fomentar uma cultura focada em resultados.

No que tange os resultados do trabalho, estão, o desenvolvimento de indicadores de faturamento, nos quais a empresa pode se debruçar e realizar análises quanto a seu desempenho histórico e, ainda, quanto a projeção de seu faturamento, além de desenvolver análises específicas por regiões do país. Os indicadores financeiros mostram-se essenciais para que a empresa esteja atenta em relação a oscilações financeiras e as compare com diferentes cenários e suas respectivas implicações. Os indicadores comerciais, em contrapartida, apresentam o comportamento de seus concorrentes, contribuindo para que análises comparativas com outras marcas sejam realizadas de maneira eficiente. Ao analisar o indicador de *drivers,* é possível identificar quantitativamente o que mais impacta o produto em análise, possibilitando estratégias de recuperação para resultados detratores ou de potencialização para os promotores.

Em relação a aplicação do modelo CRISP-DM de mineração de dados, obtém-se como resultado, a concepção de um painel robusto e confiável, no qual se fundamenta em dados devidamente processados. O modelo se mostrou eficiente em todas as etapas, garantindo um processo de mineração completo e gerando informações fidedignas.

Ressalta-se, durante a aplicação do modelo, a identificação de problemas na base, em relação ao volume de vendas, da empresa e de suas concorrentes, para os produtos da categoria de analgésicos. Com isso, por considerar para alguns produtos a conversão em *blisters* e para outros em caixas, o volume de dados não apresentava uma unidade de medida padronizada, resultando, também, no preço unitário e faturamento dos produtos. Sem que esses problemas fossem identificados, as análises de diversos relatórios estariam sendo feitas de maneira equivocada, apresentando desempenhos significativamente maiores para alguns produtos que, na realidade, não eram factíveis. Dessa maneira, além de contribuir para os resultados apresentados na proposta de painel gerencial deste trabalho, a metodologia utilizada agregou valor para diversos outros projetos internos.

Em relação à qualidade dos resultados obtidos, reitera-se sua validação com base em comparações com os resultados apresentados em outros relatórios consolidados, caracterizados como fontes oficiais da empresa em análise. Para as validações, foram comparados os indicadores propostos com relatórios que apresentassem informações semelhantes, mas descentralizadas. Desse modo, foram necessárias comparações com mais de um relatório, a fim de estabelecer a acuracidade dos valores apresentados. Ainda, para os dados que não foram encontrados nos relatórios disponíveis, foram calculados diretamente na base, fazendo uso das mesmas formulações apresentadas no capítulo de desenvolvimento. Dessa maneira, após comparar os resultados, e identificar resultados coerentes, a proposta de painel foi validada.

Para trabalhos futuros de criação de painéis gerenciais baseados em mineração de dados, recomenda-se a aplicação de outras tarefas, como por exemplo a clusterização e associação. Para análises mais aprofundadas em relação aos estados, sugere-se a aplicação de algoritmos para identificar zonas homogêneas no território nacional. Ainda, para que questões como tratamento de dados sejam realizadas de maneira mais rápida e automatizada, sugere-se a construção, em linguagens de programação, como *Python*, de ferramentas de mineração de dados. Por fim, sugere-se a modelagem em diferentes ferramentas de BI disponibilizadas no mercado, a fim de se comparar seus desempenhos e escolher aquele que mais se adequa aos problemas a serem solucionados.

REFERÊNCIAS

ADRIAANS, P.; ZANTINGE, D. **Data mining addison wesley longman limited**. Edinbourgh Gate, Harlow, CM20 2JE, England, 1996.

AGGARWAL, C. C.; KONG, X.; GU, Q.; HAN, J.; PHILIP, S. Y. **Active learning: A survey. In: Data Classification**. Chapman and Hall/CRC, 2014.

AGRAWAL, R; IMIELINSKI, T; SWAMI, A. **Mining association rules between sets of items in large databases**. Proc. of the ACM SIGMOD, p. 207–216, 1993.

ATAMAN, M. Berk; VAN HEERDE, Harald J.; MELA, Carl F. **The long-term effect of marketing strategy on brand sales**. Journal of Marketing Research, v. 47, n. 5, p. 866-882, 2010.

AZEVEDO, Ana; SANTOS, Manuel Filipe. **KDD, SEMMA and CRISP-DM: a parallel overview**. IADS-DM, 2008.

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

CAMILO, Cássio Oliveira; SILVA, João Carlos da. **Mineração de dados: Conceitos, tarefas, métodos e ferramentas**. Universidade Federal de Goiás (UFC), v. 1, n. 1, p. 1-29, 2009.

