



UNIVERSIDADE FEDERAL DE SANTA CATARINA
CENTRO TECNOLÓGICO
DEPARTAMENTO DE AUTOMAÇÃO E SISTEMAS
CURSO DE GRADUAÇÃO EM ENGENHARIA DE CONTROLE E AUTOMAÇÃO

Mário Pedro Ghislandi

Desenvolvimento de um sistema computacional baseado em inteligência artificial para suporte à tomada de decisão de campanhas de marketing em uma rede de supermercados catarinense

Florianópolis
2023

Mário Pedro Ghislandi

Desenvolvimento de um sistema computacional baseado em inteligência artificial para suporte à tomada de decisão de campanhas de marketing em uma rede de supermercados catarinense

Relatório final da disciplina DAS5511 (Projeto de Fim de Curso) como Trabalho de Conclusão do Curso de Graduação em Engenharia de Controle e Automação da Universidade Federal de Santa Catarina em Florianópolis.

Orientador: Prof. Dr. Ricardo Rabelo

Supervisor: Tatiana Corrêa Góes Mendonça, Mestre

Florianópolis

2023

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

Ghislandi, Mário Pedro

Desenvolvimento de um sistema computacional baseado em inteligência artificial para suporte à tomada de decisão de campanhas de marketing em uma rede de supermercados catarinense / Mário Pedro Ghislandi ; orientador, Ricardo Rabelo, coorientadora, Tatiana Mendonça, 2023.

80 p.

Trabalho de Conclusão de Curso (graduação) - Universidade Federal de Santa Catarina, Centro Tecnológico, Graduação em Engenharia de Controle e Automação, Florianópolis, 2023.

Inclui referências.

1. Engenharia de Controle e Automação. 2. Inteligência artificial. 3. Segmentação. 4. Varejo. 5. Clustering. I. Rabelo, Ricardo . II. Mendonça, Tatiana. III. Universidade Federal de Santa Catarina. Graduação em Engenharia de Controle e Automação. IV. Título.

Mário Pedro Ghislandi

Desenvolvimento de um sistema computacional baseado em inteligência artificial para suporte à tomada de decisão de campanhas de marketing em uma rede de supermercados catarinense

Esta monografia foi julgada no contexto da disciplina DAS5511 (Projeto de Fim de Curso) e aprovada em sua forma final pelo Curso de Graduação em Engenharia de Controle e Automação

Florianópolis, 20 de dezembro de 2023.

Prof. Dr. Marcelo de Lellis Costa de Oliveira
Coordenador do Curso

Banca Examinadora:

Prof. Ricardo Rabelo, Dr.
Orientador
UFSC/CTC/DAS



Documento assinado digitalmente

Ricardo Jose Rabelo

Data: 20/12/2023 13:54:51-0300

CPF: ***.802.119-**

Verifique as assinaturas em <https://v.ufsc.br>

Tatiana Corrêa Góes Mendonça, Me.
Supervisora
Bistek Supermercados



Documento assinado digitalmente

TATIANA CORREA GOES MENDONCA

Data: 20/12/2023 18:19:29-0300

Verifique em <https://validar.iti.gov.br>

Prof. Lara Popov Zambiasi Bazzi Oberderfer, Me.
Avaliador
UFSC/CTC/DAS

Prof. Eduardo Camponogara, Dr.
Presidente da Banca
UFSC/CTC/DAS

Com profunda gratidão, dedico esta monografia aos meus pais, cujo amor e apoio ao longo dos anos são inestimáveis. Este trabalho é o resultado do amor e confiança que vocês depositaram em mim.

AGRADECIMENTOS

Gostaria de expressar meus sinceros agradecimentos a todos que contribuíram para a realização deste trabalho. Primeiramente, agradeço a Deus pela força, sabedoria e orientação ao longo deste percurso acadêmico.

À Universidade Federal de Santa Catarina, instituição na qual tive a honra de cursar minha graduação, expresse minha gratidão pela excelência no ensino oferecido, pela infraestrutura disponibilizada e pelos professores que compartilharam seu conhecimento, fundamentais para minha formação acadêmica.

Agradeço especialmente aos meus pais por serem minha fonte de inspiração e por sempre acreditarem no meu potencial. Seu amor incondicional, apoio constante e os sacrifícios feitos para que eu pudesse alcançar este momento são inestimáveis. Sem a presença e o incentivo de vocês, esta conquista não seria possível.

Ao meu orientador, Ricardo Rabelo, pela orientação sábia, paciência e constante incentivo durante este trabalho. Sua expertise e conselhos foram fundamentais para o desenvolvimento deste estudo.

Ao Bistek Supermercados, por proporcionar um ambiente propício para aplicar os conhecimentos adquiridos, permitindo-me explorar novas ideias e aplicar na prática o que aprendi.

E, por fim, à minha querida namorada Debora, pelo apoio incondicional, compreensão nos momentos de ausência e por ser minha fonte de motivação e alegria.


A todos vocês, meu profundo agradecimento. Este trabalho não teria sido possível sem o suporte e encorajamento de cada um de vocês.

DECLARAÇÃO DE PUBLICIDADE

Florianópolis, 20 de dezembro de 2023.

Na condição de representante da instituição Bistek Supermercados na qual o presente trabalho foi realizado, declaro não haver ressalvas quanto ao aspecto de sigilo ou propriedade intelectual sobre as informações contidas neste documento, que impeçam a sua publicação por parte da Universidade Federal de Santa Catarina (UFSC) para acesso pelo público em geral, incluindo a sua disponibilização *online* no Repositório Institucional da Biblioteca Universitária da UFSC. Além disso, declaro ciência de que o autor, na condição de estudante da UFSC, é obrigado a depositar este documento, por se tratar de um Trabalho de Conclusão de Curso, no referido Repositório Institucional, em atendimento à Resolução Normativa n° 126/2019/CUn.

Por estar de acordo com esses termos, subscrevo-me abaixo.

Documento assinado digitalmente
 TATIANA CORREA GOES MENDONCA
Data: 20/12/2023 18:17:10-0300
Verifique em <https://validar.itf.gov.br>

Tatiana Mendonça
Bistek Supermercados

RESUMO

O varejo alimentar no Brasil apresenta um ambiente competitivo e dinâmico. Com base nisso, o presente trabalho visa à concepção de um sistema computacional com base em inteligência artificial, com foco na segmentação de clientes do clube de uma rede de supermercados catarinense, o objetivo principal é a segmentação dos clientes em grupos distintos para o aumento da assertividade de campanhas de marketing. Inicialmente, é apresentado o contexto atual da empresa e como um novo sistema pode melhorar os processos, são levantados os requisitos desse sistema, bem como as abordagens a serem tomadas que melhor se adequam ao problema. Após isso, é proposta uma arquitetura baseada em aprendizado não supervisionado, utilizando o algoritmo de misturas gaussianas. Assim, o projeto é caracterizado pela implementação de algoritmos de aprendizado de máquina com o intuito de aprimorar o processo de segmentação de campanhas de marketing realizadas. A implementação da solução foi realizada utilizando uma base de dados da rede de supermercados Bistek, e os resultados obtidos foram explorados e discutidos.

Palavras-chave: Palavra-chave 1. Varejo alimentar, Palavra-chave 2. inteligência artificial, Palavra-chave 3. segmentação de clientes, Palavra-chave 4. algoritmo de misturas gaussianas, Palavra-chave 5. implementação de banco de dados, Palavra-chave 6. relacionamento com o cliente.

ABSTRACT

The food retail sector in Brazil presents a competitive and dynamic environment. Based on this, the current work aims to design a computer system based on artificial intelligence, focusing on segmenting customers of a club within a supermarket network in Santa Catarina. The main objective is to segment customers into distinct groups to enhance the accuracy of marketing campaigns. Initially, the current context of the company is presented, highlighting how a new system can enhance processes. The requirements for this system are outlined, along with the most suitable approaches to address the issue. Subsequently, an architecture based on unsupervised learning is proposed, utilizing the Gaussian mixture algorithm. Therefore, the project is characterized by implementing machine learning algorithms to refine the segmentation process of conducted marketing campaigns. The solution's implementation utilized a database from the Bistek supermarket network, and the obtained results were explored and discussed.

Keywords: Keyword 1. Food retail, Keyword 2. artificial intelligence, Keyword 3. customer segmentation, Keyword 4. Gaussian mixture algorithm, Keyword 5. database implementation, Keyword 6. customer relationship.

