UNIVERSIDADE FEDERAL DE SANTA CATARINA
CENTRO TECNOLÓGICO
DEPARTAMENTO DE ENGENHARIA DE PRODUÇÃO E SISTEMAS
CURSO DE GRADUAÇÃO EM ENGENHARIA DE PRODUÇÃO CIVIL

Natalia Tomazi

Análise Preditiva de dados numa empresa B2B no setor agroalimentar

Florianópolis
2022

Natalia Tomazi

Análise Preditiva de dados numa empresa B2B no setor agroalimentar

Trabalho Conclusão do Curso de Graduação em Engenharia de Produção Civil do Centro Tecnológico da Universidade Federal de Santa Catarina como requisito para a obtenção do título de Engenheira Civil, habilitada em Produção.
Orientador: Prof. Mauricio Uriona Maldonado.

Florianópolis
2022

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

Tomazi, Natalia
Análise Preditiva de dados numa empresa B2B no setor agroalimentar / Natalia Tomazi ; orientador, Mauricio Maldonado, 2022.
82 p.

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

Inclui referências.

1. Engenharia de Produção Civil. I. Maldonado, Mauricio. II. Universidade Federal de Santa Catarina. Graduação em Engenharia de Produção Civil. III. Título.

Natalia Tomazi

Análise Preditiva de dados numa empresa B2B no setor agroalimentar

Este Trabalho Conclusão de Curso foi julgado adequado para obtenção do Título de Engenheira Civil habilitada em Produção e aprovado em sua forma final pelo Curso de Graduação Engenharia de Produção Civil.

Florianópolis, 17 de março de 2022.

Documento assinado digitalmente
Monica Maria Mendes Luna
Data: 21/03/2022 08:23:37-0300
CPF: 314.104.993-91
Verifique as assinaturas em https://v.ufsc.br

Profa. Mônica Maria Mendes Luna, Dra.
Coordenador do Curso

Banca Examinadora:

Documento assinado digitalmente
Mauricio Uriona Maldonado
Data: 19/03/2022 09:58:59-0300
CPF: 010.972.939-04
Verifique as assinaturas em https://v.ufsc.br

Prof. Mauricio Uriona Maldonado, Dr.
Orientador
Universidade Federal de Santa Catarina

Documento assinado digitalmente
Ricardo Villarroel Davalos
Data: 21/03/2022 08:43:25-0300
CPF: 932.328.619-20
Verifique as assinaturas em https://v.ufsc.br

Prof. Ricardo Villarroel Dávalos, Dr.
Avaliador
Universidade Federal de Santa Catarina

Documento assinado digitalmente
Cosme Polese Borges
Data: 19/03/2022 16:56:20-0300
CPF: 081.725.299-10
Verifique as assinaturas em https://v.ufsc.br

Cosme Polese Borges
Avaliador
Universidade Federal de Santa Catarina

Este trabalho é dedicado aos meus queridos pais e irmão.

AGRADECIMENTOS

Agradeço inicialmente aos meus pais, Jucirene e Carlos, que sempre apoiaram incondicionalmente todas as etapas do meu desenvolvimento pessoal e profissional, e estiveram ao meu lado tornando essa caminhada mais segura, leve e confortável. Também agradeço ao meu irmão Gustavo, com quem tenho o prazer de compartilhar o gosto pela Engenharia de Produção, pelo apoio, acolhimento e conhecimentos divididos.

Agradeço a Universidade Federal de Santa Catarina pela oportunidade de estar em contato com grandes profissionais, por me permitir acessar um ensino de qualidade e por contribuir para o meu crescimento como cidadã. Agradeço ao Movimento Empresa Júnior e ao PET Engenharia de Produção por me proporcionarem aprendizados, experiências e amizades que levarei para sempre comigo.

Agradeço ao meu orientador Maurício, por todo o ensinamento, atenção, disponibilidade, confiança e auxílio na construção desse estudo. Agradeço a empresa objeto de análise desse trabalho, a qual tem me proporcionado grande desenvolvimento e onde inicio minha jornada profissional.

Agradeço a todos as amizades que me acompanharam durante o período de graduação, que me incentivaram, apoiaram, entenderam minhas ausências, compartilharam momentos inesquecíveis e me inspiraram. Em especial, agradeço aos amigos Maria Cristina, Ivana, Nildo e Vinícius que com palavras de apoio foram imprescindíveis durante o processo de construção deste trabalho.

Sem dados, você é apenas mais uma pessoa com uma opinião. (DEMING, 1986)

RESUMO

O presente trabalho se desenvolve a partir da necessidade de uma empresa do setor de tecnologia de aprimorar a tomada de decisão no setor comercial. Ao longo do seu tempo de operação, a instituição vem armazenando uma grande quantidade de dados de negociações comerciais, realizando análises de cunho tradicional a partir deles. Dessa forma, o objetivo deste trabalho é apresentar, por meio do Design Science Research, o aprimoramento da tomada de decisão no processo de vendas da empresa por meio da análise preditiva de dados. Inicialmente, apresenta-se uma revisão bibliográfica sobre os principais conceitos acerca do Business Intelligence, Data Science e técnicas aplicadas ao Machine Learning. A elucidação do objetivo do estudo se dá por meio da aplicação das fases de um projeto de Data Science. Assim, após a delimitação do escopo de trabalho, segue-se a coleta e análise exploratória dos dados, chegando à aplicação dos métodos de Machine Learnig para modelagem dos dados. Para tal, a linguagem de programação R é empregada na clusterização para a segmentação dos clientes, na regressão logística com o intuito de identificar os determinantes do sucesso de uma venda, e no random forest para possibilitar a estimação do tempo de negociação. Como resultados tem-se uma clusterização que forma três grupos de clientes, um modelo de regressão logística com precisão de 76% e um modelo random forest com RMSE de 4 dias, além da identificação dos atributos que mais impactam no processo de vendas. Em adição ao alcance dos objetivos específicos traçados, a implementação da análise preditiva de dados do setor comercial gerou insumos para uma tomada de decisão mais orientada a dados.

Palavras-chave: Business Intelligence. Data Science. Clusterização. Regressão Logística. Random Forest. Vendas.

ABSTRACT

The present work is developed from the need of a company in the technology sector to improve decision making in the commercial sector. Throughout its time of operation, the institution has been storing a large amount of data from commercial negotiations, performing traditional analyzes from them. In this way, the objective of this work is to present, through Design Science Research, the improvement of decision making in the company's sales process through predictive data analysis. Initially, a literature review is presented on the main concepts about Business Intelligence, Data Science and techniques applied to Machine Learning. The elucidation of the objective of the study takes place through the application of the phases of a Data Science project. So, after defining the scope of work, the collection and exploratory analysis of the data follows, reaching the application of Machine Learning methods for data modeling. For that, the R programming language is used in clustering for customer segmentation, in logistic regression in order to identify the determinants of the success of a sale, and in the random forest to enable the estimation of negotiation time. In the results and discussions section, the quality of the modeling is evaluated, concluding the statistical significance of the built models. Also, the main points of interpretation are exposed together with the final results of the models. In addition to reaching the specific objectives outlined, the implementation of predictive analysis of data in the commercial sector generated inputs for a more data-oriented decision making.

Keywords: Business Intelligence. Data Science. Clustering. Logistic Regression. Random Forest. Sales.

LISTA DE FIGURAS

LISTA DE QUADROS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

ABRAS Associação Brasileira de Supermercados

AUC *Area Under the Curve* (Área sob a curva)

BI *Business Intelligence* (Inteligência de Negócios)

B2B *Business To Business* (De Empresa Para Empresa)

CLV *Customer Lifetime Value* (Valor do Ciclo de Vida do Cliente)

CRM *Customer Relationship Management* (Gestão de Relacionamento com o Cliente)

FP Falso Positivo

FN Falso Negativo

ML *Machine Learning* (Aprendizado de Máquina)

RMSE *Root Mean Squared Error* (Raiz do Erro Médio Quadrado)

ROC *Receiver Operating Characteristic Curve* (Curva Característica de Operação do Receptor)

R2 R-quadrado

SaaS *Software as a Service* (Software como Serviço)

VN Verdadeiro Negativo

VP Verdadeiro Positivo

SUMÁRIO

1 INTRODUÇÃO

Este capítulo tem como objetivo contextualizar o tema abordado no estudo, justificar a pesquisa, definir seus objetivos e apresentar a estrutura do trabalho.

1.1 CONTEXTUALIZAÇÃO

Segundo Becker (2002), um dos mercados que possui grande quantidade de soluções tecnológicas a serem exploradas é o alimentar. Desde a produção no campo, passando pela indústria de alimentos, até a disponibilização da refeição ao consumidor final, todos os agentes podem se beneficiar de alguma forma de soluções inovadoras. Ainda, apesar dessa extensa cadeia de atores possuir características e necessidades diferentes, é fundamental o seu trabalho em conjunto no uso de tecnologias para garantir a qualidade e segurança do alimento, assim como o atendimento a legislações como a Instrução Normativa Conjunta 02/2018, que trata da obrigatoriedade da rastreabilidade de alimentos (NACIONAL, 2018). Assim, a produção e comércio de alimentos conseguem se beneficiar do uso de produtos na forma de Programas como Serviço (do inglês, *Software As A Service [SaaS]*), construídos com o objetivo de atender as necessidades do segmento.

No que tange a contratação dessas soluções, as vendas de Empresa para Empresa (do inglês, *Business to Business [B2B]*), representam as parcerias comerciais construídas entre empresas - fornecedores e usuários da ferramenta SaaS. De acordo com Araújo (2019), nesse tipo de relação os clientes não buscam satisfazer as suas próprias necessidades, e sim atender a uma demanda direta ou indireta de seus consumidores finais. Ademais, outra característica é o envolvimento de mais pessoas durante o processo de vendas, uma vez que o usuário da solução pode não ser o decisor final da compra na empresa. Esses aspectos, além de alongarem o período de negociação, influenciam fortemente na taxa de sucesso do negócio, tornando as vendas B2B mais complexas do que aquelas direcionadas diretamente às pessoas físicas. Por esse motivo, uma prática comum nesse setor é a segmentação do processo de vendas em algumas etapas, como a prospecção e qualificação do cliente em potencial, oferta, negociação e fechamento da proposta comercial e por fim o pós-venda (ROSS; TYLER, 2017). Por conta desse desenho processual e do envolvimento de diferentes equipes em uma mesma transação comercial, o registro de informações se torna algo fundamental para o alinhamento das partes e sucesso da negociação.

A partir dessa prática, e pelo fato de registros bem detalhados beneficiarem o relacionamento com os clientes a longo prazo, é comum que as empresas acumulem uma grande quantidade de dados. Nesse sentido, segundo Atkins et al (2016) há muitos exemplos no segmento das vendas em que a análise de dados está gerando melhorias em crescimento, eficiência e eficácia, entretanto a maioria das organizações ainda não considera que faz um uso adequado desse potencial estratégico.

No que tange a empresa alvo de estudo do presente trabalho, esta se caracteriza por seu viés tecnológico ao oferecer soluções voltadas para a cadeia agroalimentar, conforme é apresentado na Figura 1. Assim, por meio de programas (do inglês, *softwares*) que atendem a demandas de rastreabilidade e de gestão da qualidade, seus principais clientes são produtores, distribuidores e indústrias de alimentos, e também varejos supermercadistas. Com ampla atuação no Brasil e uma crescente de trabalho internacional, a empresa conta com uma área dedicada a vendas, que em conjunto com a área de marketing, se torna responsável pela geração de receitas da organização.

Figura 1 - Clientes da cadeia agroalimentar

Fonte: Elaborado pelo autor.

1.1.1 Descrição do Problema

A partir da vasta atuação da empresa ao longo da cadeia agroalimentar, o aspecto de diferenciação entre os diversos *leads* se destaca, exigindo uma adaptação no processo de negociação para cada perfil de cliente. Assim, durante a fase inicial do processo de vendas surge a necessidade de conhecer melhor as características dos potenciais clientes, para que seja possível o desenho de estratégias especificas de abordagem.

Já ao se considerar a conotação gerencial do setor comercial e seu impacto nas demais áreas da empresa, como financeiro e operação, se torna essencial o planejamento de vendas, considerando a perspectiva de período e taxa de sucesso da negociação. Dessa forma, fica evidente a utilidade de uma estrutura que permita prever a probabilidade de sucesso e o ciclo de uma venda. Além disso, para que a negociação possa ser otimizada e gere melhores resultados, há a demanda de direcionamento de recursos para os fatores que mais impactam no sucesso da venda, exigindo que se tenha clareza acerca desses aspectos.

Nos últimos anos o registro de informações relacionadas aos clientes e negociações comerciais foi intensificado por meio do uso de ferramentas específicas, gerando um acúmulo de dados no setor. Por esse motivo, a empresa tem buscado formas de fazer um uso mais estratégico desses dados, com o objetivo de tornar mais eficiente o trabalho dos envolvidos e maximizar a receita da organização. Portanto, considerando características como probabilidade de sucesso da venda e tempo de negociação, a pergunta de pesquisa que se pretende responder é como aprimorar a tomada de decisão para apoiar o processo de vendas?

1.2 OBJETIVOS

Nas seções abaixo estão descritos o objetivo geral e os objetivos específicos deste Trabalho de Conclusão de Curso.

1.2.1 Objetivo Geral

Aprimorar a tomada de decisão no processo de vendas de uma empresa de tecnologia por meio da análise preditiva de dados.

1.2.2 Objetivos Específicos

- Realizar segmentação de clientes.
- Identificar os determinantes do sucesso de uma venda.
- Estimar o tempo de negociação de uma venda.

1.3 JUSTIFICATIVA

A partir da crescente exigência dos mercados, com destaque para as vendas B2B, a complexidade e robustez das negociações comerciais tem aumentado o volume de dados nas empresas. Se por um lado esse domínio de informações pode significar sucesso, de outro há a dificuldade no tratamento e tomada de decisão (SILVA; SAMBONGO; NOÉ, 2016). Neste sentido, o uso de conhecimento extraído de dados para a tomada de decisão pode gerar benefícios como a redução de custos e tempo, insumo para desenvolvimento de novos produtos e melhoria de qualidade (CARDOZO, 2021)**.** Por isso, a determinação de uma sistemática de análise de dados que leve à melhoria de processos se torna fundamental no contexto atual das organizações.

Em relação ao uso de dados em empresas, segundo Guazzelli (2012) a abordagem tradicional consiste na manipulação das informações registradas de forma a compreender observações passadas, e com isso usar da experiência humana para tentar melhorar eventos futuros. Por outro lado, a análise preditiva usa de registros de acontecimentos passados para criar modelos que possam ser empregados no entendimento de tendências e estimação de probabilidades futuras. Ou seja, a análise preditiva de dados possibilita maior isenção de julgamento humano em tomadas de decisão, contribuindo assim para a cultura de orientação a dados da organização.

Quanto ao foco de utilização das informações, ao se tratar do setor comercial de uma empresa, o início do processo de vendas se dá com as primeiras interações geradas pelo marketing com o potencial cliente (do inglês, *lead)*. Então, garantir que esse primeiro contato, assim como toda a relação de negociação, seja adaptada ao perfil do *lead*, é fundamental para contribuir para o sucesso da relação e consequente venda. Nesse sentido, a técnica de agrupamento (clusterização) possibilita maior entendimento das características desses clientes em potencial.

Já no que tange o impacto das negociações nas áreas comercial, financeira e de operação, é de extrema importância que se possa prever o volume e quando novas vendas serão efetuadas, uma vez que esse tipo de previsão permite o planejamento dos setores. É a partir desse contexto que é justificado o uso dos métodos de regressão para maior entendimento dos parâmetros de sucesso da negociação e ciclo de venda.

Nesse sentido, a aplicação de técnicas relacionadas a Engenharia de Produção, como a clusterização, a regressão logística e a floresta aleatória (do inglês, *random forest*), proporciona

os meios capazes para construção de modelos que apoiam a tomada de decisão. Dessa forma, busca-se converter dados em ideias estratégicas para o ganho de eficiência e sucesso de negociações, aumentando assim a geração de receita.

No que se relaciona a empresa objeto de estudo, tendo em vista a já utilização tradicional de dados no processo de gestão, a aplicação de técnicas de Aprendizado de Máquina (do inglês, *Machine Learning [ML]*) visa refinar e aperfeiçoar a tomada de decisão orientada a dados. Nesse sentido, a melhoria de processos que permitam maior produtividade, e consequente lucratividade, vai ao encontro do objetivo da área comercial estabelecido em planejamento estratégico, que é aumentar o volume de novas vendas de forma a contribuir para o faturamento saudável da organização.

De acordo com Guazzelli (2012)**,** empresas que atingem um alto nível de maturidade analítica possuem maior geração de receita, maior valorização de mercado e maior lucratividade do que aquelas sem processo estabelecidos de uso de dados. Em conclusão, ao se considerar o contexto de vendas, o estudo realizado se torna pertinente não apenas para a organização em questão, mas também para empresas de tecnologia no geral que possuem uma estruturação comercial.

1.4 ESTRUTURA DO TRABALHO

O presente trabalho é composto por seis capítulos que visam explorar a contextualização e desenvolvimento de uma solução acerca do uso de dados no setor comercial de uma empresa de tecnologia. De forma inicial, a atual seção aborda as circunstâncias da problemática, apresentando os objetivos e justificativas do estudo.

Já o capítulo dois expõe a fundamentação teórica necessária para entendimento do trabalho. São apresentados conceitos acerca da Inteligência de Negócios (do inglês, *Business Intelligence [BI]*), como dados, inteligência competitiva e aplicação do BI em vendas, um aprofundamento sobre a temática Ciência de Dados (do inglês, *Data Science*), através da análise preditiva e *Machine Learning*, e por fim a caracterização das técnicas de *ML* e seus métodos avaliativos a serem adotados, no contexto da clusterização, regressão logística e *random forest*.

A terceira seção é responsável por exibir a metodologia empregada no estudo, isso por meio do enquadramento metodológico, da delimitação de materiais e métodos e do detalhamento dos procedimentos metodológicos a serem adotados. Em seguida, o capítulo quatro, após delimitar o escopo de trabalho, discorre sobre a exploração dos dados.