CANUTO, OCTAVIANO; CLESIO, XAVIER. **Specialization and competitiveness in Brazilian foreign trade**. Revista Momento Económico, n. 119, 2000.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. **CRISP-DM 1.0: Step-by-Step Data Mining Guide**. SPSS (2000).

CIOS, K. J; PEDRYCZ, W; SWINIARSKI, R. W; KURGAN, L. A. **Data Mining – A Knowledge Discovery Approach**. Springer, 2007.

Corrêa, P. C., Oliveira, G. H. H., Botelho, F. M., Goneli, A. L. D., & Carvalho, F. M. **Modelagem matemática e determinação das propriedades termodinâmicas do café (Coffea arabica L.) durante o processo de secagem.** Revista Ceres, v. 57, p. 595-601, 2010.

DINO. **US$ 965 milhões em 2018: o promissor mercado dos dados requer integração**. [*S. l.*], 8 fev. 2018. Disponível em: https://www.Mundodomarketing. com.br/noticias-corporativas/conteudo/125458/us$-965-milhoes-em-2018-o-promissor-mercado-dos-dados-requer-integracao. Acesso em: 10 dez. 2021.

DRAPER, Norman R.; SMITH, Harry. **Applied regression analysis**. John Wiley & Sons, 1998.

DUARTE, João Carlos Assunção. **Dashboard Visual uma ferramenta de Business Intelligence**. Porto, 2012.

EXAME. **O segredo das empresas que sabem usar os dados a seu favor.** [S. l.], 1 dez. 2021. Disponível em: https://exame.com/inovacao/o-segredo-das-empresas-que-sabem-usar-os-dados-a-seu-favor/. Acesso em: 4 mar. 2022.

FAYYAD, U; PIATETSKY-SHAPIRO, G; SMYTH, P. **From Data Mining to Knowledge Discovery in Databases**. American Association for Artificial Intelligence, 1996.

FERRARI, Daniel Gomes; SILVA, Leandro Nunes de Castro. **Introdução a mineração de dados**. Saraiva Educaçao SA, 2017.

GIL, Antonio Carlos. **Métodos e técnicas de pesquisa social**. 6. ed. Ediitora Atlas SA, 2008.

GOLDSCHMIDT, Ronaldo; PASSOS, Emmanuel. **Data mining**. Gulf Professional Publishing, 2005.

HAN, Jiawei; PEI, Jian; KAMBER, Micheline. **Data mining: concepts and techniques**. Elsevier, 2011.

HAND, D; MANNILA, H; SMYTH, P. **Principles of Data Mining**. MIT Press, 2001.

HOAGLIN, David C.; MOSTELLER, Frederick; TUKEY, John W. **Análise exploratória de dados técnicas robustas: um guia**. 1992.

INTAN, Rolly; YENTY, Oviliani. **Mining multidimensional fuzzy association rules from a normalized database. In: 2008 International Conference on Convergence and Hybrid Information Technology**. IEEE, 2008.

JANSSEN, M.; VOORT, H, V der.; WAHYUDI, A. **Factors influencing big data decision-making quality**. Journal of Business Research, v. 70, p. 338-345, 2016.

KAIZEN INSTITUTE. **A importância de ser VISUAL**. [S. l.], 28 ago. 2016. Disponível em: https://br.kaizen.com/blog/post/2016/09/28/a-importancia-de-ser-visual. Acesso em: 17 dez. 2021.

KOH, Yun Sing; RAVANA, Sri Devi. **Unsupervised rare pattern mining: a survey. ACM Transactions on Knowledge Discovery from Data (TKDD)**, v. 10, n. 4, p. 1-29, 2016.

KOSHY, Elizabeth; KOSHY, Valsa; WATERMAN, Heather. **Action research in healthcare**. Sage, 2010.

LAGO, Karine; ALVES, **Laennder. Dominando o power BI.** São Paulo: DataB, v. 1, 2018.

LAROSE, D. T. **Discovering Knowledge in Data: An Introduction to Data Mining.** John Wiley and Sons, Inc, 2005.

LAUREANO, Raul; CAETANO, Nuno; CORTEZ, Paulo. **Previsão de tempos de internamento num hospital português: aplicação da metodologia CRISP-DM**. 2014.

LIMA, Rafaela Somavila. **Criação de projeto de ciência de dados utilizando a metodologia CRISP-DM em conformidade com a LGPD**. 2021.

MAROCO, João. **Consistency and efficiency of ordinary least squares, maximum likelihood, and three type II linear regression models: A Monte Carlo simulation study.** Methodology: European Journal of Research Methods for the Behavioral and Social Sciences, v. 3, n. 2, p. 81, 2007.

MATTAR, F.N. **Pesquisa de Marketing**. São Paulo: Atlas, 1998.