LISTA DE FIGURAS

Figura 1 – Critérios da ferramenta atual para análises de clientes	21
Figura 2 – Novo processo proposto	21
Figura 3 – Diferentes tipos de segmentação	25
Figura 4 – Arquitetura do Sistema	42
Figura 5 – Diagrama de atores	44
Figura 6 – Diagrama de implantação	45
Figura 7 – Arquitetura dos Dados	46
Figura 8 – Arquitetura simplificada do Sistema	49
Figura 9 – Distribuição do Ticket Médio pré processamento dos dados	54
Figura 10 – Histogramas dos atributos dos clusters	56
Figura 11 – Distribuição do sexo por cluster	57
Figura 12 – Quantidade de clientes por cluster	58
Figura 13 – Distribuição da variável Diversidade	59
Figura 14 – Histogramas dos atributos dos novos clusters	60
Figura 15 – Distribuição de gênero por cluster	61
Figura 16 – Quantidade de clientes por cluster	62
Figura 17 – Dashboard desenvolvida	64
Figura 18 – Histograma do Ticket Médio por Cluster	65
Figura 19 – Critério de informação Bayesiano	66
Figura 20 – Características dos clusters comparadas entre si	68
Figura 21 – Coeficiente de silhueta para 3 clusters	68
Figura 22 – Visualização da Segmentação	71

LISTA DE TABELAS

Tabela 1 – Score RFV de 15 clientes	28
Tabela 2 – Tipos de algoritmos de clusterização	33
Tabela 3 – Requisitos funcionais do projeto	41
Tabela 4 – Requisitos não funcionais do projeto	41
Tabela 5 – Variáveis do modelo	64

LISTA DE ABREVIATURAS E SIGLAS

AM	Aprendizado de Máquina
CPF	<i>Cadastro de Pessoa Física</i>
CRISP DM	<i>Cross Industry Standard Process for Data Mining</i>
CRM	<i>Customer Relationship Management</i>
IA	Inteligência Artificial
ML	<i>Machine Learning</i>
RFV	Recência, Frequência e Valor

SUMÁRIO

1	INTRODUÇÃO	14
1.1	PROBLEMÁTICA	15
1.2	OBJETIVOS	17
1.2.1	Objetivo geral	17
1.2.2	Objetivos específicos	17
2	A EMPRESA	19
2.1	HISTÓRIA	19
2.2	ÁREA DE ATUAÇÃO DO PROJETO	19
3	FUNDAMENTAÇÃO TEÓRICA	22
3.1	SEGMENTAÇÃO	22
3.1.1	Tipos de segmentação	24
3.1.2	Benefícios da segmentação	26
3.1.3	Problemas da segmentação	26
3.2	MODELO RFV	27
3.3	INTELIGÊNCIA ARTIFICIAL	29
3.3.1	Aprendizado de Máquina	29
3.3.2	Tipos de Aprendizado de Máquina	30
3.3.2.1	Aprendizado de Máquina Supervisionado	30
3.3.2.2	Aprendizado de Máquina Não Supervisionado	30
3.3.2.3	Aprendizado por reforço	31
3.4	CLUSTERIZAÇÃO	31
3.4.1	Métodos de Clusterização	32
3.4.1.1	Algoritmo de Agrupamento Baseado na Partição	32
3.4.1.2	Algoritmo de Agrupamento Baseado na Hierarquia	32
3.4.1.3	Algoritmo de Agrupamento Baseado na Teoria <i>Fuzzy</i>	34
3.4.1.4	Algoritmo de Agrupamento Baseado na Distribuição	34
3.4.1.5	Algoritmo de Agrupamento Baseado na Teoria dos Grafos	34
3.4.1.6	Algoritmo de Agrupamento Baseado em grade	35
3.4.1.7	Algoritmo de Agrupamento Baseado na Teoria Fractal	35
3.4.1.8	Algoritmo de Agrupamento Baseado em Modelo	35
3.5	MÉTRICAS DE AVALIAÇÃO DE DESEMPENHO DO MODELO	36
4	ANÁLISE DE REQUISITOS E MODELAGEM DO SISTEMA PROPOSTO	38
4.1	METODOLOGIA	38
4.2	REQUISITOS	40
4.3	ESCOLHA DAS ABORDAGENS A SEREM IMPLEMENTADAS	41
4.4	ARQUITETURA DO SISTEMA	42

5	IMPLEMENTAÇÃO DO PROJETO E INFRAESTRUTURA TECNOLÓGICA	47
5.1	BASE DE DADOS	47
5.2	FERRAMENTAS UTILIZADAS	48
5.3	ARQUITETURA	48
5.3.1	Extração dos dados	49
5.3.2	Seleção de Variáveis	49
5.3.3	Pré-processamento dos dados	50
5.3.3.1	Valores ausentes	50
5.3.3.2	Tratamento de Valores Ausentes	50
5.3.3.3	Criação de Novas Variáveis	51
5.3.4	Padronização dos Dados	52
5.4	ALGORITMO DE APRENDIZADO DE MÁQUINA	53
5.4.1	Experimentos	54
6	ANÁLISE DOS RESULTADOS E IMPACTO DA SOLUÇÃO PROPOSTA	63
6.1	MODELO DE PREDIÇÃO FINAL	64
6.2	VALIDAÇÃO DOS RESULTADOS	67
7	CONCLUSÕES E SUGESTÕES PARA TRABALHOS FUTUROS	70
7.1	TRABALHOS FUTUROS	70
7.1.1	Aumento das classes identificadas	70
7.1.2	Aumento do conjunto de dados	72
7.1.3	Integração do Treinamento Online para Atualização Dinâmica dos Algoritmos	72
7.1.4	Ajuste dos atributos	73
	REFERÊNCIAS	74
	APÊNDICE A – CÓDIGO MISTURAS GAUSSIANAS	79
	APÊNDICE B – CÓDIGO PARA AJUSTE DAS VARIÁVEIS	80

1 INTRODUÇÃO

A disputa pelo consumidor, cada vez mais exigente e informado, que tem à disposição uma oferta crescente de artigos e que apresenta diferentes perfis e hábitos, está gerando uma multiplicação de estratégias e uma certa convergência de formatos no varejo.

A concorrência está em ascensão, seja entre formatos semelhantes, distintos ou mesmo entre fornecedores e varejistas, especialmente no caso de supermercados (SANTOS; COSTA, 1997).

Enquanto Johannes Gutenberg, no final do século XV, produzia 135 cópias da Bíblia em um período de três anos por meio da inovação da impressão tipográfica, a atual era digital testemunha uma notável transformação na disseminação e no consumo de informações (CHEIDA, 2017). No contexto atual, o Google registra aproximadamente seis milhões de buscas por minuto, enquanto ocorrem cerca de 230 milhões de envios de e-mails. Além disso, a plataforma YouTube vê cerca de 500 horas de vídeo postadas a cada minuto, e as trocas de mensagens online totalizam aproximadamente 16 milhões. (JACE MCLEAN, 2022).

Os dados desempenham um papel fundamental na economia moderna, impulsionando processos de negócios e otimizando resultados. Com avanços em aprendizado de máquina e inteligência artificial, eles se tornaram um recurso valioso para as empresas.

Sua importância reside na capacidade de fornecer *insights* e previsões para aprimorar o desempenho empresarial, compreendendo o comportamento do cliente, identificando tendências de mercado e otimizando a eficiência operacional. Isso permite a tomada de decisões estratégicas informadas, aprimora as experiências do cliente e estimula a inovação (FARBOODI; VELDKAMP, 2021).

Com a ascensão do big data e o surgimento de novas tecnologias, como a computação em nuvem, os custos de processamento de dados diminuíram significativamente. Isso tornou o processamento de grandes volumes de dados economicamente viável. Como resultado, algoritmos e modelos estatísticos tornaram-se mais acessíveis e capazes de oferecer níveis mais elevados de precisão. Conforme documentado por (MANYIKA, JAMES ET AL, 2011), existem cinco abordagens fundamentais para extrair valor dos dados:

1. Criação de transparência.
2. Permitir a experimentação para descobrir necessidades, expor a variabilidade e melhorar o desempenho.
3. Segmentar populações para personalizar ações.
4. Substituir e/ou apoiar a tomada de decisão humana com automação e algoritmos.

5. Inovar modelos de negócios, produtos e serviços.

A crescente competição no varejo, impulsionada por consumidores cada vez mais exigentes e bem informados, tem levado as empresas a adotar estratégias cada vez mais diversificadas. Para enfrentar essa concorrência, é vital entender e se adaptar aos diferentes perfis e hábitos dos consumidores. É nesse contexto que a combinação de big data e algoritmos de inteligência artificial se torna um recurso poderoso para as empresas. A capacidade de processar grandes volumes de dados permite segmentar populações de consumidores de forma precisa e eficaz. Essa segmentação, por sua vez, possibilita a personalização de ações de marketing e vendas, atendendo às necessidades específicas de cada grupo de consumidores. Além disso, a automação e os algoritmos auxiliam na tomada de decisões estratégicas e na criação de modelos de negócios, produtos e serviços inovadores, tornando as empresas mais ágeis e competitivas no mercado atual. Portanto, a combinação de big data e algoritmos de IA não apenas oferece uma vantagem competitiva, mas também se torna essencial para atender às complexas demandas dos consumidores modernos.

1.1 PROBLEMÁTICA

A segmentação de mercado é amplamente reconhecida como um dos pilares essenciais do marketing moderno (WIND, 1978). Essencialmente, esse conceito parte do princípio de que as empresas podem maximizar sua lucratividade ao ajustar suas estratégias de acordo com grupos específicos de clientes que compartilham características semelhantes.

Ao segmentar os clientes, as empresas podem criar estratégias mais eficazes e personalizadas, direcionadas a cada grupo específico de clientes. Isso possibilita que as empresas ofereçam produtos e serviços mais adequados, aumentando a satisfação do cliente e, conseqüentemente, a fidelidade e retenção dos mesmos. Além disso, a segmentação de clientes também permite que as empresas otimizem seus recursos de marketing, direcionando suas campanhas e promoções para os segmentos mais lucrativos e com maior potencial de retorno (DOGAN; AYÇIN; BULUT, 2018).

Em 2022, o setor supermercadista atingiu um impressionante faturamento de R\$ 695,7 bilhões, abrangendo uma ampla gama de formatos e canais de distribuição, como supermercados, hipermercados, atacarejos, mercados de vizinhança e comércio eletrônico. Esse montante representa 7,03% do Produto Interno Bruto (PIB) nacional, evidenciando o sólido desempenho desse setor crucial para a economia.

Além disso, houve um notável crescimento na quantidade de estabelecimentos, totalizando 94.706 lojas operantes, destacando o ambiente competitivo no qual as empresas estão inseridas (ABRAS, 2022).

Segundo Roberto Butragueño Ravenga, o ambiente brasileiro do varejo está sofrendo grandes transformações devido a mudanças de comportamento do consumidor, estar preparado para esse dinamismo do mercado se mostra cada vez mais como uma necessidade das empresas.