Posteriormente, a seção número cinco aborda os pormenores da construção dos modelos preditivos e da clusterização, apresentando a avaliação dessas modelagens e seus principais resultados. Por fim, o sexto capítulo se dá por meio das considerações finais do trabalho, apresentando os desafios e sugestões de estudos futuros.

2 FUNDAMENTAÇÃO TEÓRICA

Este capítulo tem como objetivo construir uma fundamentação teórica sobre o *Business Intelligence*, ressaltando aspectos da formação do conhecimento, da inteligência competitiva e de sua aplicação no processo de vendas, sobre o *Data Science*, conceituando a análise preditiva e o *Machine Learning*, e também sobre as técnicas de clusterização, regressão logística e *random forest* que serão empregadas na base de dados em estudo. Assim, espera-se expor a teoria relacionada à pesquisa e contextualizá-la em meio à literatura referência da área.

2.1 BUSINESS INTELLIGENCE

Amplamente divulgado nos últimos tempos, o conceito de *Business Intelligence* compreende um conjunto de termos como arquitetura, ferramentas, banco de dados, aplicações e metodologias de acordo com Turban et al. (2009)**.** Ainda segundo o autor, o principal objetivo do *BI* é permitir o acesso e manipulação de dados, de forma a fornecer aos interessados insumos suficientes para análises que embasem melhores decisões. Em complemento, Bezerra (2015) trata o *Business Intelligence* como a utilização de diferentes fontes de informações que por meio da análise, exploração e apresentação irão determinar estratégias e tomadas de decisões.

Diante da grande quantidade de informações que são geradas e se tornam disponíveis por meio de registros, o ambiente corporativo enfrenta dificuldades na manipulação e utilização desses para a tomada de decisão. Neste sentido, o BI reúne o conjunto de técnicas que fornece às empresas o conhecimento para a adoção de ações estratégicas (MIKROYANNIDIS; THEODOULIDIS, 2010)**.** Em consonância com os autores citados, Negash e Gray (2008) abordam o BI como uma prática de coleta e armazenamento de dados e gerenciamento de conhecimento para o suporte ao processo de decisão.

Nesta seara, o conceito de inteligência é relacionado ao *Business Intelligence* ao se explorar o uso de dados de forma competitiva pelas organizações. A inteligência é apresentada com o significado de redução de um grande volume de informações em conhecimento, que é filtrado, analisado e utilizado de forma estratégica pelas empresas (KALAKOTA; ROBINSON, 2001)**.** Ainda, segundo Lima e Souza (2003), a inteligência é definida como um processo de coleta, análise e disseminação de informação precisa, relevante, específica, atual e relacionada com a empresa, o ambiente em que está inserida e seus competidores.

Em adição ao estudo tradicional da temática, as autoras Petrini, Pozzebon e Freitas (2004) propõem uma visão mais social e política ao definir BI como um processo construído coletivamente de captura, análise e geração de conhecimento. Como resultado, a informação retida possui menor volume e se torna mais estratégica, pertencente a uma vasta contextualização, interna e externa às organizações.

A partir deste aspecto coletivo do uso de dados, o sucesso de um sistema de BI está ligado ao fato de ele ser vantajoso para a empresa como um todo (TURBAN et al., 2009). Ou seja, apesar das diferentes abordagens utilizadas em níveis estratégicos e táticos, o investimento na implantação e manutenção deve ser percebido como benéfico por toda a organização. Dessa maneira, o BI não ficará restrito ao departamento técnico, e sim servirá como forma de transformar a maneira como são conduzidas as operações do negócio, tornando as tomadas de decisões orientadas a dados.

Além da via gerencial, o BI representa o conjunto de tecnologias utilizadas para gerar a inteligência empresarial. Para Freitas et al. (2001) visualizar a informação tem o objetivo de representar graficamente os dados de modo que seja possível gerar conhecimento. Assim, as ferramentas que compõem o sistema BI proporcionam uma ampla visão do negócio, possibilitando o cruzamento de informações e uma uniforme disseminação dos dados, se tornando essenciais para a tomada de decisão e uma boa gestão organizacional (TURBAN et al., 2009).

Ao se tratar da utilidade do sistema em uma empresa, Abukari e Jog (2003) evidenciam alguns aspectos que determinam o sucesso da implementação do BI, tais como:

A. Conectar as problemáticas a serem endereçadas com os objetivos estratégicos da organização.
B. Considerar as fontes de dados já existentes antes de criar novas origens de informações;
C. Garantir a consistência da extração, transformação e armazenagem dos dados;
D. Escolher as ferramentas envolvidas de acordo com as necessidades da empresa;
E. Criar relatórios padrões, porém permitir análises sob demanda;
F. Planejar o sistema de forma que os tomadores de decisão tenham as informações adequadas.

2.1.1 Dado, informação e conhecimento

Para Davenport (1998), a compreensão da gestão da informação passa pela definição e distinção dos conceitos dado, informação e conhecimento. Segundo o autor, dados são um conjunto de registros realizados através de transações organizacionais, que podem ser estruturados, capturados, quantificados e transferíveis. Em adição, Marinheiro (2013) define os dados como fatos isolados que sozinhos não têm qualquer utilidade, estes precisam ser lapidados de forma a permitir a sua compreensão.

Após a manipulação, organização, consolidação e atribuição de propósito, os dados se transformam em informação (SORDI, 2015)**.** Por exigir uma análise e intermediação humana, de acordo com Davenport (1998) a informação gerada está sujeita à interpretação de seu criador, sendo influenciada pela realidade em que está inserida. Ou seja, as informações precisam ser discutidas dentro de um contexto para auxiliar questões específicas.

Da mesma forma que a informação é um desenvolvimento dos dados, o conhecimento é derivado das informações. Ainda segundo Davenport (1998), o conhecimento possui natureza intuitiva uma vez que é uma tradução de informação contextualizada, valores e experiências acumuladas. De forma sucinta, o autor apresenta as definições desses termos conforme o Quadro 1.

Quadro 1 - Definição de dado, informação e conhecimento

Dado	Informação	Conhecimento
Simples observações sobre o estado do mundo. • Facilmente estruturado; • Facilmente obtido por máquinas; • Frequentemente quantificado; • Facilmente transferível.	Dados dotados de relevância e propósito. • Requer unidade de análise; • exige consenso em relação ao significado; • Exige necessariamente a medição humana.	Informação valiosa da mente humana. • Inclui reflexão, síntese, contexto; • De difícil estruturação; • De difícil captura em máquinas; • Frequentemente tácito; • De difícil transferência.

Fonte: (DAVENPORT, 1998)

Após as evidências amadurecerem pelos estágios de dados, informação e conhecimento, o usuário que possui afinidade com o tema, experiência, e senso analítico é capaz de realizar o processo decisório, conforme evidenciado pela pirâmide do conhecimento sugerida por

Marinheiro (2013) Figura 2. Assim, o BI cumpre o seu papel dentro de uma organização que busca competitividade por meio da orientação a dados.

Figura 2 - Pirâmide do Conhecimento

Fonte: Adaptado (MARINHEIRO, 2013)

2.1.2 Inteligência Competitiva

Em adição ao termo inteligência, que trata de forma geral da transformação de dados em conhecimento, tem-se o conceito de competitividade. De acordo com Kupfer (2015), competitividade é a adequação das estratégias internas das empresas ao padrão de concorrência vigente no mercado. Assim, a inteligência competitiva trata da utilização de dados, passando pelo seu tratamento e desenvolvimento, para a definição de estratégias que melhorem o posicionamento da organização no ambiente em que está inserida. Segundo Hilsdorf (2018):

> Inteligência competitiva é uma forma proativa de captar e organizar informações relevantes sobre o comportamento da concorrência, mas também dos clientes e do mercado como um todo, analisando tendências e cenários, e permitindo um melhor processo de tomada de decisão no curto e longo prazo.

Por esse ângulo nasce o conceito de Organização Inteligente, que refere-se às empresas que usam BI para tomada de decisões mais rápidas e estratégicas que seus competidores (PETRINI; POZZEBON; FREITAS, 2004). Isso se dá por meio da utilização de registros, interno e externos à organização, que a aproximam de seus clientes e a afastam de seus

concorrentes. Ainda, de acordo com Turban et al. (2009), diante do ambiente de negócios atual, o acesso e utilização da maneira correta de informação de qualidade não trata apenas de aumento de lucratividade, e sim da sobrevivência da instituição.

Em contrapartida, um estudo exposto pelo Grupo Gartner (HOSTMANN; RAYNER; HERSCHEL, 2009), evidenciou que apesar de muitas empresas possuírem elementos de BI instalados, algumas ainda não são capazes de extrair todos os benefícios desse sistema. Com esse comportamento, muitas organizações estão armazenando uma grande quantidade de dados sem de fato consumi-los de forma estratégica. A partir disso, fica evidente a necessidade e importância do BI para que a informação certa seja entregue a pessoa certa e no momento certo (LIMA; DE SOUZA, 2003).

2.1.3 Business Intelligence em Vendas

A partir de desafios enfrentados pelas organizações, o *Business Intelligence* tem como objetivo dar respostas às perguntas de executivos e gestores, contribuindo de forma geral para a melhoria do rendimento do negócio. Nessa seara, o sistema tem aplicação em diversas áreas da empresa, bastando apenas uma problemática que oriente a busca de dados.

Nesse sentido, o setor comercial apresenta aspectos de operação que podem se beneficiar do uso do BI, como para identificar padrões de comportamento de clientes, entender a produtividade e capacidade da equipe, analisar padrões históricos de ciclo de vendas, sazonalidade e volume de receita e para se posicionar em comparação com concorrentes (LEADS2B, 2021).

Em adição, a Plataforma Data Science Academy (2021) apresenta os sete principais casos de uso de dados em vendas:

A. Previsão de vendas: auxilia questões como o gerenciamento de estoque, logística, produção e planejamento de mão de obra;
B. Aumento da geração de *leads*: com uma combinação de dados internos dos clientes e dados externos é possível orientar as estratégias para identificar o cliente certo no momento certo. Focar em *leads* com maior probabilidade de fechamento é uma forma de priorizar a alocação de recursos e melhorar a taxa de conversão de vendas;
C. Análise do sentimento do cliente: através de *feedbacks* dos clientes, é possível entender de forma automatizada como eles percebem a marca em questão;

D. Aumento do *cross-selling e up-selling*: o uso de dados torna capaz identificar parâmetros de venda importantes para traçar estratégias de venda de itens com maior valor para um cliente (upsell) e vendas de itens complementares (cross-sell);
E. Melhora do valor da vida útil do cliente (do inglês, *customer lifetime value [CLV]*): o *CLV* trata do tempo de relacionamento e lucratividade que o cliente trará para a empresa. Com o uso do *BI* pode-se entender as tendências de comportamento dos clientes, permitindo planejamento de ações futuras;
F. Definição correta do preço de venda: empresas estão implantando metodologias de precificação dinâmica que utilizam dados internos, do mercado em tempo real, e da concorrência de forma a obter o preço ideal de venda, tirando de cena fatores subjetivos como a experiência de vendedores;
G. Prevenção de rotatividade: por meio dos registros dos clientes que pararam de comprar, é possível entender tendências de rotatividade ou perda de clientes, o que é tão importante quanto entender o comportamento de vendas.

Nessa perspectiva, Marr (2016), descreve o uso de dados pela Amazon, uma das maiores varejistas de bens físicos e virtuais do mundo. Com uma grande gama de produtos disponíveis aos clientes, estes possuem dificuldades de identificar qual a melhor opção para atender suas necessidades, o que a empresa simplifica utilizando mecanismos de recomendações gerados a partir de dados de seus mais de um quarto de bilhão de clientes. A diferença dessa metodologia para a adotada por outras organizações é que ela não se baseia apenas nos itens já procurados, e sim no perfil do cliente, que é descoberto e atribuído a um grupo com costumes semelhantes de compra.

Outro caso da utilização de dados em vendas é relatado por Turban et al. (2009) ao descrever o método de precificação de uma rede de farmácias norte-americana. Os preços, que antes eram determinados de forma manual com modificações sugeridas pelo fabricante e decorrentes da sazonalidade, passaram a ser estabelecidos por um programa de otimização de preços. A metodologia utiliza as regras de negócios da rede e algoritmos de cálculo para recomendar automaticamente um preço para cada item de cada loja. Com a utilização do sistema, a empresa demonstrou aumentos de até 10% de receita nas farmácias da rede.

2.2 DATA SCIENCE

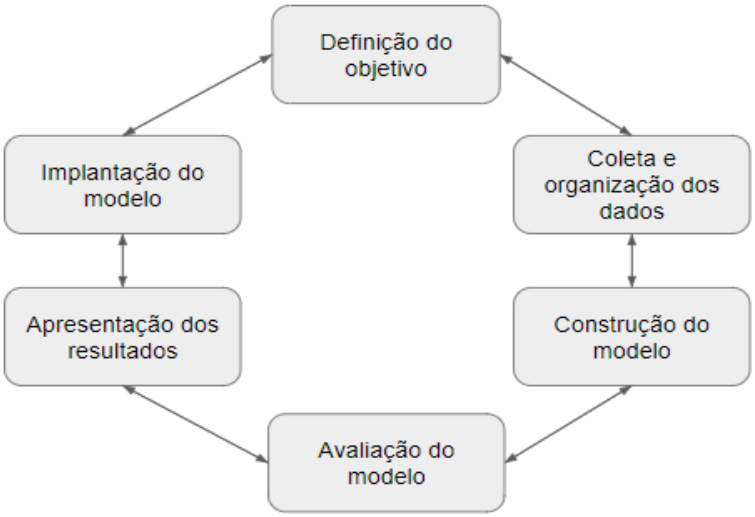
Conforme já exposto, *Business Intelligence* trata do processo de coleta e transformação de dados em conhecimentos que auxiliam na tomada de decisão, e é nesse contexto que se insere o *Data Science*. De acordo com Matos (2015), essa ciência converte dados brutos em ideias de negócios que são usados por gestores para construção de estratégias. Ainda segundo o autor, *Data Science* engloba os seguintes elementos:

A. Análise Quantitativa: modelagem matemática, análise estatística, previsões e simulações;
B. Habilidades de Programação: habilidades em programação para analisar dados brutos e torná-los acessíveis aos usuários de negócio;
C. Conhecimento de Negócios: conhecimento do ambiente de negócio, para melhor compreender a relevância dos resultados encontrados.

Em relação ao termo, uma ciência trata de sintetizar, normalizar e organizar a informação e conhecimento de forma sistemática. Do mesmo modo, a ciência de dados se objetiva a estudar o dado em todo o seu ciclo de vida (AMARAL, 2016). Sob essa perspectiva, o autor também chama atenção para a diferença entre a estatística e o *Data Science*: enquanto a primeira dedica-se à manipulação de dados, o segundo aborda processos, modelos e tecnologias que estudam os dados da produção ao seu descarte, passando pelo seu uso estratégico e aplicado ao negócio.

Em adição, Zumel e Mount (2014) defendem que o sucesso de um projeto *Data Science* depende de metas quantificáveis, fluxo de trabalho repetível, interações interdisciplinares e uma boa metodologia. Sobre esse último ponto, os autores sugerem um ciclo de etapas para realização do projeto, destacando que os limites entre os estágios são fluidos, podendo ir e voltar entre as fases até que se avance no projeto como um todo (Figura 3).

Figura 3 - Ciclo de vida de um projeto *Data Science*



Fonte: Adaptado (ZUMEL; MOUNT, 2014)

2.2.1 Análise Preditiva

Tendo em vista a captura de informações em volumes cada vez maiores e também a dificuldade de extrair padrões de dados à primeira vista, surgiram uma série de técnicas estatísticas capazes de agregar valor aos registros e realizar manipulações confiáveis. De acordo com Guazzelli (2012), ao contrário do BI tradicional que trata da utilização de dados para entendimento de um acontecimento passado, a análise preditiva é o ramo do *Data Science* que possibilita prever as tendências e estimar probabilidades de eventos futuros.

Segundo Turban et al. (2009), as ferramentas de análise preditiva auxiliam na obtenção da probabilidade de um acontecimento futuro se concretizar ou na identificação de relações e padrões. Assim, o objetivo da análise preditiva é a construção de um modelo probabilístico que possibilite predizer, com base em observações passadas, o comportamento aleatório de eventos futuros (TURKMAN, 1995).

Dessa forma, o termo análise preditiva se torna amplo por englobar uma grande variedade de técnicas estatísticas e analíticas utilizadas na construção de modelos que prevêem comportamentos futuros, assim como métodos para avaliação da qualidade dessas previsões (NYCE, 2007). Em concordância com essa definição, Guazzelli (2012) complementa o conceito de modelo preditivo como uma função matemática capaz de identificar o mapeamento entre um conjunto de variáveis de entrada de dados e uma variável de resposta.

Ainda segundo Guazzelli (2012), a quantidade e qualidade dos dados utilizados na análise preditiva é de suma importância. Apenas com uma quantidade significativa de informações que um modelo preditivo é capaz de aprender e generalizar um conjunto de registros para obtenção de previsões. Já em relação a qualidade dos dados, esse é um ponto determinante para a obtenção de modelos preditivos com maior nível de assertividade.

2.2.2 Machine Learning

Tradicionalmente, a única forma de fazer um computador executar uma operação era escrever um algoritmo que explicasse em detalhes como, porém essa condição é quebrada pelo *Machine Learnig*. Pode-se dizer que esses algoritmos, conhecidos como aprendizes, usam de inferências de dados para descobrirem sozinhos o como, chegando ao ponto de se dizer que os computadores são uma tecnologia que se constrói a si própria (DOMINGOS, 2017).

De acordo com Burkov (2019), o ML tem como objetivo imitar a inteligência humana a partir de algoritmos computacionais que se utilizam de informações fornecidas pelo seu entorno. Esses algoritmos são um conjunto de instruções precisas o suficiente para serem executadas por uma máquina (DOMINGOS, 2017).

A tecnologia *Machine Learnig* tem como base o treinamento de modelos em conjuntos de dados, que podem ser textos, números e até imagens, antes de serem implementados. Uma aplicação com essa técnica melhora de forma contínua e automática à medida que mais experiência ela adquire ao ser colocada para treinar (ROLLINS, 2021)**.** Ou seja, a partir de dados, os modelos são treinados para fazerem previsões cada vez melhores.