MCKINSEY GLOBAL INSTITUTE. **The age of analytics: Competing in a data-driven world.** [*S. l.*], 7 dez. 2016. Disponível em: https://www.mckinsey.com/business-functions/quantumblack/our-insights/the-age-of-analytics-competing-in-a-data-driven-world. Acesso em: 8 nov. 2021.

MICROSOFT. **What is Power BI?**. [*S. l.*], 28 jan. 2022. Disponível em: https://docs.microsoft.com/en-us/power-bi/fundamentals/power-bi-overview. Acesso em: 2 jun. 2022.

NAGAR, P., ATRIWAL, L., MEHRA, H.; TAYAL, S. **Comparison of generalized and big data business intelligence tools**. In: 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom). IEEE, p. 3585-3588, 2016.

NIJS, V., MISRA, K., ANDERSON, E. T., HANSEN, K.; KRISHNAMURTHI, L. **Channel pass-through of trade promotions**. Marketing Science, v. 29, n. 2, p. 250-267, 2010.

OUYANG, Weimin. **Mining positive and negative fuzzy association rules with multiple minimum supports. In: 2012 International Conference on Systems and Informatics (ICSAI2012)**. IEEE, 2012.

PARMENTER, David. **Key performance indicators: developing, implementing, and using winning KPIs**. John Wiley & Sons, 2015.

PEACOCK, Peter R. **Data mining in marketing: Part 1**. Marketing Management, v. 6, n. 4, p. 8, 1998.

PEREIRA, J.C.R. **Análise de Dados Qualitativos**. São Paulo: Edusp/Fapesp, 1999.

PINTO, Vladimir Steffen. **Desenvolvimento de solução CRISP-DM para classificação do evento prisão de coluna no processo de perfuração de poços offshore**. 2018.

RAHPEYMAI, Neda. **Data Mining with Decision Trees in the Gene Logic Database: A Breast Cancer Study**. Institutionen för datavetenskap, 2002.

REGINATO, Luciane; NASCIMENTO, Auster Moreira. **Um estudo de caso envolvendo Business Intelligence como instrumento de apoio à controladoria**. Revista Contabilidade & Finanças, v. 18, p. 69-83, 2007.

REZENDE, Denis Alcides. **Sistemas de informações organizacionais**. São Paulo: Atlas, 2005.

SCALABRIN BIANCHI, Isaías; DINIS SOUSA, Rui; PEREIRA, Ruben. **Information technology governance for higher education institutions: A multi-country study. In: Informatics**. MDPI, p. 26, 2021.

SCHRÖER, Christoph; KRUSE, Felix; GÓMEZ, Jorge Marx. **A systematic literature review on applying CRISP-DM process model**. Procedia Computer Science, v. 181, p. 526-534, 2021.

SHEARER, C. **The CRISP-DM Model: The New Blueprint for Data Mining**. Journal of Data Warehousing, v. 5, n. 4, p. 13-22, 2000.

SILVA, Edna Lúcia da; MENEZES, Estera Muszkat. **Metodologia da pesquisa e elaboração de dissertação.** 2001.

SRINIVASAN, Shuba; LESZCZYC, Peter TL Popkowski; BASS, Frank M. **Market share response and competitive interaction: The impact of temporary, evolving and structural changes in prices**. International journal of research in marketing, v. 17, n. 4, p. 281-305, 2000.

THOMPSON, Mark PA; WALSHAM, Geoff. **Placing knowledge management in context.** Journal of Management Studies, v. 41, n. 5, p. 725-747, 2004.

TRIPP, David. **Pesquisa-ação: uma introdução metodológica.** Educação e pesquisa, v. 31, p. 443-466, 2005.

TURBAN, E., KING, D. R., LANG, J., LAI, L. **Introduction to electronic commerce**. 2009.

TURBAN, E., SHARDA, R., ARONSON, J. E., KING, D. **Business intelligence: um enfoque gerencial para a inteligência do negócio**. Bookman Editora, 2009.

UBER, JOSÉ LINO. **Descoberta de conhecimento com o uso de text mining aplicada ao SAC.** Universidade Regional de Blumenau. Centro de Ciências Exatas e Naturais, 2004.

WEDEL, Michel; KANNAN, P. K. **Marketing analytics for data-rich environments**. Journal of Marketing, v. 80, n. 6, p. 97-121, 2016.

WILBUR, Kenneth C.; FARRIS, Paul W. **Distribution and market share**. Journal of Retailing, v. 90, n. 2, p. 154-167, 2014.

WIXOM, B., ARIYACHANDRA, T., GOUL, M., GRAY, P., KULKARNI, U., PHILLIPS-WREN, G. **The current state of business intelligence in academia.** Communications of the Association for information Systems, v. 29, n. 1, p. 16, 2011.