Estamos em um momento muito desafiador. Um estudo global da NielsenIQ mostra que grande parte da população foi impactada economicamente, ou seja, 63% dos lares do mundo estão fazendo mudanças para adaptar seus orçamentos. Só que no Brasil, a parcela de pessoas com queda no poder aquisitivo é muito maior, principalmente devido a inflação. Em 3 anos, a NielsenIQ contabiliza uma alta média nos preços dos alimentos de até 30%, o que impacta por demais na vida das pessoas. Já, bebidas, itens de higiene, beleza e limpeza sofreram aumentos de até 16% no ano passado. Claro que o trabalho da indústria e do varejo é um desafio gigantesco, porque é preciso compensar essa elevação, as margens e não aparecer como vilão diante os consumidores

No escopo do varejo alimentar, algumas empresas já obtiveram sucesso com estratégias de segmentação, um exemplo é da maior rede de supermercados da Turquia, Migros, que aplicou com sucesso essa estratégia, como resultado de seu eficaz esquema de segmentação, a Migros não apenas conseguiu melhorar sua própria produtividade, mas também auxiliou os fornecedores ao fornecer análises específicas mediante solicitação, realizar comparações com concorrentes e organizar campanhas conjuntas.

As altas taxas de resposta às campanhas organizadas pela Migros, baseadas nesses métodos de segmentação, indicam que essas são abordagens bem-sucedidas e viáveis. Uma campanha de envio de correspondência especial para um segmento exclusivo pode resultar em uma taxa de resposta de 36%, podendo aumentar para até 64% em alguns segmentos de estilo de vida, como os "amantes de dieta". Essa taxa está muito acima das taxas médias de resposta de campanhas regulares, que geralmente giram em torno de 1 – 2% (COOIL; AKSOY; KEININGHAM, 2007). Para efeitos de comparação, hoje na rede Bistek a taxa média de respostas gira em torno de 5%.

Com base nisso, estratégias de segmentação se mostram como um diferencial competitivo das empresas, conseguindo aproximar as redes do cliente e gerando maior fidelidade nas compras. Ainda mais por estar inserido em um ambiente com mais de 94.000 concorrentes.

O varejo enfrenta um cenário desafiador, refletido na heterogeneidade da desaceleração do segmento pelo país. Regiões como o Nordeste (-8,3%), Norte (-7,5%) e

Sudeste (-7,3%) apresentaram quedas acentuadas em unidades de venda, associadas a variações significativas na massa total de renda em 2021.

Os dados revelam um aumento do ticket médio em 8,4% no Brasil, coincidindo com uma redução na frequência de compras (-3,2%). Essa mudança na cesta de compras dos brasileiros é resultado do encarecimento dos alimentos, impulsionado pelo aumento dos custos dos insumos e pelos desafios enfrentados na cadeia de suprimentos.

A vitória no varejo atual exige mais do que uma boa execução. É necessário um foco acentuado em uma proposta de valor diferenciada. Estratégias especializadas em categorias específicas, podem oferecer experiências de compra únicas e fomentar a fidelização dos clientes.

A análise de dados emerge como capacidade essencial no varejo. Essa ferramenta oferece uma compreensão mais profunda das necessidades do consumidor e melhoram sua experiência de compra. Além disso, permitem uma execução mais eficiente das atividades varejistas (TAMASO ROBERTO E FURTADO, 2018).

1.2 OBJETIVOS

1.2.1 Objetivo geral

O presente trabalho tem como objetivo o desenvolvimento de um sistema computacional baseado em inteligência artificial para suporte à tomada de decisão de campanhas de marketing em uma rede de supermercados catarinense.

1.2.2 Objetivos específicos

1. Coletar e analisar dados de vendas das lojas da rede.
2. Definição da estratégia para a identificação de oportunidades.
3. Definição da técnica de aprendizado de máquina para a segmentação de clientes.
4. Definição da estratégia e arquitetura para a integração do processo com o time de marketing.
5. Definição das funcionalidades presentes no sistema.
6. Implementação e avaliação do sistema computacional.

ESTRUTURA DO DOCUMENTO

O documento apresenta uma estrutura organizada e lógica, facilitando a compreensão e a navegabilidade do conteúdo. A introdução apresenta os principais elementos

do trabalho, incluindo a identificação da problemática, a definição dos objetivos gerais e específicos, e a justificativa da relevância da investigação.

A seção "A Empresa" apresenta informações detalhadas sobre a organização relacionada ao projeto, criando um contexto institucional para a compreensão do ambiente de aplicação das soluções propostas. A "Fundamentação Teórica" fornece uma base conceitual sólida para o leitor, apresentando informações relevantes relacionadas ao tópico do trabalho.

As seções subsequentes abordam as etapas práticas do projeto, começando pela análise de requisitos e modelagem do sistema proposto, que detalham como as necessidades identificadas serão traduzidas em soluções tecnológicas. A implementação do projeto e a infraestrutura tecnológica utilizada são descritas com informações técnicas e práticas.

A seção "Análise dos Resultados e Impacto da Solução Proposta" oferece uma avaliação crítica e detalhada dos resultados obtidos, demonstrando o impacto real e potencial das soluções propostas na prática. As conclusões resumem os principais achados do estudo e oferecem *insights* e sugestões para trabalhos futuros.

2 A EMPRESA

2.1 HISTÓRIA

A história da empresa Bistek Supermercados é uma narrativa que remonta ao final do século XIX na Itália, na província de Bérgamo, onde viveu um jovem chamado Césare Ghislandi. Em sua infância, Césare enfrentava a desafiadora tarefa de comprar "figo seco" na venda local, mas sua pronúncia peculiar levou os vendedores a conhecerem-no como "Fistech". Em 1892, Césare imigrou para Nova Veneza, uma colônia italiana no sul do estado de Santa Catarina.

Césare, conhecido por sua habilidade como pedreiro e sua inclinação para contar histórias cativantes, compartilhou sua peculiaridade linguística com seus amigos. Isso levou a risadas e à adoção do apelido "Bistech."

Césare teve um filho chamado Bruno, apelidado de "Naco". Em 28 de agosto de 1968, Adelino Ghislandi, filho de Césare, inaugurou uma venda no centro da cidade chamada Comercial Adelino Ghislandi. Em 1972, essa loja recebeu o nome fantasia de "Bistek," e começou a receber novidades de São Paulo e Porto Alegre, locais para os quais Adelino viajava regularmente.

Em 10 de setembro de 1976, Adelino decidiu fazer uma sociedade para expandir o negócio, embora essa parceria tenha sido posteriormente desfeita. A loja foi reaberta em 1º de novembro de 1979, evoluindo de uma loja de confecções para um supermercado que recebeu o nome "BISTEK."

Esses eventos marcaram o primeiro passo na trajetória que resultaria na consolidação da Rede BISTEK Supermercados. Nos anos subsequentes, a rede se expandiu, inaugurando lojas em várias cidades catarinenses, incluindo Cocal do Sul, Criciúma, Lages, Blumenau, Brusque, Joinville, São José, Florianópolis, Navegantes, Itajaí, Balneário Camboriú e, em 2020, a primeira loja no Rio Grande do Sul, em Porto Alegre, seguida por uma unidade em Torres em dezembro do mesmo ano.

Essa trajetória de crescimento e sucesso ao longo dos anos tornou o Bistek Supermercados uma marca de destaque no varejo do sul do Brasil, desempenhando um papel significativo na economia regional, empregando milhares de colaboradores e gerando um impressionante faturamento anual. Além de suas realizações em produtos e serviços, a empresa também investe continuamente em tecnologia para aprimorar o atendimento ao cliente e fortalecer os relacionamentos com os consumidores.

2.2 ÁREA DE ATUAÇÃO DO PROJETO

O presente trabalho foi desenvolvido dentro do projeto de desenvolvimento de ferramentas para o *Customer Relationship Management* (CRM), visando melhorar as experiências dos clientes clube Bistek, programa de vantagens oferecido pelo super-

mercado.

A área de Tecnologia da Informação (TI) responde por toda a Infraestrutura, Sistemas, Desenvolvimentos, Projetos e Inovação da Rede Bistek Supermercados, incluindo as 24 lojas, 3 CDs, Central de Produção de Padaria, Frigorífico, escritórios e Matriz. Este setor encontra-se centralizado na loja 12, situada na Costeira do Pirajubaé, Florianópolis, Santa Catarina, mas também conta com parte da equipe em Criciúma, além do apoio dos TI regionais. O projeto engloba um esforço multidisciplinar que abrange não apenas as áreas de negócios, como é o caso do CRM, mas também as áreas de sistemas e infraestrutura da própria TI.

Atualmente, a análise de clientes para campanhas de marketing é conduzida com base no conhecimento prévio dos gestores sobre o comportamento dos clientes e nas características desejadas. Esta abordagem visa ativar campanhas de produtos de acordo com o perfil dos consumidores. Por exemplo, direcionando produtos de necessidade para clientes com alta frequência de compras, que passam pouco tempo na loja, ou oferecendo descontos para clientes de baixa frequência, que realizam compras de reabastecimento e permanecem mais tempo no estabelecimento.

O processo inicia-se com a seleção da loja a ser estudada. No entanto, a ferramenta atual não permite a seleção de múltiplas lojas, o que requer a repetição do processo para cada loja individualmente. Por exemplo, se a ação for direcionada a uma mesorregião, como o Vale do Itajaí, é necessário executar esse processo para cada uma das lojas nessa região e, posteriormente, cruzar os dados obtidos. Este procedimento demanda um tempo considerável, aproximadamente 30 minutos por loja. Para avaliar uma região mais ampla, como a Grande Florianópolis, são necessárias mais de 4,5 horas. É importante ressaltar que esse processo não resulta na segmentação direta do cliente, mas sim em uma lista de consumidores que se adequam aos parâmetros estabelecidos pelo analista, pode-se visualizar esses parâmetros na Figura 3, seguindo o princípio de Pareto, onde 20% dos clientes representam 80% do segmento desejado é aplicada uma segmentação identificando os perfis de clientes.