Segundo Makridakis, Spiliotis e Assimakopoulos (2018), atualmente umas das principais aplicações do ML são os sistemas de previsão, que se baseiam em processos de tentativa e erro até que estejam suficientemente refinados. Também, esse treino deve se dar a partir de dados que sejam o bastante para colaborar para a qualidade do modelo, porém que não sejam demais a ponto de causar elevados custos para a operação.

Ao se tratar de negócios, de acordo com Domingos (2017) num geral as organizações iniciam suas operações realizando as funções de forma manual, a medida que crescem passam a utilizar da programação tradicional para atender suas demandas, até que a fase de *Machine Learnig* é alcançada, quando não é mais possível descrever todas as regras do negócio. Ainda segundo o autor, o ML também permite a análise de problemas mais complexos, uma vez que

com o uso de muitos dados o desafio passa a ser reunir as informações de forma coerente para a gestão.

Em consonância, Rollins (2021) descreve que o *Machine Learning* oferece valor potencial às organizações que possuem um grande volume de dados e querem entender melhor pequenas mudanças de comportamento, preferência ou satisfação de *leads* e clientes.

2.2.2.1 Tipos de Aprendizagem de Máquina

Ao se tratar de *Machine Learning*, para que se garanta o melhor modelo de previsão possível, diferentes técnicas são adotadas a depender da natureza da problemática e do volume e arranjo dos dados disponíveis. De acordo com Burkov (2019), existem três tipos de aprendizagens, que são apresentadas a seguir:

A. Aprendizado Supervisionado: tem como objetivo prever uma variável dependente a partir de uma lista de variáveis independentes. Nesse tipo de sistema a resposta que é buscada pelo modelo está contida entre os dados que são usados no treinamento. São exemplos dessa abordagem as técnicas regressão linear e logística e árvore de decisão;
B. Aprendizado Não-supervisionado: empregado quando o problema envolve uma grande quantidade de informações não rotuladas, então é aplicado um processo iterativo para que se encontre uma representação informativa desses dados. As técnicas de clusterização e mapas auto-organizados se encaixam nessa metodologia;
C. Aprendizado por Reforço: nessa conduta a máquina observa os cenários futuros possíveis e escolhe uma ação a ser tomada. Após, ela recebe uma recompensa de acordo com a escolha do estado, obtendo assim a informação desejada, e esse processo se repete até que a máquina seja capaz de escolher a melhor ação dentre os possíveis cenários futuros. Essa técnica é adotada quando a decisão é sequencial e o objetivo é longo prazo.

Em relação a esses tipos de aprendizagem de ML, Rollins (2021) destaca que o Aprendizado Supervisionado é onde a grande parte dos problemas já estão bem definidos, tornando-o uma das técnicas mais utilizadas atualmente. Por outro lado, o Aprendizado Não-

Supervisionado é empregado em contextos em que conseguir dados relacionados é impossível ou muito custoso.

2.3 TÉCNICAS

De forma geral, segundo James et al. (2021), problemas que envolvem uma resposta quantitativa são considerados como de regressão, e aquelas que abrangem soluções qualitativas como de classificação. Porém, há alguns métodos, como é o caso da regressão logística, que ao serem aplicados resultam em uma variável quantitativa que é interpretada como qualitativa, dando a característica de classificação ao método. Neste sentido, a seguir são apresentadas as técnicas de classificação clusterização e regressão logística e a técnica *random forest* como método de regressão.

2.3.1 Clusterização

Segundo Zumel e Mount (2014), a análise de cluster tem como objetivo formar grupos de observações em que os dados de determinado cluster sejam mais semelhantes a outros dados do mesmo cluster do que as observações de outros aglomerados. Ou seja, busca-se formar classes agrupando os registros mais semelhantes possíveis, de forma que se tenha uma identificação clara do perfil de determinado cluster, gerando um maior entendimento do conjunto de dados.

Neste sentido, a clusterização pode ser utilizada quando se é necessário reduzir o número de objetos para um número de subgrupos característicos, tornando mais eficiente e eficaz a descrição dos aspectos peculiares de cada grupo identificado (CASSIANO, 2014). Ainda segundo a autora, outra utilidade do método é na identificação de relacionamentos entre as observações, auxiliando na exploração dos dados.

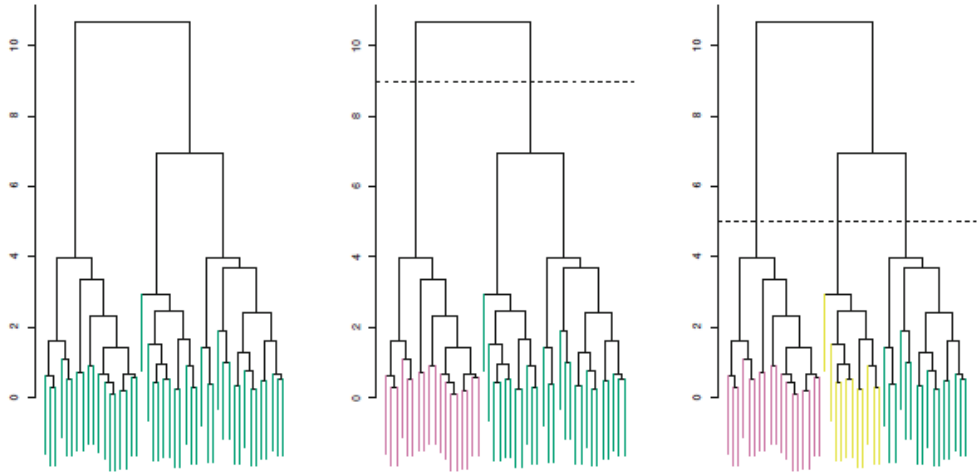
Quanto à técnica de clusterização, a partir de uma base de dados relevante, tem-se que a avaliação exaustiva de todas as possibilidades de configurações de cluster é computacionalmente inviável. Dessa forma, métodos aproximados têm sido propostos como forma de fornecer soluções com significativo acerto e redução de complexidade (OCHI; DIAS; SOARES, 2004). Assim, surgem as técnicas de particionamento e hierárquicas.

No método de clusterização particionado, segundo Ochi, Dias e Soares (2004), uma forma de identificar a similaridade é medir a distância entre os valores de um atributo de cada

elemento. Isso faz com que o conjunto de dados seja segmentado em k subconjuntos por meio da avaliação de uma função objetivo. Dessa forma, os elementos migram de grupos de forma iterativa até que essa avaliação demonstre que os clusters atendem ao problema em questão, podendo ser melhorados de forma gradativa.

Já no que tange a clusterização hierárquica, pode-se não ter previamente o número de clusters desejados, uma vez que os grupos vão sendo construídos de forma gradativa por meio de aglomerações ou divisões de elementos (JAMES et al., 2021). Ainda segundo o autor, essa hierarquia de clusters é tradicionalmente representada em forma de árvore das observações, chamada de dendrograma (Figura 4).

Figura 4 - Dendrogramas com diferentes definições de cluster



Fonte: (JAMES et al., 2021)

Na Figura 4, de acordo com James et al. (2021), o dendrograma da esquerda representa o agrupamento hierárquico inicial de um conjunto de dados. Já a imagem central, em que o tracejado indica que o dendrograma foi cortado em uma altura de nove, representa dois clusters distintos. Por fim, o dendrograma ao lado direito simboliza a formação de três grupos diferentes quando o corte é realizado na altura cinco. Ou seja, após a consolidação do dendrograma o número de clusters será definido a partir da escolha da altura do corte.

Em relação a usabilidade do método hierárquico, Ochi, Dias e Soares (2004) ressaltam a facilidade de lidar com as medidas de similaridade, possibilitando a aplicação em problemas com dados numéricos e categóricos. Já a desvantagem relacionada a técnica diz respeito a

imprecisão do critério de parada, e também ao fato de que a maioria dos algoritmos não revisitam os grupos já formados no decorrer das execuções.

2.3.2 Regressão Logística

A construção de modelos de previsão trata da relação entre variáveis dependentes em função de variáveis independentes. A partir disso, a modelagem por regressão logística é a técnica usada para prever probabilidades ou taxas e seus coeficientes são indicativos do comportamento esperado do evento (ZUMEL; MOUNT, 2014). Ou seja, além de se obter a probabilidade de acontecimento de determinado fato, é possível compreender a influência que cada variável tem sobre o caso em análise.

Nesse sentido, o atributo fim, aquele que se busca prever, é denotado de variável dependente, enquanto aqueles atributos que possuem influência, sendo usados para fazer a previsão, são intitulados de variáveis independentes. De acordo com Battisti e Smolski (2019), a regressão logística é utilizada em três casos: quando a variável dependente é binária, detectando a presença ou não de alguma característica, quando ela é categórica ordenada, ou seja, há uma hierarquia dentre as variáveis de resposta, ou quando essa variável é categórica não ordenada, não possuindo um relação de ordem entre elas.

Ainda sobre isto, segundo Zumel e Mount (2014) o modelo de regressão logística visa encontrar a combinação mais vantajosa quanto aos dados de entrada, buscando assim os melhores coeficientes para prever a variável dependente. Também de acordo com os autores, na utilização dessa técnica para construção de modelos computacionais, a previsão de valores fica restrita ao intervalo entre zero e um, assumindo um caráter probabilístico.

Como embasamento do método, a curva logística é utilizada para representação da relação entre a variável dependente e as independentes nesse tipo de regressão, a partir dos coeficientes estimados (BATTISTI; SMOLSKI, 2019). Isso explica o porquê dos valores previstos estarem limitados entre zero e um, conforme demonstra a relação da curva logística na Figura 5.

Figura 5 - Curva logística

Fonte: (BATTISTI; SMOLSKI, 2019)

Por se tratar do interesse em mensurar a probabilidade de um evento ocorrer, segundo Zumel e Mount (2014), a regressão logística se torna de suma importância em muitas áreas. Exemplos disso são a aplicação da técnica em marketing e comercial, buscando prever quais leads possuem maior chance de venda, no setor financeiro, no intuito de destacar possíveis transações fraudulentas e até mesmo em setores relacionados à saúde na identificação de doenças.

2.3.2.1 Avaliação do modelo de Regressão Logística

Em relação a análise da regressão logística, o critério de otimização utilizado é denominado máxima verossimilhança, em que a probabilidade dos dados do modelo são maximizados e se busca estimar os coeficientes mais prováveis (BURKOV, 2019). Ainda segundo o autor, a qualidade do ajuste do modelo pode ser avaliado com o uso da matriz de confusão e da curva ROC.

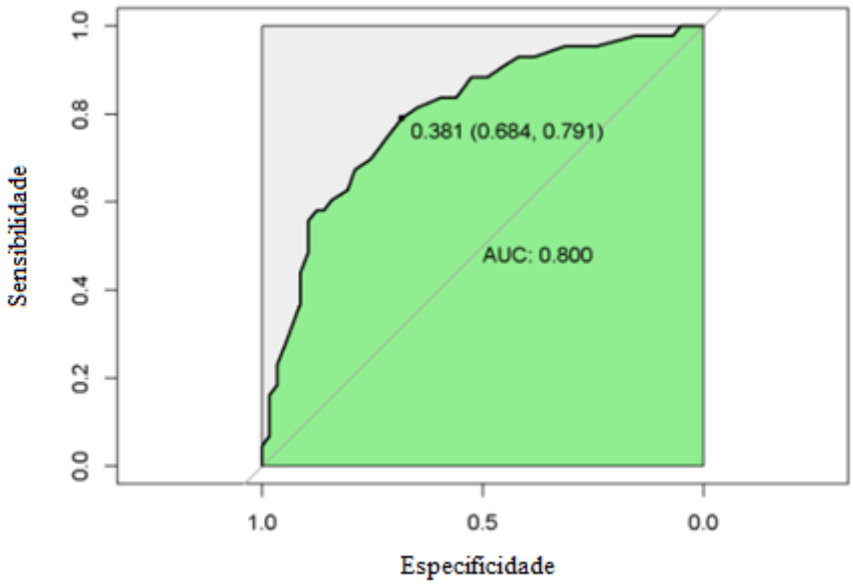
De acordo com James et al. (2021), a utilização da matriz de confusão trata da projeção do modelo, aplicado para predição de uma variável dicotômica, em uma tabela de classificação. Para tal, a matriz construída apresenta o resultado da aplicação da regressão logística para a variável resposta em relação às observações reais, medindo o desempenho do modelo na classificação de eventos e não eventos. Ou seja, constrói-se uma tabela com o número de observações de Verdadeiro Positivo (VP), Falso Positivo (FP), Falso Negativo (FN) e

Verdadeiro Negativo (VN). Ainda segundo os autores, através dessa matriz são traçados os seguintes parâmetros:

A. Precisão: é uma proporção das predições corretas sobre o total [VP/(VP+FP)];
B. Sensibilidade: representa a capacidade do modelo em identificar o evento por meio da proporção de verdadeiros positivos [VP/(VP+FN)];
C. Especificidade: representa a capacidade do modelo em identificar o não evento por meio da proporção de verdadeiros negativos [VN/(FP+VN)].

Já no que tange a utilização da curva ROC para avaliação da regressão logística, esta consiste em mensurar a capacidade de predição do modelo por meio da sensibilidade e especificidade (BATTISTI; SMOLSKI, 2019). Esta curva é produzida com a relação entre a taxa de Verdadeiros Positivos e Falsos Positivos obtidos na previsão do modelo. Ainda de acordo com os autores, o indicador área sob a curva ROC (do inglês, *Area Under The Curve [AUC])* demonstra a probabilidade de que o modelo efetue predições positivas em relação a negativas, estando sempre entre zero e um e quanto mais próximo de um, melhor. Sobre isto, o melhor posicionamento dentro do gráfico é na porção superior esquerda, como mostra o ponto destacado na Figura 6.

Figura 6 - Curva ROC



Fonte: (BATTISTI; SMOLSKI, 2019)

2.3.3 Random Forest

De acordo com James et al. (2021), *random forest* é uma derivação do método da árvore de decisão que propõe alguns ajustes de melhoria. Árvore de decisão trata de uma técnica de classificação de dados com base em valores de variáveis de entrada em que se constrói uma hierarquia de declarações se-então (TURBAN et al., 2009)**.** Segundo os autores, uma árvore consiste em ramificações e nós, em que uma ramificação significa o resultado de um teste de classificação, enquanto os nós representam o teste em um atributo (nós intermediários) ou a opção de classe final para um padrão (nó folha no fim).

Neste sentido, cada árvore é construída a partir da mesma relação de dados, possibilitando a utilização de conjuntos semelhantes de recursos em árvores individuais, alterando-se a ordem ou os valores de divisão, provocando uma correlação entre as árvores (ZUMEL; MOUNT, 2014). A partir disso, a abordagem *random forest* visa descorrelacionar as árvores, randomizando o conjunto de recursos disponíveis para utilização. Segundo James et al. (2021), em cada divisão da árvore, durante a construção do *random forest,* o algoritmo é forçado a considerar apenas um subconjunto de preditores, fazendo com que a média das árvores resultantes seja menos variável.

No método, a combinação de predição de um conjunto de árvores de decisão origina apenas uma única resposta de saída, que tende a apresentar melhor desempenho do que as obtidas com cada modelo separado. O *random forest*, em relação a árvore de decisão, apresenta como principais vantagens a diminuição do erro de previsão, redução da ocorrência de sobreajuste (quando o modelo se ajusta aos dados históricos mas é incapaz de prever novos resultado) e a capacidade de previsão de base de dados mais extensas, porém para isso se torna um modelo mais complexo, exigindo mais recursos computacionais, e apresentando maior dificuldade de implementação e interpretação (MAURYA et al., 2020)**.**

Segundo Ponte, Caminha e Furtado (2020), este tipo de modelagem é usualmente empregada para classificação, mas também pode ser aplicada para regressão, estudo de importância e seleção de variáveis. Ao se tratar de classificação, a resposta final é obtida pela maioria, já para regressão a predição final é a média entre as predições individuais. Assim, ainda segundo os autores, a qualidade da predição de uma árvore depende da predição das suas folhas, que está relacionada com a distribuição dos valores da variável alvo.

2.3.3.1 Avaliação do modelo de Random Forest

Quanto a análise da qualidade da modelagem pelo random forest, de acordo com James et al. (2021), uma das métricas de erro mais utilizadas é o Erro Quadrático Médio (do inglês, *Root Mean Squared Error [RMSE])*, que calcula a raiz quadrática média dos erros entre os valores reais e os preditos. Este método pode ser interpretado como um desvio padrão, demonstrando o quanto a previsão do modelo está desajustada, e dependendo do método utilizado, muitas vezes é o que o algoritmo de ajuste está tentando minimizar (ZUMEL; MOUNT, 2014). Ainda segundo os autores, o *RMSE* é expresso nas mesmas unidades de medida que os seus valores Y, e apresenta como vantagem o fato de que seus valores são relativos aos dados utilizados, fazendo com que o método seja de fácil aplicação.

Outra metodologia também empregada na avaliação do modelo random forest é o R-quadrado (R2), ou coeficiente de determinação. Esta é uma medida estatística que representa o quão próximos os dados estão da linha de regressão ajustada, sendo a porcentagem de variação do atributo de resposta que é explicado pelo modelo (JAMES et al., 2021). Nesse sentido, o R2 está sempre entre 0 e 100%, sendo que quanto mais próximo de 100 indica que o modelo explica a variabilidade dos dados de resposta, e quanto mais próximo de 0, que a qualidade de predição do modelo não é elevada. De acordo com Zumel e Mount (2014)**,** o R2 é adimensional e pode ser interpretado como uma versão normalizada do RMSE. Também, ao se plotar os valores reais e os valores preditos em um gráfico tem-se uma visualização da qualidade do modelo, em que quanto mais próximos os pontos estiverem da linha de regressão ajustada, maior é o valor de R2.

3 METODOLOGIA DA PESQUISA

Esta seção apresenta a caracterização da pesquisa, os materiais e métodos adotados, assim como o procedimento metodológico aplicado.

3.1 ENQUADRAMENTO METODOLÓGICO

De acordo com Silva e Menezes (2005), a pesquisa se classifica quanto à sua natureza, à forma de abordagem do problema, aos seus objetivos e aos procedimentos técnicos. Assim, o presente trabalho se caracteriza como de natureza de pesquisa aplicada, uma vez que trata de problemas específicos com o objetivo de gerar soluções com aplicações práticas. Ao se considerar a forma de abordagem do problema, este se apresenta como uma pesquisa quantitativa pois exige o uso de técnicas estatísticas, a exemplo da clusterização e análise de regressão, dando um caráter quantificável às informações em estudo (MENEZES; SILVA, 2005). Já no que se refere ao objetivo deste trabalho, segundo Gil (1991), por buscar o entendimento de relações entre variáveis e por fazer uso de formas padronizadas de coleta de dados, este se apresenta como uma pesquisa descritiva.

Do ponto de vista dos procedimentos metodológicos, a pesquisa pode ser vista como *Design Science Research*. Assim se define, pois, trata de um problema prático envolvido em uma questão organizacional, e possui como objetivo a construção de um artefato do tipo modelo para constituir uma solução viável para a problemática. Para isso utiliza-se de métodos rigorosos, tanto na construção como na avaliação dos artefatos (LACERDA et al., 2013). Ainda, conforme o Quadro 2, este procedimento metodológico diferencia-se de outros como o Estudo de Caso, que tem como objeto de estudo algum fenômeno social, e da Pesquisa-Ação, que atua na resolução de um problema específico gerando conhecimento sobre.

Quadro 2 – Comparativo *Design Science Research*, Estudo de Caso e Pesquisa-Ação

Características	*Design Science Research*	Estudo de Caso tradicional	Pesquisa-Ação tradicional
Objetivos	Desenvolver artefatos que permitam soluções satisfatórias aos problemas práticos.	Auxiliar na compreensão de fenômenos sociais complexos.	Resolver ou explicar problemas de um determinado sistema gerando conhecimento para a prática e para a teoria.
	Prescrever e Projetar	Explorar, Descrever e Explicar	Explorar, Descrever e Explicar
Principais Atividades	• Conscientizar • Sugerir • Desenvolver • Avaliar • Concluir	• Definir Estrutura Conceitual • Planejar o(s) caso(s) • Conduzir Piloto • Coletar Dados • Analisar Dados • Gerar Relatório Miguel (2007, p. 221)	• Planejar a Ação • Coletar Dados • Analisar dados e Planejar ações • Implementar Ações • Avaliar Resultados • Monitorar (Contínuo) Turrioni e Mello (2010)
Resultados	Artefatos (Constructos, Modelos, Métodos, Instanciações)	Constructos Hipóteses Descrições Explicações	Constructos Hipóteses Descrições Explicações Ações
Tipo de Conhecimento	Como as coisas deveriam ser	Como as coisas são ou como se comportam.	Como as coisas são ou como se comportam.
Papel do Pesquisador	Construtor e Avaliador do Artefato	Observador	Múltiplo, em função do Tipo de Pesquisa-Ação
Base Empírica	Não obrigatória	Obrigatória	Obrigatória
Colaboração Pesquisador-Pesquisado	Não obrigatória	Não obrigatória	Obrigatória
Implementação	Não obrigatória	Não se Aplica	Obrigatória
Avaliação dos Resultados	Aplicações Simulações Experimentos	Confronto com a Teoria	Confronto com a Teoria
Abordagem	Qualitativa e/ou Quantitativa	Qualitativa	Qualitativa

Fonte: (LACERDA et al., 2013)

3.2 MATERIAIS

Para o entendimento do comportamento do processo de vendas da empresa em estudo, partiu-se da utilização de dados das negociações comerciais já realizadas. Nesse sentido, o registro das atividades exercidas na rotina do setor se dá por meio de um *software* de Gestão de Relacionamento com o Cliente (do inglês, *Customer Relationship Management [CRM]*), ferramenta que possui a finalidade de auxiliar na gestão do relacionamento com os stakeholders. Apesar da implementação de um novo *CRM* em janeiro de 2021, o HubSpot, os dados referentes às negociações anteriores a esse ano foram importados na nova ferramenta, permitindo de forma geral o acesso a todos os dados históricos.

No que tange o uso dessa ferramenta, o seu acesso se dá de forma online e está disponível para todos os colaboradores por meio de login individual, mesmo que com diferentes níveis de acesso a depender das funções exercidas. Sob essa perspectiva, durante a rotina de trabalho do setor comercial é comum a execução de atividades por meio da ferramenta, como a realização de chamadas, agendamento de reuniões e envio de propostas comerciais. Dessa forma, com exceção do viés subjetivo da negociação, todo o processo de venda, desde o primeiro contato com o cliente até a assinatura do contrato, possui registro de dados na plataforma.

Para a utilização desse tipo de informação no presente trabalho, foi necessária a extração de um relatório contendo os dados secundários de venda do *CRM*, e posterior análise em outro software que possuísse maior capacidade de refinamento de análise. Assim, procedeu-se com a configuração de um relatório em que as linhas representassem as negociações em análise, e as colunas as principais características do processo de vendas, gerando um arquivo Excel que é exemplificado na Figura 7.

Figura 7 - Relatório de dados do CRM

Pipeline	Etapa	Fase do ciclo de vida	Valor	Estado	Segmento	Solução	Data de criação	Data de fechamento	Origem da Oportunidade	Tags do Negócio	Tags de Venda	Company ID	Deal ID	Line item ID
Listas de Varejo	Ganho	Customer	650.0	SP										
Listas de Varejo	Ganho	Customer	1572.0	PR	Produtor									
Listas de Varejo	Ganho	Other	0.0	SP										
Listas de Varejo	Ganho	Customer	2150.0	PR	Produtor	Rastreador	06/07/2021	07/07/2021	Outbound	Regional SUL NE NO	REALIZADO	5408038263	5621258342	1712236058
Listas de Varejo	Perdido	Customer	2040.0	ES	Distribuidor									
Listas de Varejo	Perdido	Opportunity					26/02/2021	28/04/2021	Outbound	Regional CENTRO		5462363752	4375788274	
Listas de Varejo	Ganho	Customer	540.0	SP		Rastreador	27/08/2021	06/09/2021	Outbound	Regional CENTRO	REALIZADO	6856681000	6045921032	1929463470
Listas de Varejo	Follow-up	Other	3670.0	RS	Distribuidor	Rastreador	06/09/2021	30/09/2021	Inbound	Regional SUL NE NO		5154886297	6151331487	1959859773
Listas de Varejo	Perdido	Opportunity	3170.0	SC	Indústria	Implantação	22/04/2021	05/05/2021	Inbound	Regional SUL NE NO		5921925776	5075227018	1396797698

Fonte: Elaborado pelo autor.

À respeito do espaço amostral, foi utilizado o histórico de dados de dois anos (2020 e 2021) por se considerar que representa o período em que o processo de vendas passou a ter mais características similares às atuais. Por conta disso, foi realizada uma análise da necessidade de limpeza dos dados anteriores à implantação do HubSpot, já que havia a possibilidade de discrepância no formato em relação aos dados registrados diretamente no novo CRM.

3.3 MÉTODOS

Em relação aos métodos que foram adotados para o atingimento dos objetivos traçados, fez-se uso da clusterização, regressão logística e *random forest* através da linguagem de programação R com o uso de pacotes voltados para o estudo estatísticos.

3.3.1 Clusterização

Inicialmente, a clusterização para a segmentação dos clientes foi empregada com o intuito de gerar insumos para uma abordagem mais específica ao perfil de cada *lead*. Partindo de uma base de dados, essa técnica consiste em reunir em clusters elementos que sejam similares entre si, de forma que possuam mais caraterísticas em comum com os elementos do seu grupo do que com outros grupos. Ainda, esse procedimento pode ser aplicado de forma supervisionada ou não, sendo a segunda maneira a adotada no presente trabalho, uma vez que é necessário encontrar a melhor estrutura de segmentação para os elementos (JAMES et al., 2021).

3.3.2 Regressão Logística

Já para estimar a probabilidade de sucesso de uma venda, foi empregada a regressão logística na construção do modelo. Para tal, de acordo com James et al. (2021) utiliza-se uma função logística que possui como saída valores entre 0 e 1, produzindo uma curva em forma de S. Em relação aos coeficientes desta função, devem ser estimados a partir dos dados disponíveis para treinamento do modelo. Assim, com a definição da equação de regressão logística se torna possível o cálculo de taxas e probabilidades.

3.3.3 Random Forest

O método aplicado para desenvolver o objetivo de estimar o tempo de negociação de uma venda foi o random forest. Segundo James et al. (2021), esta técnica se relaciona com a árvore de decisão, uma vez que consiste na randomização de atributos empregados na construção de uma série de ramos. A partir disso, o resultado da análise consiste na combinação da predição do conjunto de árvores originadas.

3.3.4 Linguagem de programação R – Pacotes

Para aplicação dos métodos de clusterização, regressão logística e random forest foi utilizada a linguagem R com o apoio dos pacotes apresentados da Tabela 1

Tabela 1 - Pacotes R utilizados para construção do trabalho

Nome	Descrição	Aplicação	Referência Bibliográfica
tidyverse	Compartilhar representações de dados comuns e design de 'API'	Carregamento dos dados	Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686
readxl	Importe arquivos do Excel para o R	Carregamento dos dados	Hadley Wickham and Jennifer Bryan (2019). readxl: Read Excel Files. R package version 1.3.1. https://CRAN.R-project.org/package=readxl
lubridate	Permitir análise e a manipulação de datas	Cálculo do ciclo de venda	Garrett Grolemund, Hadley Wickham (2011). Dates and Times Made Easy with lubridate. Journal of Statistical Software, 40(3), 1-25. URL https://www.jstatsoft.org/v40/i03/.
caret	Análise de componentes em um conjunto e retorno de classes	Clusterização	Max Kuhn (2021). caret: Classification and Regression Training. R package version 6.0-90. https://CRAN.R-project.org/package=caret
skimr	Alternativa às funções de resumo padrão	Análise Exploratória de Dados	Elin Waring, Michael Quinn, Amelia McNamara, Eduardo Arino de la Rubia, Hao Zhu and Shannon Ellis (2021). skimr: Compact and Flexible Summaries of Data. R package version 2.1.3. https://CRAN.R-project.org/package=skimr
tidymodels	Coleção de pacotes para modelagem e análise estatística.	Organização dos dados em treino e teste	Kuhn et al., (2020). Tidymodels: a collection of packages for modeling and machine learning using tidyverse principles. https://www.tidymodels.org
themis	Melhorar desempenho do modelo por meio do balanceamento de dados.	Balanceamento de dados.	Emil Hvitfeldt (2021). themis: Extra Recipes Steps for Dealing with Unbalanced Data. R package version 0.1.4. https://CRAN.R-project.org/package=themis

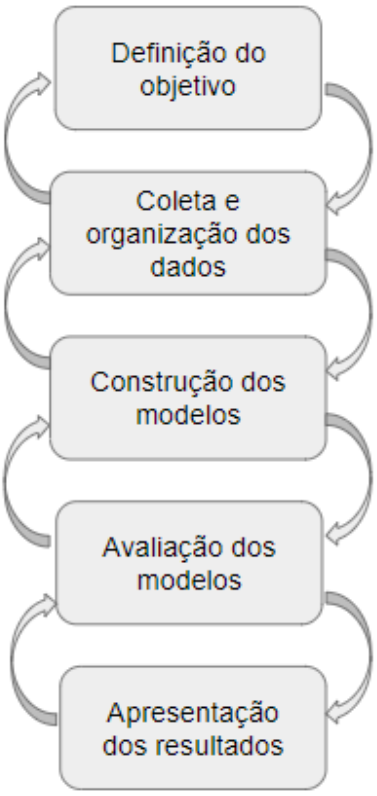
knitr	Ferramenta com design mais flexível e diferentes recursos mais preciso de gráficos.	Avaliação dos modelos	Yihui Xie (2021). knitr: A General-Purpose Package for Dynamic Report Generation in R. R package version 1.36.
MASS	Funções e conjuntos de dados para suportar Venables e Ripley	Modelagem *random forest*	Venables, W. N. & Ripley, B. D. (2002) Modern Applied Statistics with S. Fourth Edition. Springer, New York. ISBN 0-387-95457-0
randomForest	Implementa o algoritmo de floresta aleatória de Breiman	Modelagem *random forest*	A. Liaw and M. Wiener (2002). Classification and Regression by randomForest. R News 2(3), 18--22.

Fonte: Elaborado pelo autor.

3.4 PROCEDIMENTOS METODOLÓGICOS

A presente seção tem como objetivo descrever as etapas adotadas durante a realização deste trabalho. Para construção de tal, adotou-se como referência as fases de um projeto de Data Science apresentadas por Zumel e Mount (2014) (Figura 8). Segundo os autores, o limite entre as etapas pode variar e algumas atividades se sobreporem, entretanto a visão geral dos estágios permanece.

Figura 8 - Etapas do projeto



Fonte: Elaborado pelo autor.

Assim, a Etapa 1: Definição do objetivo consiste em definir o objetivo do projeto, criando uma delimitação clara do trabalho a ser exercido. Para tanto é preciso conhecer o máximo possível do contexto em que a problemática está inserida, como sua importância, o que é feito atualmente e os recursos necessários para a sua resolução (ZUMEL; MOUNT, 2014).

Em seguida, parte-se para a Etapa 2: Coleta e organização dos dados, que segundo Zumel e Mount (2014), é uma das etapas mais importantes pois reunirá as informações que servirão de insumo para resolução do problema. Nesse momento será configurado o relatório para extração do CRM dos registros históricos de negociações comerciais, e posteriormente será analisada a necessidade de normalização dos dados, principalmente em relação a anos distintos.

A partir do objetivo definido e dos registros normalizados, inicia-se a Etapa 3: Construção dos modelos. De acordo com Zumel e Mount (2014), nesse momento os dados serão explorados de forma a trazer percepções úteis sobre a problemática, levando em conta os objetivos específicos traçados no presente trabalho. Assim, antes de se iniciar as análises acerca das variáveis que envolvem o processo de vendas, a segmentação dos clientes, momento em

que os elementos semelhantes serão reunidos por meio da clusterização (OCHI; DIAS; SOARES, 2004), será adotada como forma de melhor compreensão do perfil dos leads. Assim, espera-se contribuir para a tomada de decisão em relação à abordagem dos clientes em potencial.

Ao se considerar o estudo das variáveis que influenciam no processo de negociação comercial, tem-se que a probabilidade de sucesso de uma venda e o tempo estimado de negociação podem ser traçados a partir do uso da regressão logística e random forest. Segundo Zumel e Mount (2014), estas técnicas podem ser empregadas quando o objetivo não é apenas prever um resultado, mas também compreender a relação entre as variáveis de entrada.

Após construir os modelos, na Etapa 4: Avaliação dos modelos, faz se necessária uma avaliação dos mesmos com o intuito de verificar se as questões levantadas na primeira etapa de definição do objetivo foram atendidas (ZUMEL; MOUNT, 2014). Depois que os resultados encontrados fizerem sentido no contexto do problema, parte-se para a Etapa 5: Apresentação dos Resultados em que os modelos são apresentados com recomendações de aplicação para resolução dos problemas reais.

4 ANÁLISE EXPLORATÓRIA DE DADOS

O presente capítulo tem como objetivo descrever a execução das duas primeiras etapas do projeto de *data science*, que foram apresentadas no capítulo 3, na empresa objeto de estudo. Iniciando com a revisão e consolidação do objetivo do atual trabalho, em seguida são descritos os passos de coleta e organização dos registros, culminando na manipulação de dados.

4.1 DEFINIÇÃO DO OBJETIVO

A empresa objeto de estudo neste trabalho, que possui atuação tecnológica no setor agroalimentar, possui uma gama de soluções voltadas para os diferentes elos dessa cadeia. Ao se tratar dos produtos *softwares* disponíveis para a venda, as opções constituem de forma geral duas famílias de produtos: uma voltada à rastreabilidade para a base da cadeia (produtos e distribuidores de alimentos) e outra com atuação mais transversal atendendo requisitos de gestão da qualidade e processo.

Pelo fato de as duas soluções apresentarem características de negociação muito distintas, como equipe de vendas, abordagem, público alvo e ticket médio, percebeu-se a inviabilidade de conduzir as análises das duas ferramentas em conjunto. Assim, optou-se pelo estudo do produto voltado à rastreabilidade, uma vez que este apresenta mais tempo de mercado, possuindo maior quantidade de dados de negociações, e também representa parte significativa da receita da empresa.

Em relação ao setor comercial da empresa em estudo, o processo de vendas passou por algumas modificações de estrutura, abordagem e equipe nos últimos anos. Dessa forma, definiu-se como tempo amostral para o estudo, o período a partir de 2020, por apresentar características semelhantes a metodologia de vendas abordada atualmente.

Dessa maneira, o presente trabalho tem como objetivo aplicar a análise preditiva nos registros relacionados ao processo de venda da empresa. Com isso espera-se uma maior compreensão, baseada em dados, da dinâmica comercial, que possibilite a geração de ideias e melhorias de eficiência no setor.

Assim sendo, com o intuito de compreender melhor os usuários da ferramenta e para que se possa traçar estratégias de abordagem mais assertivas, busca-se primeiramente a distinção dos clientes em potencial em grupos de acordo com suas características. Em outros

termos, tem-se o propósito de realizar a clusterização dos leads que estiveram envolvidos em algum processo de negociação da solução.

Por fim, para que se possa construir estratégias relacionadas ao processo de negociação, espera-se compreender melhor os fatores que impactam nessa dinâmica. Nesse sentido, pensando na organização financeira da empresa, é importante que se tenha uma previsão comercial de sucesso ou insucesso da negociação e quando essa venda será finalizada. Assim, os aspectos escolhidos para análise por meio de modelos preditivos são o resultado da negociação (ganho ou perdido na venda) e o tempo do ciclo de venda.

4.2 COLETA E ORGANIZAÇÃO DOS DADOS

De acordo com Zumel e Mount (2014), a etapa de coleta e organização dos dados consiste em identificar as informações que serão necessárias extraindo-as dos locais em que estão armazenadas. Também inclui a limpeza dos dados, reparando registros e variáveis de transformação conforme necessário.