Após todo esse processo os dados são visualizados em planilhas eletrônicas e gerados análises e gráficos para a campanha.

O trabalho atual concentra-se em auxiliar na resolução do gargalo existente no CRM por meio das análises dos clientes. Esta proposta visa reduzir o tempo despendido pelos analistas nessa tarefa e fornecer dados cruciais para embasar as decisões da empresa. Em vez de depender exclusivamente do conhecimento tácito dos gerentes e analistas, busca-se uma abordagem orientada por dados, visando padronizar e tornar mais uniforme toda a operação. Esse novo processo pode ser esquematizado na Figura 2

Figura 1 – Critérios da ferramenta atual para análises de clientes

Alterar Análise de Produtos

Código: 10.302 Status: Processado - 14/08/23 13:34 ANA.BARP

Descrição: PUBLICO DE ADOCANTES XILITOL 2023

Mala: ...

Campanha: ...

Assunto:

Seção:

Grupo:

Subgrupo:

Produtos:

Finalizadora:

Fornecedor: ... TODOS

Loja: ... TODAS

Período: a

Ranking Máximo: Somente clientes com cartão CLUBE BISTEK

Ordem: Quantidade Total Compras Rentabilidade

Fonte: Software Intellisys, captura em 14/11/2023

Figura 2 – Novo processo proposto



Fonte: Autor

3 FUNDAMENTAÇÃO TEÓRICA

3.1 SEGMENTAÇÃO

A segmentação, conceitualmente, envolve a ação de dividir um conjunto em grupos homogêneos, ou seja, agrupar pessoas que compartilham perfis semelhantes. Essa prática é realizada com o propósito de personalizar a oferta de produtos, comunicações e o modo de atendimento de forma apropriada aos clientes que possuem necessidades semelhantes (KOTLER, 1998).

É um processo de identificação e classificação de grupos de consumidores que compartilham características suficientemente semelhantes. Esse processo possibilita o planejamento e a criação de produtos ou serviços que atendam de forma específica às necessidades de cada grupo identificado (GIANESI; CORRÊA, 1994).

Para adquirir uma posição competitiva, inúmeras organizações têm aderido à estratégia de marketing segmentado. Ao invés de dissipar seus recursos em esforços de marketing dispersos, essas empresas direcionam seus esforços para públicos consumidores aos quais podem satisfazer de maneira efetiva. Essa nova abordagem permite uma alocação mais assertiva de recursos, otimizando o impacto das estratégias comerciais (KOTLER *et al.*, 2018).

Grandes empresas, principalmente no varejo, necessitam de estratégias específicas para criar proximidade com seus clientes e conseguir vender seus produtos e serviços. O desafio reside na capacidade de identificar esse público, uma tarefa que se torna ainda mais complexa considerando sua ampla dispersão geográfica pelo país, bem como sua diversidade de necessidades, desejos e a crescente variação de hábitos e costumes (RICHERS, 2000).

Ao identificar um conjunto de indivíduos com características similares, presume-se que as necessidades de consumo desse grupo sejam semelhantes. A seleção cuidadosa do modelo de segmentação é crucial para alcançar os objetivos estabelecidos.

Os resultados da segmentação podem ser utilizados para embasar um amplo conjunto de decisões de negócios e posicionamento da empresa, alterando estratégias e campanhas de marketing e se aproximando do consumidor. Segundo (TYNAN; DRAYTON, 1987), a segmentação de mercado é frequentemente utilizada para abordar as seguintes questões:

- **Definir um mercado:** A segmentação de mercado desempenha um papel essencial na análise do cenário empresarial sob a ótica do consumidor. Ela reconhece que os consumidores não formam um grande grupo uniforme, mas, em vez disso, consistem em diversos segmentos com diferentes preferências e necessidades. Isso significa que produtos ou serviços que eram considerados concorrentes di-

retos podem ser percebidos de maneira diferente pelos clientes, dependendo de suas preferências e percepções individuais.

- **Justificar princípios de ação para marcas e produtos:** Direciona estratégias para aumentar a retenção de clientes, converter compradores de outras marcas ou atrair um novo grupo de compradores para os produtos e/ou serviços ofertados.
- **Posicionar as marcas e os produtos:** Dado a variedade de segmentos de mercado, as empresas devem concentrar seus esforços nos grupos que apresentam maior potencial de retorno.
- **Identificar lacunas no mercado:** A segmentação de mercado também possibilita identificar grupos de consumidores cujas necessidades não estão sendo supridas. Essas demandas podem ser atendidas por meio do lançamento de novos produtos ou pela adaptação de um produto ou serviço já existente.

E ainda, segundo (KOTLER *et al.*, 2018), para serem úteis, os segmentos de mercado devem atender preferencialmente aos cinco critérios a seguir:

- **Mensuráveis:** O tamanho, o poder de compra e as características dos segmentos devem ser passíveis de mensuração.
- **Substanciais:** Os segmentos devem ser grandes e rentáveis o suficiente para serem atendidos. Um segmento deve ter o maior grupo homogêneo possível.
- **Acessíveis:** Deve ser possível alcançar e atender ao segmento.
- **Diferenciáveis:** Os segmentos são conceitualmente distintos e respondem de maneira diferente a cada elemento e programa do mix de marketing.
- **Acionáveis:** Deve ser possível desenvolver programas efetivos para atrair e atender aos segmentos.

A segmentação de mercado apresenta diversos desafios e complexidades para os profissionais de marketing. Aqui estão alguns dos principais desafios associados à segmentação de mercado:

- **Complexidade na Compreensão do Cliente:** A segmentação de mercado evoluiu de um foco simples na compreensão das necessidades dos clientes para uma exploração mais intrincada dos estilos de vida, valores, tarefas a serem realizadas, estados de necessidade e ocasiões dos clientes. Essa complexidade torna desafiador para os profissionais de marketing identificar e compreender com precisão as diversas necessidades e preferências dos diferentes segmentos de clientes.

- **Análise de Dados em Tempo Real:** A disponibilidade de dados digitais em tempo real sobre os clientes abriu novas maneiras de compreendê-los. No entanto, analisar e dar sentido a essa vasta quantidade de dados pode ser uma tarefa complexa, que acaba demandando muito tempo desses profissionais, enquanto que esse esforço poderia estar sendo despendido em outras tarefas.
- **Integração com as Funções do Negócio:** A segmentação de mercado tem o maior impacto quando é compreendida e utilizada por todas as funções relevantes dentro de uma organização. No entanto, se a segmentação não estiver ativamente envolvida e incorporada ao planejamento estratégico e aos processos de tomada de decisão, ela pode não alcançar todas as funções relevantes. Alcançar o máximo impacto requer alinhamento dos stakeholders e o envolvimento de funções-chave desde o início de um programa de segmentação.
- **Treinamento e Aplicação de Insights:** Após a conclusão de um programa de segmentação, é crucial fornecer treinamento adequado a todos os usuários finais e equipá-los com ferramentas para aplicar os insights do cliente em suas decisões diárias. Sem treinamento adequado e suporte, os resultados da segmentação podem ser ignorados ou subutilizados por funções comerciais críticas.
- **Comportamento do Cliente em Constante Evolução:** O comportamento do consumidor está em constante mudança, influenciado por diversos fatores, como avanços tecnológicos, tendências sociais e condições econômicas. Os profissionais de marketing precisam adaptar continuamente suas estratégias de segmentação para acompanhar essas mudanças constantes no comportamento e preferências dos clientes.
- **Equilibrar Metas de Curto e Longo Prazo:** Embora a segmentação de mercado seja frequentemente usada para impulsionar vendas e participação de mercado a curto prazo, também é essencial para a expansão e crescimento do negócio a longo prazo. Os profissionais de marketing precisam encontrar um equilíbrio entre atingir metas imediatas de vendas e alinhar estratégias de segmentação com os objetivos de longo prazo da organização.

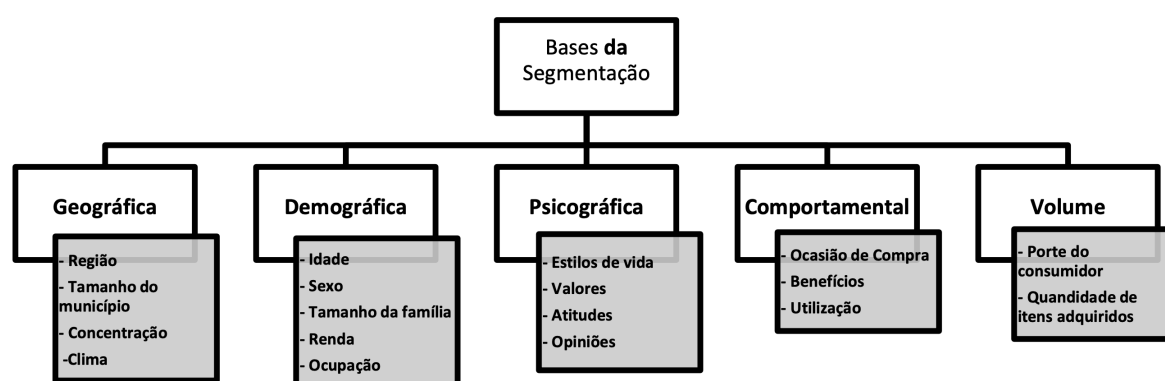
Em resumo, a segmentação de mercado é um processo complexo que requer uma compreensão profunda dos clientes, análise eficaz de dados, integração com funções comerciais e a capacidade de se adaptar a mudanças no comportamento do cliente (CHAUBEY; SUBRAMANIAN, 2020).