Em relação ao setor comercial da empresa em estudo, grande parte das atividades desenvolvidas no dia-a-dia são executadas no CRM, gerando uma grande quantidade de dados sobre as negociações. Para extração desses dados, prosseguiu-se com a configuração na plataforma de um relatório que retornasse as informações necessárias para o trabalho, conforme Figura 9. Dessa forma, no relatório extraído cada linha representa uma negociação finalizada, totalizando 5.484 negócios no período entre 01/01/2020 e 08/10/2021.

Figura 9 - Informações selecionadas no CRM para exportação

Fonte: Elaborado pelo autor.

A partir do relatório removido do CRM, deu-se início ao tratamento dos dados utilizando o Excel. A primeira limpeza realizada foi a exclusão de linhas que continham a palavra “teste” no nome do negócio. Em seguida, foi feita a correção de algumas variáveis que possuíam forma de escrita diferente na plataforma CRM utilizada em 2020, deixando todos os dados padronizados. Depois, foram excluídas as informações excedentes na variável “Tags do Negócio”, deixando apenas a indicação à qual regional de vendas o negócio pertence. Por fim, foram excluídas as linhas em que o “Valor” era igual a zero, por representarem negociações que não chegaram na etapa de precificação antes de serem finalizadas, e demais linhas em que a variável “Segmento” não possuía informação.

Em seguida foram criadas novas colunas no arquivo para representar de forma consolidada informações contidas nas outras variáveis. A primeira foi “Ciclo de Venda”, sendo o resultado da subtração entre a “Data de fechamento” e a “Data de criação do negócio”. A partir do “Nome do negócio” extraiu-se a informação da solução negociada e se possuía característica de Upsell, originando duas novas colunas e permitindo a filtragem das linhas para manter apenas os registros de negociações do produto alvo do presente estudo. E por fim, criou-se uma coluna “Categoria varejo atendido” em que os varejos foram divididos em categorias de

faturamento de acordo com informações da Associação Brasileira de Supermercados (ABRAS). O conjunto de variáveis resultantes e sua explicação é apresentado no Quadro 3.

Quadro 3 - Variáveis da base de dados

Variável	**Significado**	**Categorias**
Pipeline	Local onde os negócios ficam armazenados dentro do CRM	Comercial Padrão Listas de Varejo
Etapa do negócio	Informa se a negociação foi finalizada com uma venda ou não	1 (Ganho) 0 (Perdido)
Nome do negócio	Título que o vendedor usa para identificar a negociação	-
Valor	Representa o valor total que está sendo negociado	Mínimo: R$ 6,25 Máximo: R$ 26031,00
Segmento	Categoria de mercado a qual o lead faz parte	Produtor + Distribuidor Supermercado/Atacado Indústria Consultoria/Auditoria Restaurante/Food Service Outros
Varejos Atendidos	Supermercados que são clientes do lead	-
Data de criação	Data em que a negociação foi iniciada	-
Data de fechamento	Data em que a negociação foi finalizada	-
Origem da Oportunidade	Representa a fonte a qual captou a oportunidade de negociação	Inbound Outbound
Tags do Negócio	Identifica qual equipe de vendas é responsável pela negociação	Regional SUL NE NO Regional CENTRO INTERNACIONAL
Company ID	Identificador do objeto empresa no CRM	-
Deal ID	Identificador do objeto negócio no CRM	-
Ciclo de Venda	Tempo de duração da negociação	Mínimo: 0 Máximo: 1108
Upsell	Identifica se a negociação é em relação ao upgrade da solução de um cliente	sim não
Categoria varejo atendido	Identifica qual o perfil de supermercado que é cliente do lead	A (alto faturamento) B C (baixo faturamento) sem informação

Fonte: Elaborado pelo autor.

4.3 ANÁLISE EXPLORATÓRIA DOS DADOS

Segundo Amaral (2016), a análise exploratória dos dados, que é um procedimento comum na área do *Machine Learning*, tem o objetivo de criar familiaridade com os dados permitindo que as primeiras hipóteses sejam criadas. Com esse objetivo, os registros foram carregados para utilização de um software em linguagem R, onde também será realizada a construção dos modelos. Assim, nesta seção serão abordadas as características da base de dados em estudo.

Após a limpeza inicial, a base de dados resultou em uma seleção de 2443 linhas, e nove variáveis de interesse selecionadas para o estudo, sendo elas:

- Variáveis categóricas: Pipeline, Etapa do negócio, Segmento, Origem da Oportunidade, Tags do Negócio, Upsell e Categoria varejo atendido.
- Variáveis numéricas: Valor e Ciclo de Venda.

No que se refere aos atributos categóricos, o Quadro 4 apresenta a categoria de referência de cada variável desse tipo.

Quadro 4 - Categoria de referência das variáveis categóricas

Variável	**Categoria de Referência**
Pipeline	Comercial Padrão
Segmento	Consultoria/Auditoria
Origem da Oportunidade	Inbound
Tags do Negócio	INTERNACIONAL
Upsell	não
Categoria varejo atendido	A
Etapa do negócio	0

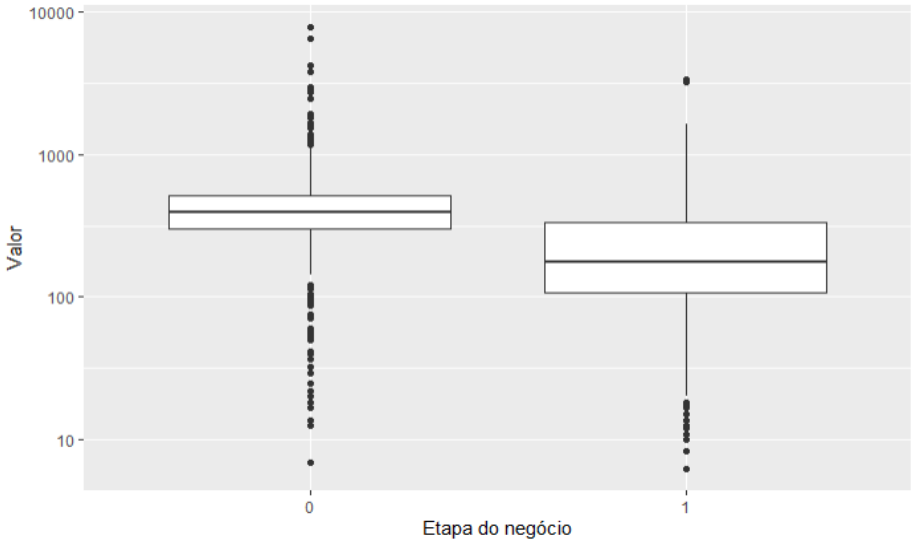
Fonte: Elaborado pelo autor.

Com o objetivo de compreender a distribuição dos valores que as variáveis assumem nos registros, seguiu-se com a manipulação dos dados para que se tenha de forma resumida ou

visual essas informações. Já pensando no modelo de regressão logística que será construído em relação a variável “Etapa do Negócio”, ela será usada como um dos eixos dos gráficos para que tenha maior entendimento dos dados em relação a ela. Para tal, a função *table* foi utilizada para resumo dos dados e a função *ggplot* para a composição dos gráficos.

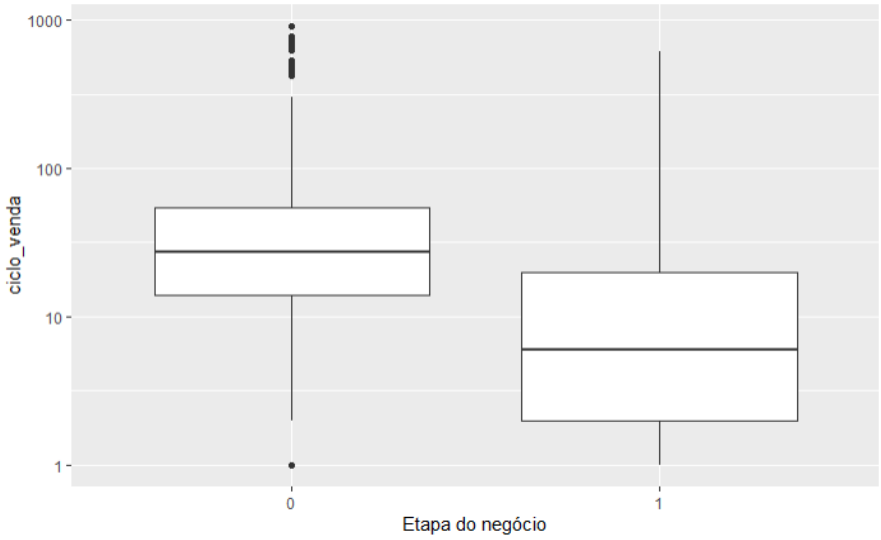
Para as variáveis numéricas, Valor e Ciclo de Venda, o gráfico utilizado para evidenciar a concentração dos dados foi o diagrama de caixa. Conforme a Figura 10, percebe-se que os valores dos negócios perdidos estão de forma geral definidos como maiores e apresentam maior número de pontos fora da curva. Já em relação ao Ciclo de Venda, ganhos e perdidos não possuem grande discrepância, com uma concentração maior de negócios ganhos com um ciclo de venda menor (Figura 11).

Figura 10 - Análise do Valor em relação à Etapa do negócio



Fonte: Elaborado pelo autor.

Figura 11 - Análise do Ciclo de Venda em relação à Etapa do negócio

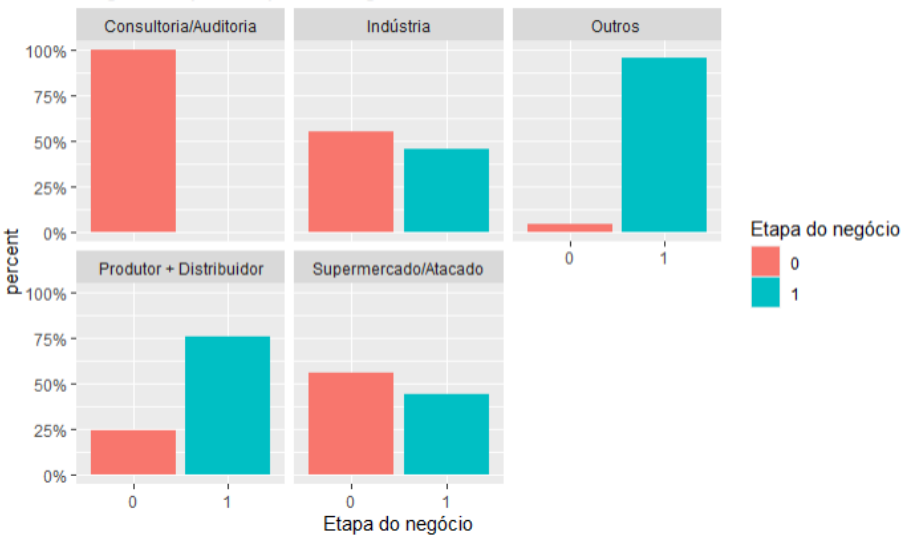


Fonte: Elaborado pelo autor.

No que diz respeito a variável “Tags do negócio”, 53% dos dados se referem a categoria “Regional SUL NE NO”, enquanto os outros 47% se apresentam como “Regional CENTRO”. Ainda sobre essa variável, as duas categorias apresentam um valor aproximado de 75% correspondente a negócios ganhos.

Em relação ao segmento, 95% dos registros são de leads da classe “Produtor + Distribuidor” em que a maior parte são vendas finalizadas com sucesso, como mostra a Figura 12. Em seguida, o segmento “Outros” é o único que também apresenta maior registro de ganhos, porém esses negócios representam apenas 2% da base. Por fim, “Consultoria/Auditoria”, “Indústria” e “Supermercado/Atacado” são constituídos em maior parte por negociações perdidas.

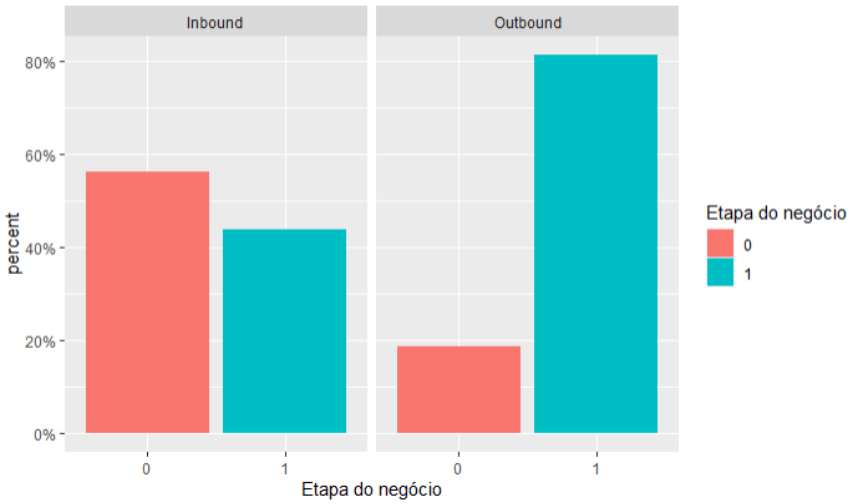
Figura 12 - Análise do Segmento em relação à Etapa do negócio



Fonte: Elaborado pelo autor.

Quanto à origem, os 83% dos negócios provenientes da categoria "Outbound" apresentaram de forma geral resultado positivo quando a negociação, ao contrário da origem "Inbound" que é composta em sua maioria por negócios perdidos (Figura 13). Em relação ao Pipeline em que se encontra o negócio, 60% dos dados estão na classe "Listas de Varejo", o que corrobora com a análise da variável anterior, uma vez que esse funil engloba em sua maioria leads que possuem ligação com uma demanda *compliance* de varejo, possuindo origem outbound e historicamente detendo grande taxa de sucesso em vendas.

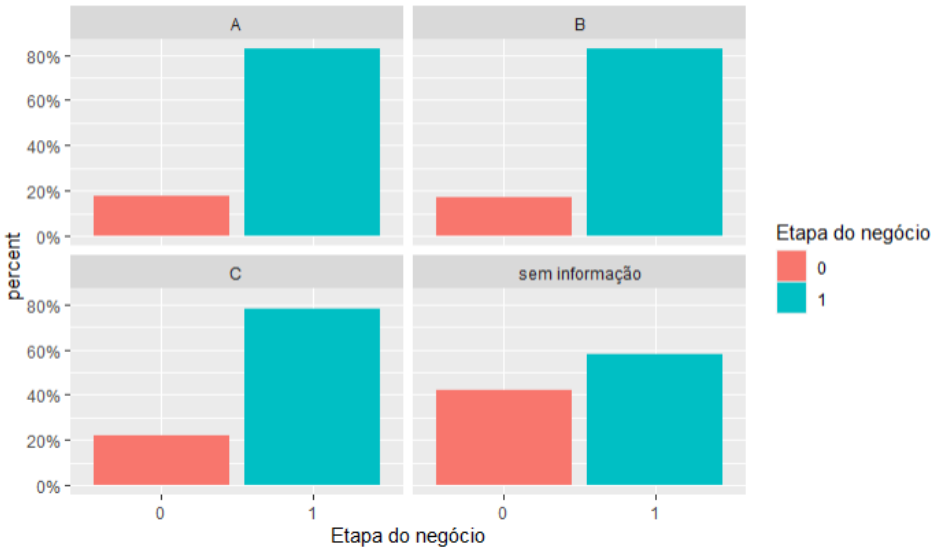
Figura 13 - Análise da Origem da oportunidade em relação à Etapa do negócio



Fonte: Elaborado pelo autor.

Em relação a variável “Upsell” os 89% de negócios que estão presentes na categoria “Não” apresentam uma porção menor de negócios ganhos, evidenciando a tendência de sucesso nos negócios que se caracterizam como um upsell. Já quanto ao perfil de varejo que o lead atende, em todas as categorias o percentual de negócios ganhos é maior, como é apresentado na Figura 14. Também, dentre as quatro categorias, a classe A, aquela com os varejos de maior porte, é responsável por 50% dos registros, destacando o quanto a relação com varejos dessa dimensão pode influenciar na procura pela solução de rastreabilidade.

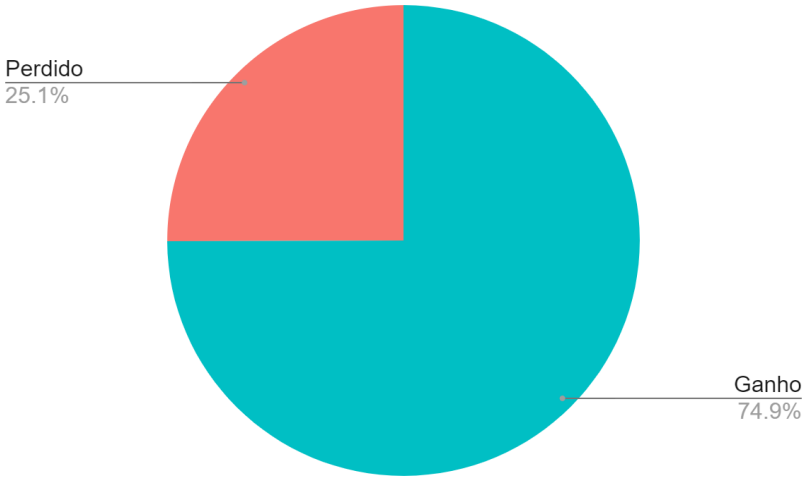
Figura 14 - Análise da Categoria de Varejo em relação à Etapa do negócio



Fonte: Elaborado pelo autor.

Por fim, a Figura 15 expõe a proporção dos dados em relação à Etapa do negócio, evidenciando a grande quantidade de negociações finalizadas com sucesso no conjunto de dados. Essa situação é esperada tendo em vista que os registros de negócio no CRM se iniciam a partir do momento em que os leads passam pela qualificação de pré-venda, eliminando contatos sem potencial aparente de venda.

Figura 15 - Análise da Etapa do negócio



Fonte: Elaborado pelo autor.

Nesse sentido, a grande quantidade de registros referentes a uma mesma categoria de uma variável evidencia um caráter de desbalanceamento dos dados, relevante para aplicação de alguns métodos de *Machine Learning*. Para a construção de um modelo de regressão logística, por exemplo, essa característica prejudica a precisão da modelagem, sendo necessária a aplicação de alguma técnica para balanceamento dos dados.

5 CONSTRUÇÃO E AVALIAÇÃO DOS MODELOS

A atual seção visa apresentar a execução das etapas três, quatro e cinco da metodologia exposta no capítulo 3. A partir dos dados do setor comercial da empresa em estudo, os modelos preditivos são construídos, avaliados e seus resultados expostos.

5.1 SEGMENTAÇÃO DOS CLIENTES

A utilização da técnica de clusterização se deu com o intuito de agrupar leads que possuem perfis semelhantes, permitindo a criação de insumos que embasem estratégias especificas de abordagem.

5.1.1 Construção da Clusterização

Para construção da segmentação de leads, utilizou-se a base de dados de negociações previamente tratada e normalizada, sendo selecionadas para o estudo as seguintes variáveis: Pipeline, Etapa do negócio, Valor, Segmento, Ciclo de Venda, Origem da Oportunidade, Tags do Negócio, Upsell e Categoria varejo atendido.

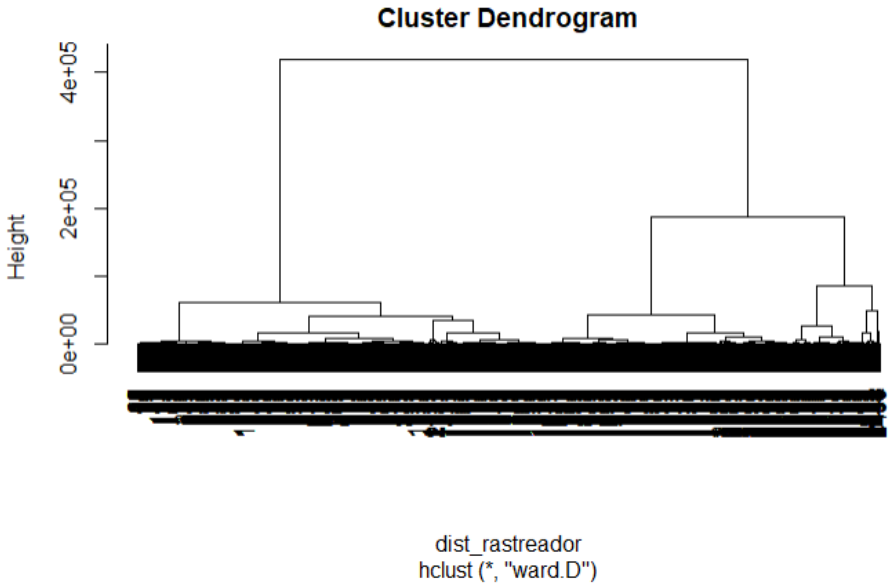
Para construção dos clusters foi adotado o método não supervisionado e de análise de cluster hierárquica por meio da função *hclust* da biblioteca *cluster.* Essa função opera interativamente, iniciando com um cluster para cada observação e unindo dois clusters mais semelhantes a cada fase, até que se tenha apenas um único cluster. Ainda, as distâncias entre os grupos são calculadas de acordo com o método de agrupamento específico escolhido. Assim, os parâmetros de entrada na função são os dados de análise e a técnica para agrupar a ser adotada.

Em relação ao método de agrupamento, este foi escolhido por tentativa e erro dentre as opções disponíveis no pacote (ward.D, ward.D2, single, complete, average, mcquitty, median e centroid). Nesse sentido, a cada teste era observado a composição e a similaridade de tamanho entre os clusters formados, sendo adotado por final o método “ward.D” por apresentar o melhor resultado.

A aplicação da função na base de dados deu origem ao dendrograma apresentado na Figura 16, que representa os grupos formados ao longo das interações à medida que a altura cresce. Após análise gráfica, optou-se pelo corte do dendrograma na altura 100000,

consolidando a formação de três clusters diferentes. O primeiro grupo é composto por 55% dos dados, o segundo por 33% e o último por 12% das observações.

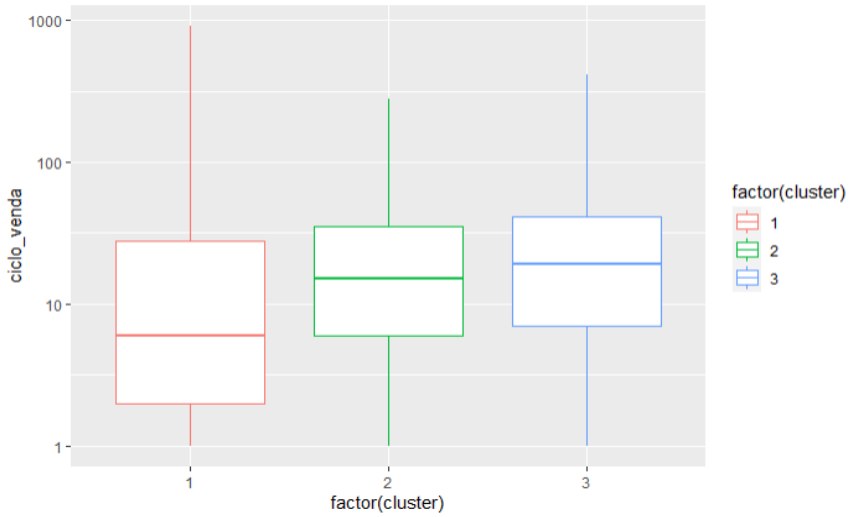
Figura 16 - Dendrograma da clusterização



Fonte: Elaborado pelo autor.

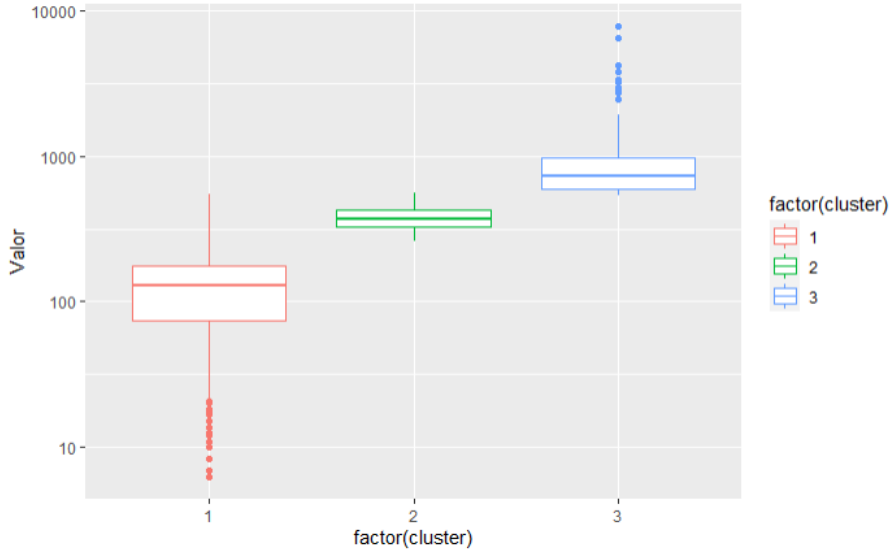
Em seguida, com o objetivo de compreender as características semelhantes que construíram os grupos, foram plotados gráficos para análise de cada variável em relação à constituição dos clusters, conforme apresentado no Apêndice A. Com isso, observou-se que as variáveis Valor e Ciclo de Venda apresentam influência relevante na formação dos clusters, conforme Figura 17 e 18, uma vez que as demais não apresentam valores de variação significativa entre os três grupos.

Figura 17 - Representação do Ciclo de Venda de cada cluster



Fonte: Elaborado pelo autor.

Figura 18 - Representação do Valor de cada cluster



Fonte: Elaborado pelo autor.

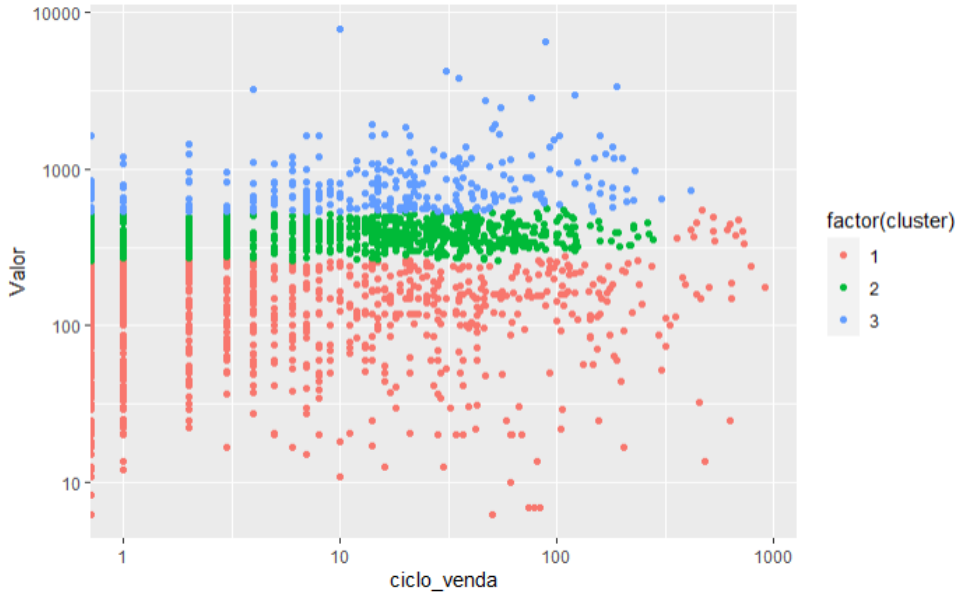
Ao observar as Figuras 17 e 18 percebe-se que em relação a variável Valor há uma maior distinção entre as observações de cada cluster, enquanto o Ciclo de Venda apresenta uma diferenciação maior apenas entre o cluster número um e os outros dois.

5.1.2 Resultado da Clusterização

A clusterização aplicada na base de dados foi capaz de gerar três grupos que reúnem observações semelhantes entre si, tendo suas diferenças ressaltadas pelas variáveis Valor e Ciclo de Venda. Como exposto pela Figura 19, a caracterização de cada cluster se dá da seguinte maneira:

A. Cluster 1: formado por 1334 negociações que apresentam de forma geral Valor até R$250,00 ou valores maiores com o ciclo de venda também mais elevado, partindo de 350 e chegando a 1000 dias.
B. Cluster 2: composto por 817 negociações que apresentam Valor entre R$250,00 e R$550,00 e ciclo de venda até no máximo 350 dias.
C. Cluster 3: constituído por 292 negociações de Valor acima de R$550,00 porém que possuem ciclo de venda até aproximadamente 350 dias.

Figura 19 - Representação do Valor e Ciclo de Venda de cada cluster



Fonte: Elaborado pelo autor.

5.2 DETERMINANTES DO SUCESSO DA VENDA

Com o intuito de criar uma estrutura que permita a predição do sucesso ou insucesso de uma negociação, a aplicação da regressão logística foi empregada para modelagem da variável dicotômica “Etapa do Negócio”.

5.2.1 Construção do modelo de Regressão Logística

Com a definição da variável resposta como a Etapa do Negócio, os demais atributos presentes na base de dados, as variáveis independentes, são testados e organizados de forma a estimar a variável dependente relacionada ao resultado da venda.

Para modelagem com a regressão logística, assim como é comum em aplicações de Machine Learning, considerou-se a divisão da base de informações disponível em um conjunto de dados para treinamento e outro para teste. De acordo com Zumel e Mount (2014), o primeiro conjunto é composto pelos registros que serão usados para alimentar a construção da regressão e o segundo grupo são os dados usados no modelo resultante para verificar a precisão da modelagem. Ainda, foi utilizado o método da validação cruzada em que os dados foram divididos em dez grupos que a cada interação formava um conjunto diferente de dados para treino e teste.

Conforme identificado na etapa de análise exploratória dos dados, os dados possuem um desbalanceamento em relação ao atributo “Etapa do Negócio”, apresentando 75% das negociações com resultado positivo de venda. Portanto, foi empregado o algoritmo Smote , que através da técnica de over-sampling replica observações em menor quantidade para se equalizar em número de classificações, resultando assim em um conjunto de dados balanceados para o treino da modelagem.

A partir desses ajustes, a construção do modelo de regressão logística relacionado ao sucesso da venda foi realizada por meio do uso da função *glm.* Enquanto o atributo “Etapa do Negócio” foi introduzido como variável dependente, os seguintes atributos tomaram lugar das variáveis independentes no modelo: Pipeline, Valor, Segmento, Origem da Oportunidade, Tags do Negócio, Ciclo de Venda, Upsell e Categoria varejo atendido. Os coeficientes resultantes do modelo assim como os valores de p são apresentados na Figura 20.

Figura 20 - Detalhes sobre as variáveis independentes do modelo

	term	estimate	std.error	statistic	p.value
	<chr>	<dbl>	<dbl>	<dbl>	<dbl>
1	(Intercept)	-109.	2625.	-0.0415	9.67e- 1
2	Valor	-0.00268	0.000229	-11.7	7.49e-32
3	Ciclo_Venda	-0.0128	0.00119	-10.8	3.70e-27
4	Pipeline_Listas.de.Varejo	-0.114	0.144	-0.789	4.30e- 1
5	Segmento_Indústria	109.	2625.	0.0415	9.67e- 1
6	Segmento_Outros	111.	2625.	0.0424	9.66e- 1
7	Segmento_Produtor...Distribuidor	109.	2625.	0.0417	9.67e- 1
8	Segmento_Restaurante.Food.Service	NA	NA	NA	NA
9	Segmento_Supermercado.Atacado	110.	2625.	0.0420	9.66e- 1
10	Origem_Outbound	1.51	0.151	10.0	1.23e-23
11	Tags_Regional.CENTRO	0.0306	0.0992	0.309	7.58e- 1
12	Tags_Regional.SUL.NE.NO	NA	NA	NA	NA
13	Upsell_Sim	0.423	0.206	2.06	3.97e- 2
14	Categoria_Varejo_B	-0.105	0.160	-0.657	5.11e- 1
15	Categoria_Varejo_C	-0.378	0.175	-2.16	3.09e- 2
16	Categoria_Varejo_sem.informação	-0.712	0.143	-4.98	6.23e- 7

Fonte: Elaborado pelo autor.

A estimação dos coeficientes resultantes representa para cada variável o efeito da categoria em relação à categoria de referência. Como o modelo de regressão logística considera a aplicação da função logarítmica, os valores dos coeficientes resultantes não podem ser interpretados diretamente, porém o seu sinal representa se a associação é positiva ou negativa em relação a variável dependente em análise. Nesse sentido, de acordo com a Figura 20, os atributos Valor, Ciclo de Venda, Pipeline com categoria Listas de Varejo e Categoria varejo atendido com valores B, C e sem informação apresentam influência negativa em relação ao sucesso da venda.

Quanto ao valor de p, este é proveniente de um teste de avaliação do modelo e tem como objetivo medir o grau de significância de cada coeficiente da equação logística, sendo estatisticamente mais significativo quanto menor o valor de p. Assim, a Figura 20, por meio dos valores de p menores do que 0,001, indicam como atributos significativos para o estudo do sucesso da venda o Valor, Ciclo de Venda, a Origem da Oportunidade e a Categoria varejo atendido.

Por conta da forma logarítmica da regressão logística, os coeficientes foram transformados por meio da base “e” para melhor interpretação dos seus resultados que é exposto na Figura 21.

Figura 21 - Coeficientes transformados em relação a forma logarítmica

term	estimate	std.error	statistic	p.value
(Intercept)	0.000000e+00	2624.52	-0.04	0.97
Categoria_Varejo_sem.informação	4.900000e-01	0.14	-4.98	0.00
Categoria_Varejo_C	6.900000e-01	0.18	-2.16	0.03
Pipeline_Listas.de.Varejo	8.900000e-01	0.14	-0.79	0.43
Categoria_Varejo_B	9.000000e-01	0.16	-0.66	0.51
Ciclo_Venda	9.900000e-01	0.00	-10.79	0.00
Valor	1.000000e+00	0.00	-11.74	0.00
Tags_Regional.CENTRO	1.030000e+00	0.10	0.31	0.76
Upsell_Sim	1.530000e+00	0.21	2.06	0.04
Origem_Outbound	4.540000e+00	0.15	10.02	0.00
Segmento_Indústria	1.988175e+47	2624.52	0.04	0.97
Segmento_Produtor…Distribuidor	3.062633e+47	2624.52	0.04	0.97
Segmento_Supermercado.Atacado	7.999603e+47	2624.52	0.04	0.97
Segmento_Outros	2.244213e+48	2624.52	0.04	0.97
Segmento_Restaurante.Food.Service	NA	NA	NA	NA
Tags_Regional.SUL.NE.NO	NA	NA	NA	NA

Fonte: Elaborado pelo autor.

Através da Figura 21, tem-se os valores dos coeficientes das variáveis significativas da equação logística que pode ser interpretado, por exemplo, da seguinte maneira: a chance de a variável Etapa do Negócio assumir valor 1 aumenta 4,54 vezes quando a variável Origem da Oportunidade assume o valor “Outbound”. Da mesma maneira, o sucesso da venda diminui pouco mais de uma vez quando o Valor aumenta em uma unidade, e 0,99 vezes em relação ao Ciclo de Venda. Já em relação a Categoria do Varejo, quando ela assume o valor “sem informação” o sucesso de venda diminui em 0,49 vezes.

5.2.2 Avaliação do modelo de Regressão Logística

O primeiro método empregado na avaliação do modelo de regressão logística foi a Matriz de Confusão, em que as observações da base de teste são comparadas com o resultado da aplicação do modelo construído nessa mesma base. A matriz de confusão resultante é apresentada na Figura 22.

Figura 22 - Matriz de confusão

		Observação	
		0	1
Predição	0	99	94
	1	54	364

Fonte: Elaborado pelo autor.