3.1.1 Tipos de segmentação

Existem diversas formas de segmentar o mercado, cada uma levando em conta dimensões e pesos diferentes para os dados, que dependem fortemente do produto

ou serviço prestado. As diferentes formas são classificadas por (KOTLER *et al.*, 2018), por (CHURCHILL; PETER, 2005) e citados por (BERNUZZI, 2022), que definem as segmentações como sendo geográfica, segmentação demográfica, psicográfica, comportamental e segmentação por volume, Figura 3.

Figura 3 – Diferentes tipos de segmentação



Fonte: (BERNUZZI, 2022) p.14

- **Segmentação Geográfica:** Nesse tipo de segmentação, a divisão do mercado é feita de acordo com parâmetros geográficos, delimitada em regiões, países, estados, bairros, ou outras divisões geográficas (KOTLER *et al.*, 2018).
- **Segmentação Demográfica:** A segmentação demográfica é fundamentada nas características da população, abrangendo elementos como idade, sexo, escolaridade, nível de renda, ocupação, entre outros. Essas variáveis demográficas são amplamente empregadas pelas empresas devido à sua praticidade e, frequentemente, as organizações optam por combinar duas ou mais delas, como idade e escolaridade, ou sexo, ocupação e nível de renda, para uma segmentação mais precisa do mercado. (CHURCHILL; PETER, 2005).
- **Segmentação Psicográfica ou Socioeconômica:** A segmentação psicográfica é orientada pelo estilo de vida dos consumidores, considerando elementos como classe social, personalidade, atitudes e percepções individuais. Ela proporciona uma descrição profundamente detalhada do mercado-alvo, oferecendo uma compreensão íntima do funcionamento interno dos potenciais consumidores.
- **Segmentação comportamental:** Os clientes são agrupados de acordo com seu grau de familiaridade com um produto, suas atitudes em relação a ele, seu padrão de uso e suas respostas. Muitos especialistas em marketing consideram que as variáveis comportamentais são os pontos iniciais mais eficazes para a construção de segmentos de mercado (KOTLER *et al.*, 2018).

- **Segmentação por volume:** nesse tipo de segmentação, o mercado é dividido entre consumidores de pequeno, médio e grande porte, considerando-se a quantidade de itens que eles adquirem.

3.1.2 Benefícios da segmentação

O principal objetivo de uma empresa ao desenvolver e implementar uma estratégia de segmentação de mercado é atender às necessidades dos clientes de forma mais eficaz e assertiva, o que melhora suas vantagens competitivas. De acordo com (WEINSTEIN, 1995), a análise e a estratégia de segmentação de mercado oferecem quatro benefícios significativos para as empresas:

1. **Desenvolver produtos que atendam de forma eficaz às necessidades do mercado:** Empresas que adotam essa abordagem colocam o consumidor em primeiro lugar, projetando produtos que se alinham com as demandas do público-alvo.
2. **Elaborar estratégias promocionais eficazes e de baixo custo:** Isso permite a criação de campanhas publicitárias apropriadas, direcionadas aos veículos de mídia corretos, resultando em uma alocação mais eficiente dos recursos de marketing.
3. **Avaliar a concorrência, especialmente a posição de mercado da empresa:** A pesquisa de segmentação oferece um mecanismo de inteligência competitiva para comparar o desempenho da empresa em relação aos padrões do setor, possibilitando uma tomada de decisões mais informada.
4. **Fornecer *insights* para as estratégias de marketing em vigor:** Ao revisar regularmente as estratégias, a empresa pode capitalizar novas oportunidades e evitar possíveis ameaças, adaptando-se de maneira ágil ao ambiente de negócios em constante evolução.

3.1.3 Problemas da segmentação

A adoção da estratégia de segmentação no mundo dos negócios não ocorre sem desafios e limitações. A transição de um modelo de marketing de massa para um focado na segmentação é um processo que demanda tempo, investimento e requer um comprometimento corporativo substancial. Além disso, é importante notar que a segmentação resulta em perfis genéricos, não individuais, e, portanto, não pode substituir completamente as interações pessoais para alcançar compradores específicos (RICHERS, 2000).

Os projetos de segmentação de mercado são empreendimentos que podem demandar um período significativo para sua conclusão. Dada a sua forte dependência

dos dados, tornam-se especialmente sensíveis a quaisquer anomalias ou deficiências nos mesmos. Assim, para que os dados possam ser considerados confiáveis e viáveis para uso em projetos de segmentação, é imperativo que atendam a uma série de requisitos fundamentais.

Primeiramente, a precisão dos dados é de suma importância, uma vez que qualquer imprecisão pode levar a conclusões equivocadas e ações inapropriadas. A consistência dos dados, tanto ao longo do tempo como entre diferentes fontes, é igualmente crucial, a fim de evitar ambiguidades e conflitos. A atualidade dos dados é um requisito essencial, uma vez que informações desatualizadas podem levar a decisões inadequadas.

Além disso, os dados devem ser completos, ou seja, devem abranger todas as informações necessárias para a análise e segmentação. A disponibilidade dos dados quando necessário é um fator determinante para a eficácia do projeto. Portanto, é imperativo que os dados estejam prontamente acessíveis para apoio à tomada de decisões.

3.2 MODELO RFV

O modelo Recência, Frequência e Valor (RFV) é composto principalmente por três variáveis: **Recência**, **Frequência** e **Valor**. De acordo com (WEI; LIN; WU, 2010) essas dimensões são conceituadas como:

- **Recência**: Representa a quantidade de dias que se passou desde a última data de compra realizada em uma das lojas, considerando dias corridos.
- **Frequência**: Representa a quantidade de vezes que um cliente realizou as suas compras durante o período analisado.
- **Valor**: Definido como o montante total ou médio, em unidades monetárias, gasto pelo cliente em suas compras.

Com base nessas dimensões, é viável a atribuição de pontuações aos clientes, permitindo, desse modo, a segmentação em distintos grupos. Além disso, essas métricas podem ser empregadas em conjunto com outras informações disponíveis para fins de segmentação. Portanto, cabe à empresa a responsabilidade de estabelecer a abordagem a ser adotada na compreensão e categorização de comportamentos semelhantes.

Um possível caminho, de acordo com (NIMBALKAR; SHAH, 2013), é de ordenar as três dimensões de maneira ascendente/descendente e atribuir uma pontuação aos clientes. Cada uma das três dimensões é utilizada para classificar os clientes com base em seus respectivos scores. No caso da dimensão *recência*, a ordenação segue

Tabela 1 – Score RFV de 15 clientes

CID	Rec.	R	CID	Freq.	F	CID	Mon.	M	CID	RFM
12	1	5	9	15	5	9	2430	5	1	544
1	3	5	2	10	5	12	1410	5	2	454
15	4	5	12	9	5	8	950	5	3	111
7	5	4	11	8	4	2	940	4	4	222
11	5	4	1	6	4	11	840	4	5	333
2	6	4	10	5	4	1	540	4	6	222
10	10	3	5	4	3	10	190	3	7	433
5	14	3	7	3	3	5	169	3	8	115
14	17	3	13	3	3	7	130	3	9	155
4	21	2	14	2	2	4	64	2	10	343
13	24	2	4	2	2	6	55	2	11	444
6	32	2	6	2	2	13	54	2	12	555
9	33	1	15	1	1	14	44	1	13	232
3	45	1	3	1	1	15	32	1	14	321
8	50	1	8	1	1	3	30	1	15	511

Fonte: O autor

uma lógica crescente, de forma que os consumidores são dispostos na lista com base em suas compras mais recentes. Quanto às outras dimensões, a ordenação segue um critério decrescente, mantendo no topo da lista aqueles clientes que efetuam as compras mais frequentes ou que apresentam maior volume de gastos.

Na segunda etapa os atributos são agrupados em 5 grupos de 20%, e é designado a pontuação 5 ao grupo que mais contribui, em seguida, 4 ao próximo grupo com maior contribuição, e assim por diante, até 1, conforme tabela 1.

Após determinar a pontuação ainda é possível utilizar um algoritmo de clusterização para identificação dos possíveis grupos de clientes. Naturalmente, os clientes com interações mais recentes, mais frequentes e de maior valor são apontados como clientes mais valiosos.

Pode-se notar que clientes com pontuação (454), possuem Recência alta (4), Frequência alta (5), e Valor alto (4), assim podem ser caracterizados como clientes *campeões*, aqueles que visitam muito a loja e gastam muito. Outra análise é identificar clientes com *score* baixos, como (115), que podem ser clientes que estão comprando os produtos em outras lojas, pois possuem uma Frequência e uma Recência baixas, ambas (1), mas um valor alto (5), logo eles consomem muito o produto mas deixaram de comprar na loja.

Ao tabular as possibilidades, identificamos 125 combinações distintas, resultando em 125 estratégias potenciais para aplicar em cada combinação. Alternativamente, uma abordagem mais simplificada envolve a divisão de cada dimensão em tercis, com a atribuição de apenas 3 notas (sendo 3 a mais alta e 1 a mais baixa). Isso reduz o número de combinações para 27, tornando o processo mais gerenciável e

adaptável às necessidades específicas.

3.3 INTELIGÊNCIA ARTIFICIAL

Segundo (RUSSELL; NORVIG, 2004), a Inteligência Artificial (IA) é uma área da computação que busca empregar métodos e dispositivos capazes de reproduzir a capacidade humana de raciocínio e solução de problemas. Sua origem remonta à Segunda Guerra Mundial, quando surgiu a necessidade de desenvolver tecnologias voltadas para a indústria bélica. A IA representa, assim, uma área de estudo que se propõe a replicar a inteligência humana por meio de sistemas computacionais avançados, com aplicações diversas em áreas como a indústria, medicina, ciências da computação, entre outras.

E ainda, segundo (LOBO, 2018), a IA por meio da aplicação de algoritmos desenvolvidos por especialistas, demonstra a habilidade de identificar um problema ou tarefa a ser executada. Ela é capaz de analisar dados e tomar decisões, simulando a capacidade humana. Embora sistemas computadorizados de apoio à tomada de decisões existam há várias décadas, os avanços na velocidade de processamento e capacidade de armazenamento dos computadores possibilitaram a análise de grandes volumes de dados em frações de segundo.

3.3.1 Aprendizado de Máquina

De acordo com (MONARD; BARANAUSKAS, 2003), o Aprendizado de Máquina (AM) é uma área da inteligência artificial que tem como propósito principal o desenvolvimento de técnicas computacionais voltadas para o processo de aprendizado. Além disso, busca a criação de sistemas capazes de adquirir conhecimento de maneira autônoma, fundamentando-se na premissa de que tais sistemas podem aprimorar seu desempenho por meio da análise de dados. Em sua essência, um sistema de aprendizado corresponde a um software que toma decisões com base em informações e experiências acumuladas, resultantes da resolução bem-sucedida de problemas anteriores.

Segundo (MITCHELL, 1997), um processo de aprendizado se dá a partir da experiência adquirida a respeito de uma classe de tarefas, o que resulta em um aumento progressivo no desempenho do processo à medida que mais experiências são acumuladas. Quando se aplica esse princípio de aprendizado a algoritmos computacionais, emerge o conceito de AM, frequentemente referido pela sigla *Machine Learning* (ML).

No âmbito do AM, sistemas computacionais são treinados para extrair conhecimento de eventos anteriores. Isso é realizado por meio de um processo de inferência conhecido como indução, que permite que esses algoritmos deduzam conclusões a partir de um conjunto de exemplos. Dessa forma, esses algoritmos têm a capacidade

de induzir uma função ou hipótese que seja competente na resolução de um problema com base em dados que representam instâncias do problema subjacente.

3.3.2 Tipos de Aprendizado de Máquina

3.3.2.1 Aprendizado de Máquina Supervisionado

O Aprendizado de Máquina Supervisionado refere-se a algoritmos que se baseiam na utilização de um conjunto de dados para treinamento, com o objetivo de identificar padrões nas relações entre entradas e saídas. Após aprenderem essas relações, esses algoritmos são capazes de fornecer saídas adequadas para novas entradas. Em outras palavras, essas técnicas de aprendizado supervisionado trabalham com dados que incluem as respostas esperadas (ou valores da variável desejada). O sistema se adapta de modo que, quando fornecida uma entrada, ele seja capaz de gerar uma saída correspondente com base nas características do dado.

No aprendizado supervisionado, o conjunto de treinamento contém tanto as entradas quanto as saídas, sendo as entradas consideradas como as variáveis a partir das quais o modelo é construído e as saídas representando as respostas esperadas para cada amostra de treinamento. Um exemplo de aprendizado supervisionado seria a classificação de e-mails como *spam* ou *não spam* em um filtro de spam de e-mail. Nesse caso, o conjunto de dados de treinamento consistiria em uma coleção de e-mails previamente classificados como *spam* ou *não spam* com base em seu conteúdo. As entradas seriam os atributos dos e-mails, como palavras-chave, remetentes, formatação, etc., e as saídas seriam as categorias *spam* ou *não spam*. O modelo de aprendizado de máquina é treinado com base nessas entradas e saídas para aprender a distinguir entre os dois tipos de e-mails. Esse tipo de aprendizado é denominado supervisionado, pois há conhecimento prévio das saídas correspondentes para cada exemplo, possibilitando a avaliação da capacidade do modelo gerado para prever os valores de saída para novos dados. Os modelos preditivos seguem essa abordagem, em que a saída desejada é conhecida para fins de treinamento e posterior previsão.

3.3.2.2 Aprendizado de Máquina Não Supervisionado

No aprendizado não supervisionado, o conjunto de dados de treinamento é composto exclusivamente por entradas, sem informações de saída correspondentes. Essa abordagem é utilizada para descobrir estruturas e padrões subjacentes em um conjunto de dados não rotulado. O algoritmo de aprendizado é desafiado a identificar tendências e situações que podem ocorrer naturalmente nos dados. Um exemplo prático é a tarefa de identificar espécies a partir de um conjunto de dados de DNA, onde as espécies e sua quantidade não são previamente conhecidas.

Este tipo de aprendizado é valioso quando se deseja explorar relações implícitas em um conjunto de dados e descobrir agrupamentos naturais de informações. Os algoritmos não supervisionados, conhecidos como modelos descritivos, podem revelar insights significativos e estruturas intrínsecas que podem passar despercebidos em análises manuais.

O aprendizado não supervisionado permite explorar e descrever dados não rotulados, buscando por estruturas e tendências intrínsecas, um exemplo comum de Aprendizado de Máquina Não Supervisionado é a segmentação de clientes em um conjunto de dados de compras ou comportamento de consumo. Nesse caso, os dados de entrada podem incluir informações sobre as compras anteriores dos clientes, como tipos de produtos adquiridos, frequência de compra, valor gasto, entre outros. Um algoritmo de aprendizado não supervisionado pode ser aplicado a esses dados para agrupar os clientes em segmentos com base em padrões de comportamento semelhantes. O algoritmo tentará identificar grupos de clientes que compartilham características semelhantes, como preferências de produtos, hábitos de compra ou frequência de compra.

3.3.2.3 Aprendizado por reforço

O aprendizado por reforço é uma área da aprendizado de máquina que se concentra em como agentes inteligentes devem tomar ações em um ambiente para maximizar a noção de recompensa acumulativa. Ele difere dos paradigmas tradicionais de aprendizado de máquina, como o aprendizado supervisionado e o aprendizado não supervisionado. Em vez de depender de pares de entrada/saída rotulados ou da correção explícita de ações sub ótimas, o aprendizado por reforço se concentra em encontrar um equilíbrio entre *exploration*, que envolve experimentar novas opções que podem resultar em resultados melhores no futuro, e *exploitation*, que envolve escolher a melhor opção conhecida com base em experiências passadas.

Nesse contexto, um agente é modelado para operar em um ambiente e tomar ações que resultem em recompensas. A única supervisão que o agente possui é a recompensa que recebe por suas ações. O objetivo principal é maximizar essa recompensa, e o agente tem a liberdade de escolher suas ações, sem receber instruções detalhadas. O agente se orienta pelo *feedback* das recompensas e busca, de forma autônoma, maximizá-las.

3.4 CLUSTERIZAÇÃO

A clusterização é um processo de classificação não supervisionada de padrões, que podem ser representados por observações, itens, dados, vetores, entre outros, em grupos denominados clusters. Em contraste com o conceito de classificação, a cluste-

rização é uma técnica mais fundamental, uma vez que não pressupõe a existência de categorias predefinidas para os grupos. A premissa subjacente à clusterização é que os elementos pertencentes a um mesmo cluster devem exibir uma alta semelhança entre si, enquanto devem ser substancialmente diferentes daqueles pertencentes a outros clusters. A principal vantagem da utilização da clusterização reside na capacidade de agrupar dados semelhantes de forma eficaz, permitindo uma descrição mais precisa das características inerentes a cada grupo, o que, por sua vez, proporciona uma compreensão mais profunda do conjunto de dados original.

3.4.1 Métodos de Clusterização

A clusterização denota um amplo conjunto de técnicas empregadas na identificação de subgrupos, também chamados de *clusters*, em um conjunto de dados. Essa abordagem visa particionar observações de forma a criar grupos distintos, nos quais as observações compartilhem semelhanças consideráveis, enquanto exibem notáveis diferenças em relação a observações de outros grupos. A clusterização representa um problema amplamente reconhecido na literatura de análise de dados, sendo amplamente aplicado em contextos como segmentação de clientes, classificação e análise de tendências. Vale ressaltar que se trata de um problema de aprendizado não supervisionado, uma vez que a intenção é identificar a estrutura subjacente, no caso, os *clusters* distintos, fundamentando-se exclusivamente nos dados disponíveis (WITTEN; JAMES, 2013). Pode-se verificar os principais algoritmos de clusterização atuais na Tabela 2.

3.4.1.1 Algoritmo de Agrupamento Baseado na Partição

A ideia fundamental desses algoritmos de clusterização é considerar o centro dos pontos de dados como o centro do *cluster* correspondente. *K-means* e *K-medoids* são os dois mais renomados desta categoria. O ponto principal do *K-means* está na atualização do centro do *cluster*, representado pelo centro dos pontos de dados, por meio de cálculos iterativos, e esse processo iterativo seguirá até que sejam atendidos determinados critérios de convergência. O *K-medoids* é uma melhoria do *K-means*, projetado para lidar com dados discretos, no qual o ponto de dados mais próximo do centro dos pontos de dados é selecionado como representante do *cluster* correspondente.