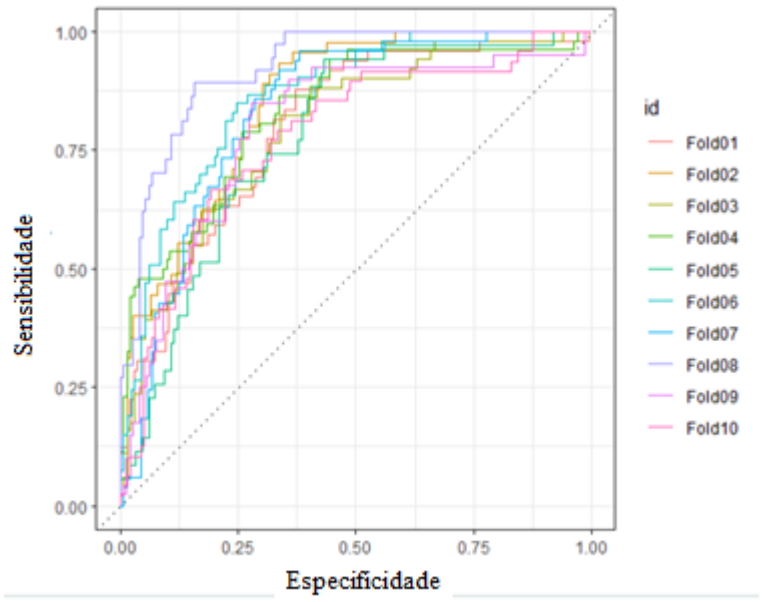
Pela matriz, percebe-se o número significativo de predições verdadeiras, sendo 99 VN e 364 VP, contra as predições incorretas, representadas pelos 54 FP e 94 FN. Através desses valores são extraídos os parâmetros a seguir:

A. Precisão: 76%
B. Sensibilidade: 67%
C. Especificidade: 79%

Logo, de acordo com a análise da matriz de confusão, o modelo apresentou um bom desempenho, tendo em vista o valor elevado da precisão, constando acerto da predição em 76% dos casos. Ao se tratar dos casos de acertos do valor 1 (sensibilidade), o valor de 67% também se mostrou satisfatório, e dos acertos do valor 0 (especificidade), a taxa de 79% também revela boa performance.

O segundo método aplicado para análise do modelo de regressão logística foi o da curva ROC, construída a partir do resultado da aplicação do modelo definido para os grupos de validação cruzada construídos, conforme apresentado na Figura 23.

Figura 23 - Curva ROC



Fonte: Elaborado pelo autor.

O desenho da curva ROC indica a boa qualidade do modelo com o traçado se aproximando da parte superior esquerda do gráfico. Além disso, o parâmetro AUC, que se relaciona com a área sob a curva, resultou em um valor de 0,83, demonstrando a alta probabilidade de sucesso nas predições do modelo.

5.2.3 Resultado da modelagem por Regressão Logística

Quanto à probabilidade de sucesso de uma venda, o modelo de regressão logística construído apresentou taxa de 76% de precisão pelo método da matriz de confusão e valor de 0,83 para o parâmetro AUC da curva ROC, evidenciando a alta probabilidade de que o modelo acerte em suas predições. Em específico, a modelagem mostrou que são determinantes para o sucesso de uma venda os atributos Valor, Ciclo de Venda, Origem da Oportunidade e Categoria varejo atendido por possuírem valor de p menor que 0,001. Seus coeficientes e influência são apresentados na Tabela 2.

Tabela 2 - Influência dos previsores no sucesso da venda

Variável	Influência	Coeficiente
Valor	negativa	1
Ciclo de Venda	negativa	0,99
Origem da Oportunidade: Outbound	positiva	4,54
Categoria varejo atendido: sem informação	negativa	0,49

Fonte: Elaborado pelo autor.

Os resultados dos coeficientes do modelo são coerentes com o contexto de negociação conhecido, uma vez que ao aumentar o valor negociado a tomada de decisão do cliente se torna mais difícil e ao se estender o ciclo de venda, nota-se uma característica de complexidade na negociação. Quanto a origem outbound, esta se relaciona com um aspecto pré determinado da negociação, em que a empresa aborda clientes com necessidade *compliance*, corroborando para o sucesso da venda. Nesse mesmo sentido, o fato de não possuir a informação de que varejo o cliente atende, é um indicativo que este não está sendo cobrado para contratação da rastreabilidade, diminuindo assim as chances de compra do produto.

5.3 ESTIMAÇÃO DO TEMPO DE NEGOCIAÇÃO

Com o objetivo de criar um modelo que fosse capaz de estimar o tempo que uma negociação levará para ser concluída, a base histórica de dados foi submetida a um algoritmo de random forest, tendo como atributo de resposta o Ciclo de Venda.

5.3.1 Construção do modelo Random Forest

Como observado na etapa de análise exploratória de dados, a variável ciclo de venda das observações possui um amplo espectro de variação, indo de -13 a 1108 dias. Os valores negativos para essa variável representam erros de registros na base, uma vez que não é possível que se tenha um ciclo de venda negativo, e por isso essas linhas foram excluídas da análise. Também, com o intuito de normalizar os dados e permitir melhor análise do atributo, os valores de ciclo de venda foram submetidos a uma transformação logarítmica.

Para aplicação do random forest, assim como na modelagem com a regressão logística, inicialmente os registros foram segmentados em uma porção para treino do modelo e outra para teste, e o método da validação cruzada foi aplicado. As variáveis selecionadas para aplicação no modelo foram: Pipeline, Valor, Segmento, Origem da Oportunidade, Tags do Negócio, Upsell e Categoria varejo atendido.

Com esses refinamentos realizados, o modelo de treinamento foi construído a partir da abordagem tradicional da função como método *rf*, para se referir ao random forest, com a métrica RMSE como medida de avaliação e foi definido a avaliação de 14 valores como número de atributos utilizados para construir cada árvore (*mtry*). Após a execução do treinamento, o algoritmo retorna a avaliação de cada *mtry* testado, em que se pode identificar o menor RMSE resultante como sendo 4, conforme apresentado na Figura 24, e que se torna o valor a ser adotado no modelo.

Figura 24 - Avaliação do modelo para 14 valores de mtry

```
mtry  RMSE       Rsquared   MAE
 2    0.6033378  0.2445364  0.4857139
 3    0.5927100  0.2527835  0.4753388
 4    0.5906737  0.2538635  0.4722779
 5    0.5912603  0.2525888  0.4717413
 6    0.5939594  0.2476937  0.4733815
 7    0.5962373  0.2445587  0.4739586
 8    0.5999916  0.2389568  0.4761336
 9    0.6027632  0.2362127  0.4769236
10    0.6069260  0.2311689  0.4786234
11    0.6103481  0.2273883  0.4801300
12    0.6134960  0.2243407  0.4813854
13    0.6155302  0.2228766  0.4824878
14    0.6169100  0.2220653  0.4830905
15    0.6169489  0.2229574  0.4832732

RMSE was used to select the optimal model using the smallest value.
The final value used for the model was mtry = 4.
```

Fonte: Elaborado pelo autor.

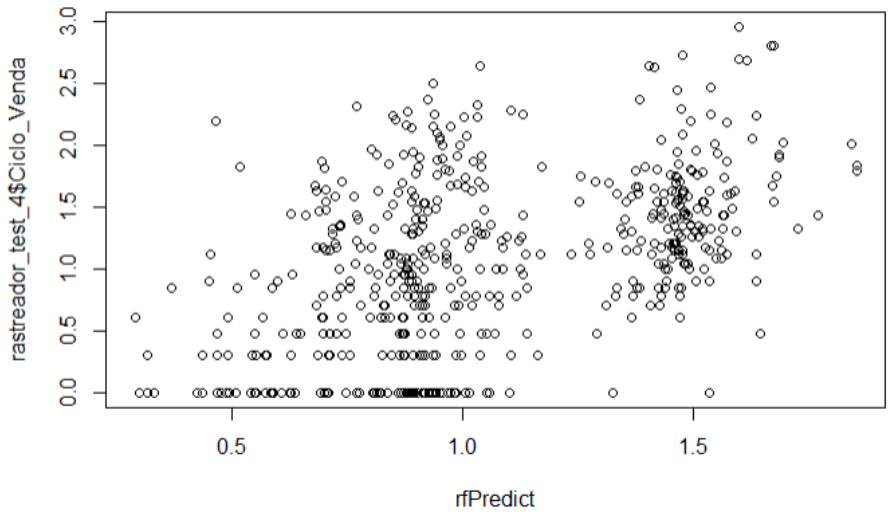
5.3.2 Avaliação do modelo Random Forest

Os métodos aplicados para a análise da qualidade do modelo de random forest construído foram o RMSE e o R2. Essas métricas foram coletadas a partir da aplicação do modelo construído na base de dados de teste, resultando em um RMSE igual a 0,60 e um R2 igual a 0,26.

Como observado, o valor RMSE para a modelagem random forest construída para a predição da variável Ciclo de Venda apresentou valor 0,60. Porém, como havia sido aplicada uma transformação logarítmica na base de dados, esse valor deve ser transformado respeitando essa aplicação de Log10. Assim, pode-se considerar que o resultado real do RMSE é de 3,98, representando que o modelo apresenta um desvio de aproximadamente quatro dias ao prever o ciclo de venda de uma negociação.

Já no que se refere a métrica R2, que apresentou valor resultante de 0,26, esta indica que o modelo construído explica em média 26% da variabilidade dos dados de resposta. Esse indicador vai ao encontro da Figura 25, que demonstra graficamente a relação entre as observações reais e as previsões.

Figura 25 - Relação entre a variável Ciclo de Venda e as predições

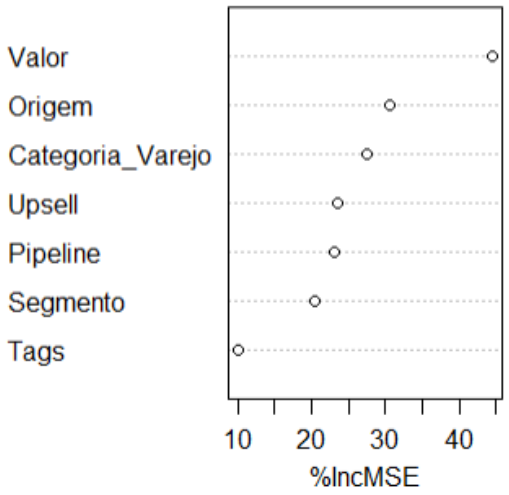


Fonte: Elaborado pelo autor.

5.3.3 Resultado da modelagem por Random Forest

Em relação ao objetivo de estimar o tempo de negociação de uma venda, o modelo construído a partir da técnica de random forest pode ser executado inserindo-se as variáveis Pipeline, Valor, Segmento, Origem da Oportunidade, Tags do Negócio, Upsell e Categoria varejo atendido e a importância de cada atributo para a estimação do Ciclo de Venda a partir do modelo é apresentada na Figura 26.

Figura 26 - Relevância dos atributos no modelo random forest



Fonte: Elaborado pelo autor.

Assim, conforme observado, o Valor é a variável mais relevante para o Ciclo de Venda, o que faz sentido na conjuntura de vendas, uma vez que as negociações tendem a se tornar mais complexas e se estenderem quando o preço é mais elevado. Em seguida, a Origem da Oportunidade e a Categoria de varejo atendido são os atributos que mais impactam no modelo, se relacionando com a característica *compliance* da negociação. Ou seja, mesmo com o RMSE indicando um desvio de quatro dias da predição em relação à realidade, o modelo construído aponta direções de entendimento e atuação em relação à duração do ciclo de venda.

6 CONCLUSÕES E RECOMENDAÇÕES

Esta seção visa apresentar as considerações finais deste trabalho, em que são revisados os estudos realizados e os resultados atingidos, além de expor as limitações e sugestões para pesquisas futuras relacionadas a área.

6.1 CONCLUSÕES

Diante da necessidade de insumos para uma tomada de decisão mais orientada a dados no setor comercial de uma empresa de tecnologia, o presente trabalho teve como objetivo geral implantar a análise preditiva de dados na área em estudo. Esta se deu a partir do armazenamento de informações executado pelo setor, que tornou possível o atingimento do propósito do estudo.

Neste contexto, alguns aspectos se destacam quanto à relevância no processo comercial, embasando os objetivos específicos abordados no estudo. Inicialmente, é importante que se possa segmentar os clientes de forma a entender suas peculiaridades e a partir disso, basear ações de primeira abordagem de lead e de processos de vendas. Para atingimento de tal finalidade, um algoritmo de clusterização foi aplicado à base de dados, possibilitando a criação de grupos que possuem registros semelhantes entre si. Como resultado, tem-se clara a relevância do Valor e do Ciclo de Venda do negócio no entendimento do comportamento dos leads.

Ao se considerar a conclusão do processo de negociação, as maiores necessidades da empresa estão em conseguir estimar a probabilidade de sucesso e o tempo necessário para finalização da venda. Neste sentido, os objetivos específicos dois e três tiveram como propósito elucidar as variáveis que impactam nesses processos, assim como criar modelos capazes de prever esses atributos para negociações futuras.

Quanto ao intuito de identificar os determinantes do sucesso de uma venda, este foi atingido por meio da construção de um modelo de regressão logística que possui como atributo de resposta a probabilidade de sucesso da negociação, e consequente previsão de perda ou ganho do negócio. A modelagem permitiu identificar a pertinência dos atributos Valor, Ciclo de Venda, Origem da Oportunidade e Categoria varejo atendido no desfecho da venda. Também, a avaliação do modelo criado constatou, a partir da precisão de 76% e do índice 0,83 para o parâmetro AUC, relevância estatística para o estudo.

O objetivo específico relacionado à estimação do tempo de negociação de uma venda, o Ciclo de Venda, foi concluído com a aplicação do algoritmo de random forest. Este apresentou valor de R2 de 26%, caracterizando a avaliação estatística como um pouco defasada, porém permitiu identificar os atributos mais importantes na constituição do Ciclo de Venda - o Valor, a Origem da Oportunidade e a Categoria de varejo atendido.

Ao fim da aplicação da metodologia proposta, o objetivo principal desta pesquisa foi atingido, gerando estruturas, conhecimentos e tópicos orientadores quanto à melhoria de processos relacionados às vendas da empresa. Além disso, a implantação de técnicas de machine learning é capaz de dar suporte a cultura de tomada de decisão orientada a dados.

Por fim, é importante salientar a relevância do Trabalho de Conclusão de Curso no processo de formação da autora, dado que o exercício de aplicação dos conhecimentos adquiridos ao longo da graduação em um contexto real do mercado de trabalho promove o estímulo da conexão entre teoria e prática, e os fundamentos necessários para resolução de futuros desafios profissionais.

6.2 RECOMENDAÇÕES

Ainda que o propósito desta pesquisa tenha sido atingido, observaram-se algumas limitações e pontos de melhoria a serem aplicados em trabalhos futuros. O fato de a base de dados ser composta principalmente por registros relacionados ao processo de negociação, faz com que os modelos construídos sejam embasados em características definidas pelos leads e pela própria empresa objeto de estudo. Assim, uma vez que o setor comercial trabalhe no armazenamento de informações dos leads como porte, número de funcionários, maturidade e tamanho da operação, recomenda-se que essas características de perfil dos clientes sejam utilizadas na construção de modelos com a finalidade de refinar estatisticamente os resultados e trazer visões diferentes acerca das questões tratadas.

Também sobre a base de dados, durante a etapa de organização dos registros disponíveis, observou-se a necessidade de tratamento ou eliminação de observações por conta de inconsistências. Para solucionar tal problemática, pode-se investir em uma maior organização da estrutura do CRM, e até mesmo em rotinas de auditoria de registros, que garantam que todos os aspectos das negociações sejam armazenados de forma correta.

Quanto a clusterização realizada, esta pode ter seu resultado aprofundado através de um estudo qualitativo em que clientes são entrevistados como forma de validar os grupos de leads

identificados. Já acerca da modelagem por regressão logística e random forest, as suas saídas podem servir como norteadoras de ações de investimento da empresa. Ao se alocar recursos na melhoria de aspectos da negociação que sabidamente causam mais impactam no sucesso e ciclo da venda, a eficiência do processo de vendas é elevada.

REFERÊNCIAS

ABUKARI, K.; JOG, V. **Business Intelligence in action**. [s.l: s.n.]. v. 77

AMARAL, F. **Introdução à Ciência de Dados**. 1. ed. Rio de Janeiro: Alta Books, 2016.

ARAÚJO, R. **Social media em vendas B2B: influência na satisfação do cliente**. Lisboa: Escola Superir de Lisboa, jul. 2019.

ATKINS, C. et al. **Unlocking the power of data in sales**. Disponível em: <https://www.mckinsey.com/business-functions/marketing-and-sales/our-insights/unlocking-the-power-of-data-in-sales>. Acesso em: 29 set. 2021.

BATTISTI, I. D. E.; SMOLSKI, F. M. DA S. **Regressão Logística - Software R**. Chapecó: Universidade Federal da Fronteira Sul, 2019.

BECKER, E. A Tecnologia da Informação Aplicada à Produção de Alimentos. **Revista da Unifebe**, v. 7, n. 7, 2002.

BEZERRA, A. A. Implantação e Uso de Business Intelligence: Um Relato de Experiência no Grupo Provider. **Revista Gestão.Org**, v. 13, p. 233–243, 2015.

BURKOV, A. **The Hundred-Page Machine Learning Book**. 1. ed. Quebec: Grupo Porto, 2019.

CARDOZO, É. **Modelo de Avaliação do Nível de Maturidade das Capabilidades Organizacionais de Business Intelligence**. Belo Horizonte: Universidade Federal de Minas Gerais, 2021.

CASSIANO, K. **Análise de Séries Temporais Usando Análise Espectral Singular (SSA) e Clusterização de Suas Componentes Baseada em Densidade**. Rio de Janeiro: Pontifícia Universidade Católica do Rio de Janeiro, 2014.

DAVENPORT, T. **Ecologia da Informação**São PauloFutura, , 1998. Disponível em: <http://pubsonline.informs.org/do/10.1287/8943f842-86f8-4d42-9a64-9a7cd07b31f5/abs/>. Acesso em: 7 jan. 2022

DOMINGOS, P. **O Algoritmo Mestre: Como a busca pelo algoritmo de machine learning definitivo recriará nosso mundo**. 1. ed. São Paulo: Novatec Editora, 2017.

DSA. **7 Principais Casos de Uso de Data Science em Vendas**. Disponível em: <https://blog.dsacademy.com.br/7-principais-casos-de-uso-de-data-science-em_vendas/>. Acesso em: 8 jan. 2022.