3.4.1.2 Algoritmo de Agrupamento Baseado na Hierarquia

O princípio destes tipos de algoritmos de clusterização é construir as relações hierárquicas entre os dados para fins de agrupamento. Suponha-se que, inicialmente, cada ponto de dados represente um *cluster* individual e, em seguida, os dois *clusters*

Tabela 2 – Tipos de algoritmos de clusterização

Categoria	Algoritmo Típico
Algoritmo de Agrupamento Baseado na Partição	K-means, K-medoids, PAM, CLARA, CLARANS
Algoritmo de Agrupamento Baseado na Hierarquia	BIRCH, CURE, ROCK, Chameleon
Algoritmo de Agrupamento Baseado na Teoria Fuzzy	FCM, FCS, MM
Algoritmo de Agrupamento Baseado na Distribuição	DBCLASD, GMM
Algoritmo de Agrupamento Baseado na Densidade	DBSCAN, OPTICS, Mean-shift
Algoritmo de Agrupamento Baseado na Teoria dos Grafos	CLICK, MST
Algoritmo de Agrupamento Baseado na Grade	STING, CLIQUE
Algoritmo de Agrupamento Baseado na Teoria Fractal	FC
Algoritmo de Agrupamento Baseado em Modelo	COBWEB, GMM, SOM, ART

Fonte: Adaptado de (XU, D.; TIAN, 2015)

mais próximos são fundidos em um novo *cluster* até que reste apenas um. Ou seja, é um processo inverso. Algoritmos típicos deste tipo de clusterização incluem **BIRCH** (ZHANG; RAMAKRISHNAN; LIVNY, 1996), **CURE** (GUHA; RASTOGI; SHIM, 1998),

ROCK (GUHA; RASTOGI; SHIM, 2000), e **Chameleon** (KARYPIS; HAN; KUMAR, 1999). O **BIRCH** obtém o resultado de clusterização construindo a árvore de características da clusterização, chamada *CF tree*, na qual um nó representa um *subcluster*. A árvore CF crescerá dinamicamente quando um novo ponto de dados é adicionado. O **CURE**, adequado para clusterização em larga escala, utiliza técnicas de amostragem aleatória para agrupar amostras separadamente e, posteriormente, integra os resultados. O **ROCK** é uma melhoria do **CURE** para lidar com dados de tipo de enumeração, levando em consideração o efeito da similaridade dos dados ao redor do *cluster*. O **Chameleon**, a princípio, divide os dados originais em *clusters* de tamanho menor com base no grafo de vizinhos mais próximos e, em seguida, funde os *clusters*

menores em um *cluster* de tamanho maior, com base em um algoritmo aglomerativo, até que a condição seja atendida.

3.4.1.3 Algoritmo de Agrupamento Baseado na Teoria *Fuzzy*

Esse tipo de algoritmos de clusterização se baseia em transformar o valor discreto do rótulo de pertencimento, 0, 1, em um intervalo contínuo [0, 1], a fim de descrever a relação de pertencimento entre objetos de maneira mais razoável. Algoritmos típicos deste tipo de clusterização incluem **FCM** (BEZDEK; EHRLICH; FULL, 1984), **FCS** (DAVE; BHASWAN, 1992) e **MM** (YAGER; FILEV, 1994). A ideia central do **FCM** é obter a pertinência de cada ponto de dados a todos os *clusters* otimizando a função-objeto. Diferentemente dos algoritmos tradicionais de clusterização *fuzzy*, o **FCS** considera a hiperesfera multidimensional como o protótipo de cada *cluster*, permitindo a clusterização com base na função de distância da hiperesfera. O **MM**, baseado na Função de Montanha, é utilizado para encontrar o centro do *cluster*.

3.4.1.4 Algoritmo de Agrupamento Baseado na Distribuição

A ideia fundamental é que os dados, gerados a partir da mesma distribuição, pertencem ao mesmo cluster se existirem várias distribuições nos dados originais. Os algoritmos típicos são o **DBCLASD** (XU, X. *et al.*, 1998) e o **GMM** (RASMUSSEN, 1999). A ideia central do **DBCLASD**, um algoritmo incremental dinâmico, é que se a distância entre um cluster e seu ponto de dados mais próximo satisfaz a distribuição da distância esperada gerada a partir dos pontos de dados existentes desse cluster, o ponto de dados mais próximo deve pertencer a esse cluster. A ideia central do **GMM** é que ele é composto por várias distribuições gaussianas das quais os dados originais são gerados, e os dados que obedecem à mesma distribuição gaussiana independente são considerados pertencentes ao mesmo cluster.

3.4.1.5 Algoritmo de Agrupamento Baseado na Teoria dos Grafos

De acordo com esse tipo de algoritmos de *clustering*, a clusterização é realizada no grafo, onde o nó é considerado como o ponto de dados e a aresta é considerada como a relação entre os pontos de dados. Algoritmos típicos desse tipo de clusterização incluem o **CLICK** (SHARAN; SHAMIR, 2000) e a clusterização baseada na Árvore de Abrangência Mínima (MST) (JAIN; MURTY; FLYNN, 1999). A ideia central do **CLICK** é realizar a divisão de peso mínimo do grafo com iterações para gerar os *clusters*. A geração da árvore de abrangência mínima a partir do grafo de dados é o passo chave para a análise de clusters no algoritmo de clusterização baseado em MST.

3.4.1.6 Algoritmo de Agrupamento Baseado em grade

A ideia fundamental deste tipo de algoritmo de clusterização é transformar o espaço de dados original em uma estrutura de grade com um tamanho definido para a clusterização. Os algoritmos típicos deste tipo de clusterização incluem o **STING** (WANG, W.; YANG; MUNTZ *et al.*, 1997) e o **CLIQUE** (AGRAWAL *et al.*, 1998). A ideia central do **STING**, que pode ser utilizado para processamento paralelo, é dividir o espaço de dados em muitas unidades retangulares construindo uma estrutura hierárquica, e os dados em diferentes níveis de estrutura são agrupados respectivamente. O **CLIQUE** aproveita os algoritmos de clusterização baseados em grade e os algoritmos de clusterização baseados em densidade.

3.4.1.7 Algoritmo de Agrupamento Baseado na Teoria Fractal

O termo *fractal* se refere à geometria que pode ser dividida em várias partes que compartilham algumas características com o todo. O algoritmo típico deste tipo de clusterização é o **FC** (BARBARÁ; CHEN, 2000), cuja ideia central é que a alteração de qualquer dado interno de um cluster não tem influência na qualidade intrínseca da dimensão fractal.

Neste tipo de abordagem de clusterização, o foco recai na estrutura fractal dos dados, onde cada parte se assemelha ao todo em termos de propriedades fractais. O algoritmo **FC** é projetado para identificar e agrupar dados com base na preservação da dimensão fractal, independentemente de pequenas alterações nos dados individuais.

3.4.1.8 Algoritmo de Agrupamento Baseado em Modelo

A ideia fundamental é selecionar um modelo específico para cada cluster e encontrar o melhor ajuste para esse modelo. Existem principalmente dois tipos de algoritmos de clusterização baseados em modelos, um baseado em método de aprendizado estatístico e o outro baseado em método de aprendizado de redes neurais.

Os algoritmos típicos, baseados em método de aprendizado estatístico, são **COBWEB** (FISHER, 1987) e **GMM** (RASMUSSEN, 1999). A ideia central do **COBWEB** é construir uma árvore de classificação, com base em critérios heurísticos, a fim de realizar clusterização hierárquica com a suposição de que a distribuição de probabilidade de cada atributo é independente. Os algoritmos típicos, baseados em método de aprendizado de redes neurais, são SOM [65] e ART [66–69]. A ideia central do SOM é criar um mapeamento de redução de dimensão do espaço de entrada de alta dimensão para o espaço de saída de baixa dimensão com a suposição de que existe topologia nos dados de entrada. A ideia central do ART, um algoritmo incremental, é gerar dinamicamente um novo neurônio para corresponder a um novo padrão e criar

um novo cluster quando os neurônios atuais não são suficientes. GMM foi discutido na seção 3.4.1.4.

3.5 MÉTRICAS DE AVALIAÇÃO DE DESEMPENHO DO MODELO

Um dos aspectos mais importantes na análise de clusters é a avaliação dos resultados de agrupamento para encontrar a partição que melhor se ajusta aos dados subjacentes (HALKIDI; BATISTAKIS; VAZIRGIANNIS, 2001).

Avaliar o desempenho de um algoritmo de clusterização é uma tarefa complexa e vai além da simples contagem de erros ou da aplicação das métricas de precisão comumente usadas em algoritmos de classificação supervisionada. Em particular, a avaliação de algoritmos de clusterização requer métricas que não considerem os rótulos dos clusters de forma absoluta, mas sim a capacidade desses clusters em estabelecer separações nos dados de maneira análoga a um conjunto de classes de referência ou que satisfaçam pressupostos nos quais os membros pertencentes à mesma classe apresentem maior similaridade do que aqueles pertencentes a classes distintas, de acordo com alguma métrica de similaridade estabelecida (SCIKIT LEARN ORG, 2023).

Para avaliar as soluções de clusterização, geralmente são utilizados índices de validade para medir a qualidade dos resultados. Existem dois tipos de índices de validade: índices externos e índices internos. Um índice externo é uma medida de concordância entre duas partições, em que a primeira partição representa a estrutura de clusterização conhecida a priori, e a segunda resulta do procedimento de clusterização. Os índices internos são usados para medir a qualidade de uma estrutura de clusterização sem informações externas. Para os índices externos, avaliamos os resultados de um algoritmo de clusterização com base em uma estrutura de clusters conhecida de um conjunto de dados (ou rótulos de clusters). Para os índices internos, avaliamos os resultados usando quantidades e características inerentes ao conjunto de dados (WANG, K.; WANG, B.; PENG, 2009).

No contexto do Bistek Supermercados, é evidente a presença de um sólido conhecimento de negócio referente aos potenciais agrupamentos de clientes clube, uma população que atualmente abrange aproximadamente 800.000 indivíduos. A equipe de Gestão de Relacionamento com o Cliente, emprega uma abordagem de segmentação fundamentada no modelo RFV, resultando na distinção de três segmentos distintos.