FREITAS, C. et al. Introdução à Visualização de Informações. v. 8, 2001.

GIL, A. C. **Como elaborar projetos de pesquisa**. 3. ed. São Paulo: Atlas, 1991.

GUAZZELLI, A. **O Que é a Análise Preditiva?** , 31 ago. 2012. Disponível em: <https://developer.ibm.com/br/articles/ba-predictive-analytics1/>. Acesso em: 11 jan. 2022

HILSDORF, C. **O que é Inteligência Competitiva?** Disponível em: <https://www.gov.br/infraestrutura/pt-br/assuntos/gestao-estrategica/artigos-gestao-estrategica/o-que-e-inteligencia-competitiva>. Acesso em: 8 jan. 2022.

HOSTMANN, B.; RAYNER, N.; HERSCHEL, G. Gartner's Business Intelligence, Analytics and Performance Management Framework. **Gartner**, p. 15, 2009.

JAMES, G. et al. **An Introduction to Statistical Learning with Applications in R**. 2. ed. New York: Springer, 2021.

KALAKOTA, R.; ROBINSON, M. **E-business 2.0: roadmap for success**. Boston: Addison-Wesley, 2001.

KUPFER, D. Padrões de Concorrência e Competitividade. dez. 2015.

LACERDA, D. P. et al. Design Science Research: método de pesquisa para a engenharia de produção. **Gestão & Produção**, v. 20, n. 4, p. 741–761, 26 nov. 2013.

LEADS2B. **O que é BI? Como usar o Business Intelligence em vendas?**, 17 ago. 2021. Disponível em: <https://leads2b.com/blog/business-inteligence/>. Acesso em: 8 jan. 2022

LIMA, M. C. C.; SOUZA, F. P. Inteligência competitiva como estratégia empresarial em micro e pequenas empresas. **ENEGEP**, p. 8, 2003.

MAKRIDAKIS, S.; SPILIOTIS, E.; ASSIMAKOPOULOS, V. Statistical and Machine Learning forecasting methods: Concerns and ways forward. **PLOS ONE**, v. 13, n. 3, p. e0194889, mar. 2018.

MARINHEIRO, A. **Análise e implementação de open source Business Intelligence**. Coimbra: Instituto Politécnico de Coimbra, 2013.

MARR, B. **Big Data in Practice**. 1. ed. Chichester: Aptara Inc, 2016.

MATOS, D. **Business Intelligence x Data Science**. Disponível em: <https://www.cienciaedados.com/business-intelligence-x-data-science/>. Acesso em: 8 jan. 2022.

MAURYA, ..PRAMOD KUMAR et al. Crop Value Forecasting using Decision Tree Regressor and Models. **European Journal of Molecular**, v. 7, n. 2, 2020.

MENEZES, E.; SILVA, E. **Metodologia da Pesquisa e Elaboração de Dissertação**. 4. ed. Florianópolis: [s.n.].

MIKROYANNIDIS, A.; THEODOULIDIS, B. Ontology management and evolution for business intelligence. **International Journal of Information Management**, v. 30, n. 6, p. 559–566, dez. 2010.

NACIONAL, I. **INSTRUÇÃO NORMATIVA CONJUNTA - INC Nº 2, DE 7 DE FEVEREIRO DE 2018**. Disponível em: <https://www.in.gov.br/materia>. Acesso em: 27 set. 2021.

NEGASH, S.; GRAY, P. **Business intelligence. In: Handbook on decision support systems**. 2. ed. Berlin: Springer, 2008.

NYCE, C. Predictive Analytics White Paper. **AICPCU/IIA**, p. 24, 2007.

OCHI, L.; DIAS, C.; SOARES, S. Clusterização em Mineração de Dados. 1 jan. 2004.

PETRINI, M.; POZZEBON, M.; FREITAS, M. **Qual é o Papel da Inteligência de Negócios (BI) nos Países em Desenvolvimento?** Curitiba: 1 set. 2004.

PONTE, C.; CAMINHA, C.; FURTADO, V. **Otimização de Florestas Aleatórias através de ponderação de folhas em árvore de regressão**. Anais do Encontro Nacional de Inteligência Artificial e Computacional (ENIAC).. Uberlândia: SBC, 20 out. 2020.

ROLLINS, J. B. **O que é Machine Learning e como utilizar? - IBM Brasil**. Disponível em: <https://www.ibm.com/br-pt/analytics/machine-learning>. Acesso em: 12 jan. 2022.

ROSS, A.; TYLER, M. **Receita Previsível: como implantar a metodologia revolucionária de vendas outbound que pode triplicar os resultados da sua empresa**. 1. ed. São Paulo: Autêntica Business, 2017.

SILVA, A. M.; SAMBONGO, E. T. T.; NOÉ, E. J. Análise de Ferramentas de BI para um Sistema Comercial Apoiado por uma proposta de Desenvolvimento baseado em Devops e SaaS. **III World Congress on Systems Engineering and Information Technology**, p. 5, 2016.

SORDI, J. O. DE. **Administração da informação: fundamento e práticas para uma nova gestão do conhecimento**. 2. ed. São Paulo: Saraiva, 2015.

TURBAN, E. et al. **Business Intelligence: um enfoque gerencial para a inteligência do negócio**. Porto Alegre: [s.n.].

TURKMAN, M. **Análise Preditiva: um pequena introdução**. Lisboa: Universidade de Lisboa, 1995.

ZUMEL, N.; MOUNT, J. **Practical Data Science with R**. 1. ed. Shelter Island: Manning, 2014.

GLOSSÁRIO

Cross-sell Venda de um novo produto a um cliente.

Inbound Processo de atração passiva de clientes por meio de estratégias de marketing.

Lead Cliente em potencial que se tem dados de contato.

Outbound Processo de atração ativa de clientes.

Pipeline Estrutura do processo de vendas onde a negociação fica alocada.

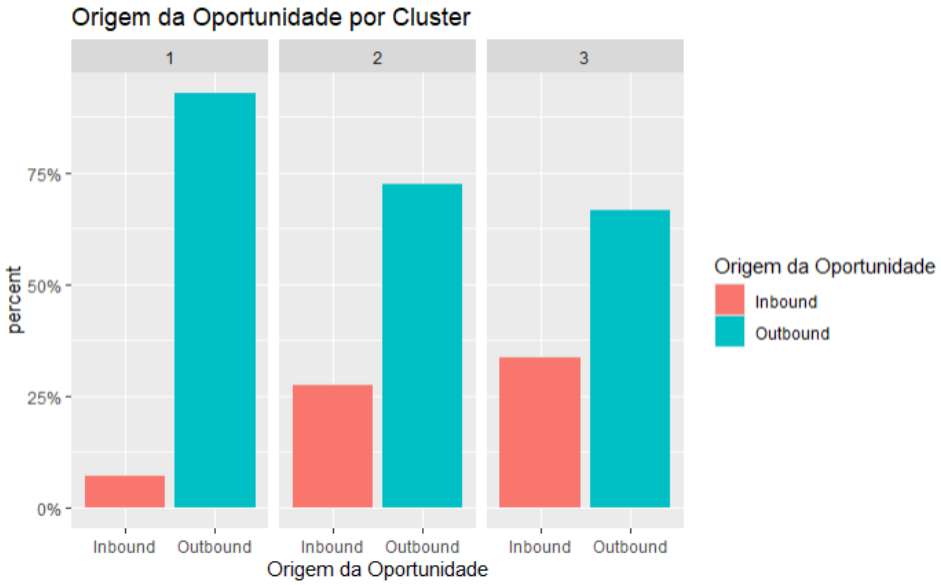
Upsell Venda de uma versão avançada do produto que o cliente já contrata.

APÊNDICE A – Gráficos dos atributos e clusters

Figura 27 - Representação da variável Upsell de cada cluster

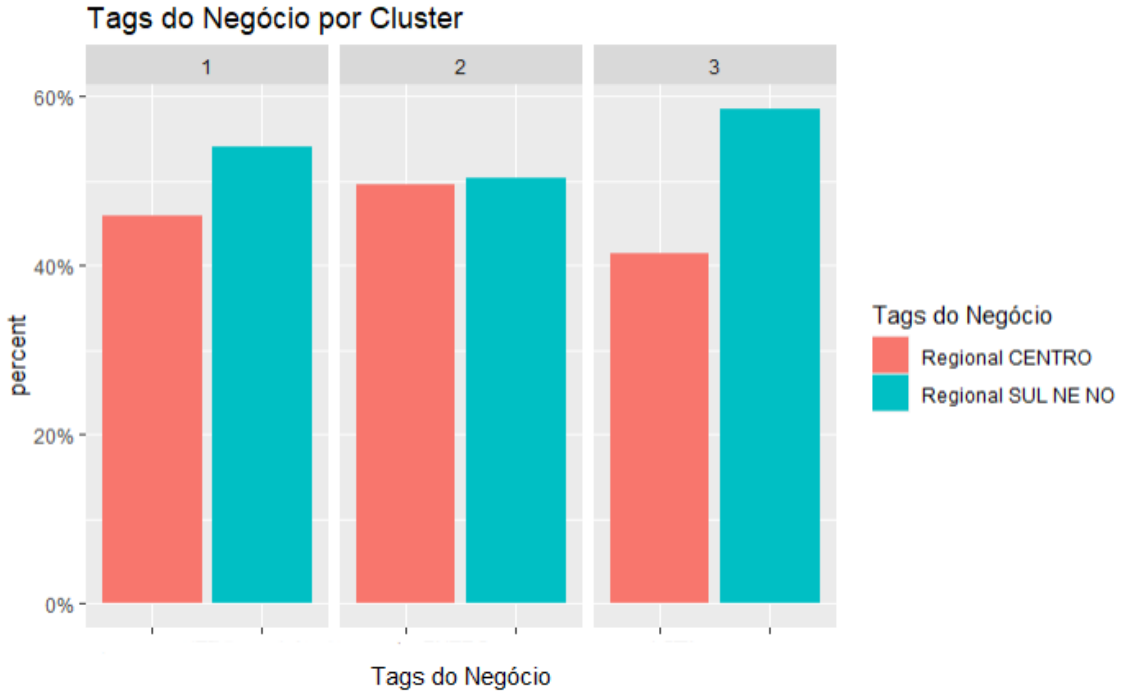
Fonte: Elaborado pelo autor.

Figura 28 - Representação da Origem da Oportunidade de cada cluster



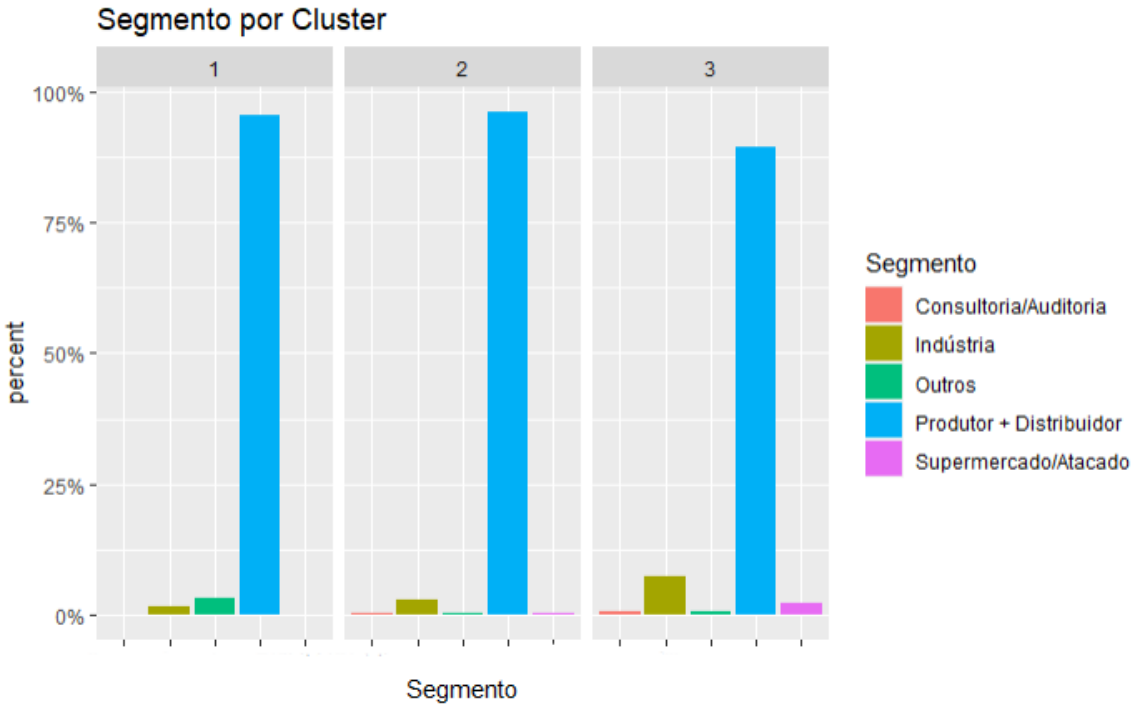
Fonte: Elaborado pelo autor.

Figura 29 - Representação das Tags de Negócio de cada cluster



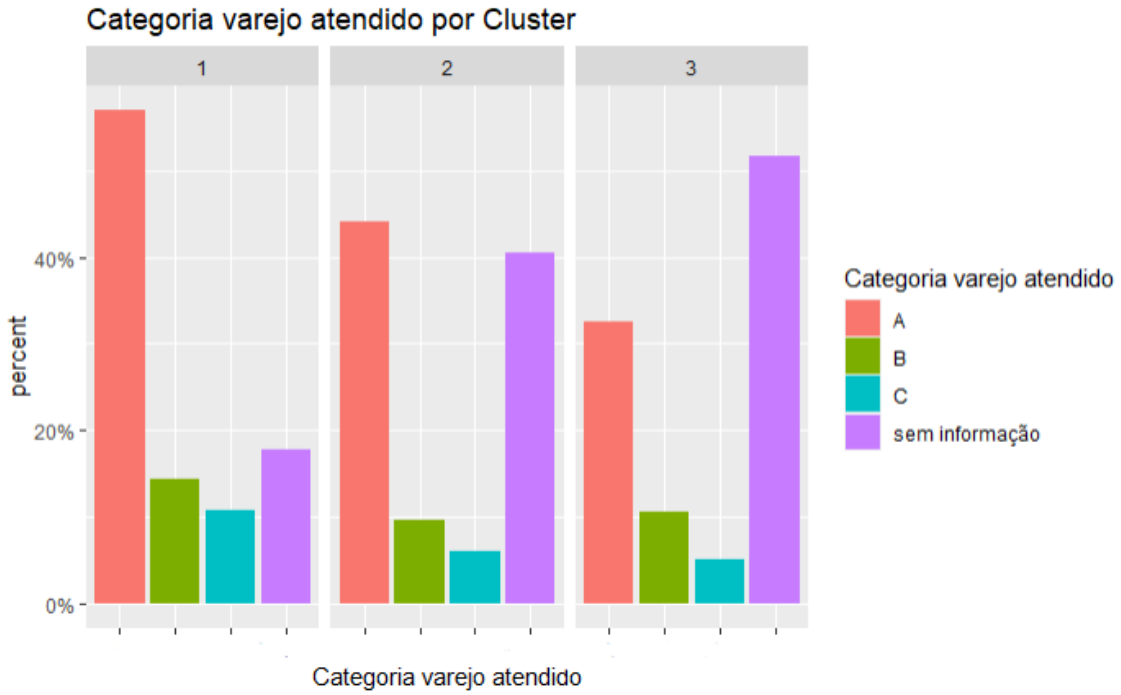
Fonte: Elaborado pelo autor.

Figura 30 - Representação do Segmento de cada cluster



Fonte: Elaborado pelo autor.

Figura 31 - Representação da Categoria varejo atendido de cada cluster

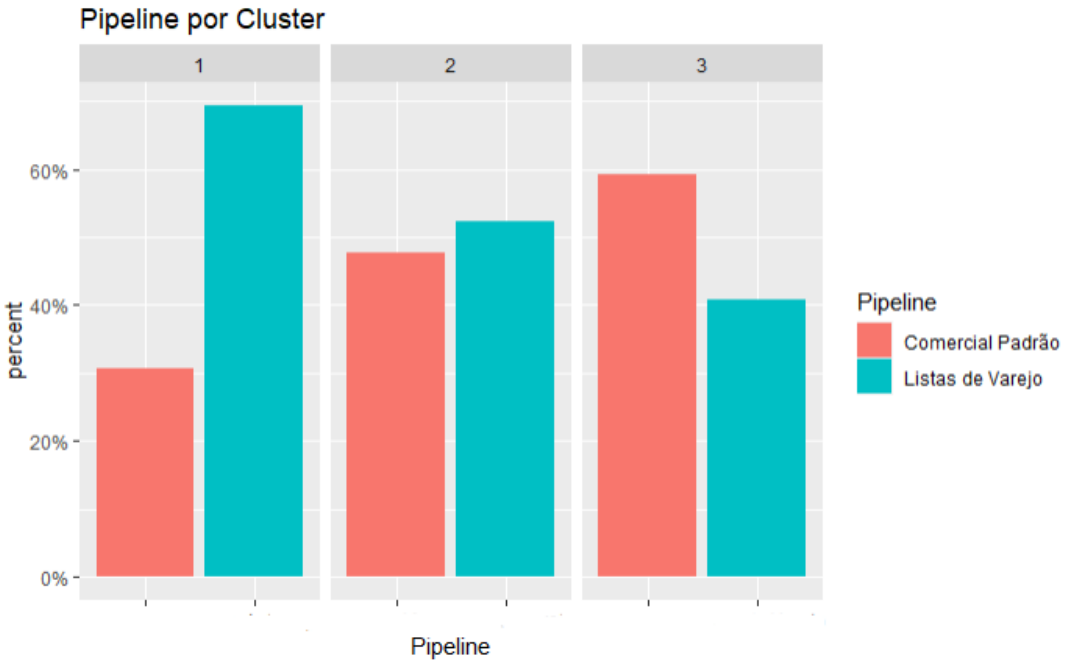


Fonte: Elaborado pelo autor.

Figura 32 - Representação da Etapa do negócio de cada cluster

Fonte: Elaborado pelo autor.

Figura 33 - Representação do Pipeline de cada cluster



Fonte: Elaborado pelo autor.