Dentre esses segmentos, destaca-se o denominado *Cestinha*, que engloba clientes caracterizados por uma alta frequência de compras, embora com um valor médio de gastos por transação inferior em relação ao outro grupo identificado como *Rancho*. Por sua vez, o grupo *Rancho* abarca clientes com uma frequência de compras inferior, porém apresentando valores médios de gastos mais substanciais.

É importante mencionar que existe ainda um terceiro grupo de clientes, que, por

falta de uma denominação específica, pode ser referido como *Outros*. Essa categoria compreende os clientes que não se enquadram nas descrições dos dois primeiros segmentos identificados. A compreensão e caracterização desses clusters são fundamentais para a formulação de estratégias de marketing eficazes e aprimoramento do relacionamento com os clientes no Bistek Supermercados.

4 ANÁLISE DE REQUISITOS E MODELAGEM DO SISTEMA PROPOSTO

4.1 METODOLOGIA

A metodologia proposta para o desenvolvimento do Projeto de Final de Curso (PFC) de inteligência artificial para o suporte à tomada de decisão de campanhas de marketing no Bistek supermercados é estruturada em etapas para garantir uma abordagem sistemática e eficiente. As etapas são as seguintes:

- **Levantamento de Requisitos:** Nessa fase, será realizada reuniões com representantes do Bistek supermercados para entender as necessidades e objetivos específicos da rede em relação ao projeto. Serão coletadas informações sobre as estratégias de marketing existentes, os dados disponíveis sobre o comportamento dos clientes, as métricas de desempenho atuais e os desafios enfrentados pela rede no cenário competitivo.
- **Análise de Dados:** Será trabalhado com os dados fornecidos pelo Bistek e, se necessário, realizar a integração de diferentes fontes de informações relevantes para a análise. Serão utilizadas técnicas de análise de dados, estatística e aprendizado de máquina para compreender os padrões de compra dos clientes, segmentar o público-alvo e identificar *insights* valiosos para a personalização das campanhas de marketing.
- **Desenvolvimento do algoritmo:** Com base nas análises de dados, será desenvolvido modelos de inteligência artificial que suportem a tomada de decisão nas campanhas de marketing. Esses modelos poderão incluir algoritmos de aprendizado de máquina, técnicas de *data mining* e outras abordagens de IA relevantes para a personalização das ofertas e adaptação das campanhas.
- **Implementação e Integração:** Os modelos de inteligência artificial desenvolvidos serão implementados em um sistema que será integrado aos processos de tomada de decisão de marketing do Bistek supermercados. A integração pode envolver o uso de APIs para acessar dados em tempo real, a conexão com sistemas existentes de gerenciamento de campanhas e a comunicação com outros sistemas relacionados.
- **Validação e Testes:** Nesta etapa, os modelos de inteligência artificial serão validados e testados com dados reais do Bistek supermercados para garantir que as previsões e recomendações sejam precisas e confiáveis. Serão realizados testes de desempenho, análises de resultados e ajustes necessários para otimizar o funcionamento do sistema.

- **Implementação Piloto:** Antes da implantação completa do sistema, será realizada uma implementação piloto em uma escala menor para avaliar a eficácia do sistema em um ambiente real. Esse piloto permitirá ajustes finos e a coleta de *feedback* dos usuários.
- **Monitoramento e Melhorias Contínuas:** Após a implantação, o sistema será monitorado continuamente para garantir sua eficácia e desempenho adequado.

Para acompanhamento e gerenciamento do projeto, foi utilizada a metodologia *O Cross Industry Standard Process for Data Mining* (CRISP DM), é uma metodologia consolidada para transformar dados empresariais em *insights* gerenciais. Desenvolvida há mais de 20 anos para lidar com o desafio do Big Data, essa abordagem foca em projetos que lidam com grandes volumes de informação.

Essa metodologia atua na mineração de dados, parte integrante da ciência de dados, utilizando estatísticas e matemática para analisar dados, identificar padrões e resolver questões empresariais. O ciclo do *CRISP DM* se divide em seis etapas:

- Entendimento do problema: Compreensão do impacto do problema na empresa e definição de objetivos.
- Compreensão dos dados: Organização e documentação dos dados relevantes para a resolução do problema.
- Preparação dos dados: Aplicação técnica na análise dos dados, escolhendo os formatos e questões técnicas.
- Modelagem: Aplicação das técnicas de Data Mining conforme os objetivos estabelecidos.
- Avaliação: Verificação dos resultados em relação aos objetivos e aplicação dos conhecimentos adquiridos.
- Implementação dos modelos na empresa: Aplicação dos insights obtidos para mudar processos e criar soluções baseadas em dados.

As primeiras três etapas concentram-se na compreensão do negócio e dos dados, além da preparação dos dados. O entendimento do objetivo do projeto e a necessidade de alinhamento entre os envolvidos são essenciais nesse estágio. Em seguida, a identificação dos dados relevantes é crucial, seguida pela organização meticulosa dos dados coletados.

As etapas finais se concentram na construção do modelo e sua implementação. A modelagem envolve a definição de quais dados serão usados e a seleção do modelo

mais adequado. A avaliação verifica se o modelo atende às expectativas e permite ajustes, se necessário. Por fim, a implementação coloca o modelo em produção, trazendo valor para o negócio.

Os benefícios incluem melhorias no relacionamento com clientes, orientação para decisões mais precisas, novos modelos de resolução de problemas e análises em tempo real. Implementar o CRISP DM envolve mapear problemas estratégicos, definir metas claras e contar com colaboradores com habilidades analíticas.

Essa metodologia flexível e validadora ajuda a lidar com a incerteza nos negócios, permitindo adaptações ao longo do tempo. Ao aplicar o CRISP DM, as empresas ganham vantagens competitivas, tornando possível aproveitar dados para impulsionar estratégias eficientes e decisões embasadas.

4.2 REQUISITOS

A fim de estabelecer os requisitos necessários para o desenvolvimento deste projeto, foi realizada uma cuidadosa análise conjunta com a equipe de gestão da área de negócios do Bistek Supermercados. O objetivo principal foi definido em conformidade com os interesses estratégicos da empresa, com um enfoque particular em três grupos distintos de clientes. É de notável relevância a identificação e caracterização desses três clusters de clientes dentro da base de clientes do Clube Bistek.

Para a condução dessa análise, foi concedido acesso aos dados da base de clientes do supermercado. Contudo, com o propósito de demonstrar o conceito deste estudo, limitou-se a utilizar os registros relativos às lojas situadas na ilha de Florianópolis, abrangendo um conjunto de três pontos de venda. Além disso, o período de análise abrangeu os meses de setembro a dezembro de 2022.

No âmbito da segmentação dos clientes, com ênfase na identificação de três grupos principais, em decorrência do conhecimento prévio da área de negócios a respeito do comportamento dos clientes, procedeu-se à validação das estratégias de segmentação, culminando na caracterização dos perfis destes clusters, conforme descrito a seguir:

- Clientes do Cluster **Cestinha**: Este segmento é constituído por uma audiência que frequenta o estabelecimento em questão com alta regularidade, realizando, em média, mais de uma compra por semana, embora com desembolsos médios de menor monta. Pode-se caracterizar este cluster como composto por clientes que recorrem ao estabelecimento para efetuar compras de natureza ágil, já munidos de um destino de compra preestabelecido.
- Clientes do Cluster **Rancho**: O cluster dos "Rancho" é composto por clientes que efetuam compras com menor regularidade, geralmente a cada uma ou duas semanas, mas que apresentam valores médios de gastos mais elevados. Este

grupo é caracterizado por clientes que frequentam o estabelecimento de forma menos frequente, porém não se restringem a itens específicos, podendo adquirir produtos além daqueles planejados.

- Terceiro Cluster, ou Cluster **Outros**: O terceiro cluster abarca clientes que não se encaixam nas categorias previamente mencionadas. Geralmente, trata-se de clientes que recorrem ao supermercado durante promoções especiais ou que podem ter deixado de frequentar o Bistek em favor de um estabelecimento concorrente mais próximo.

Conforme as Tabelas 3 e 4 pode-se observar os requisitos funcionais e não funcionais do projeto, bem como suas descrições.

Tabela 3 – Requisitos funcionais do projeto

Requisito	Descrição	Categoria	Obrigatório	Permanente
Segmentação em três clusters	O sistema deve conseguir segmentar os clientes em três principais grupos diferentes, cada um com suas características específicas	Algoritmo	(X)	(X)
Disponibilizar dados segmentados	Como o CRM trabalha com ações ativas para os clientes, ele necessita ter acesso aos dados segmentados para poder ativar tais promoções para os clientes	Sistema	(X)	(X)
Rápida visualização dos dados	Devido ao processo antigo ser demorado, o novo processo deve ser de rápido acesso para o time de CRM	Sistema	(X)	(X)

Tabela 4 – Requisitos não funcionais do projeto

Requisito	Descrição	Categoria	Obrigatório	Permanente
Identificar outros clusters, como clientes do segmento de vinho, açougue, importados e outros	Além dos principais clusters, pode-se sub segmentar esses segmentos para identificar grupos menores com características parecidas, como ticket médio por setor	Algoritmo	()	(X)
Abranger todas as lojas da rede	Aumentar o escopo de análise para outras lojas da rede	Algoritmo	()	(X)
Segmentar resultados por lojas	Poder selecionar quais são as lojas de interesse na análise presente do CRM	Sistema	()	(X)
Disponibilidade para acesso remoto	O sistema deve poder ser acessa de outros locais além da central do TI	Infraestrutura	(X)	(X)
Controle de acesso	Deve-se implementar um controle de acesso para bloquear usuários não identificados	Infraestrutura	(X)	(X)

A segmentação desses grupos de clientes desempenha um papel de suma importância na compreensão das particularidades de cada cluster e na adaptação das estratégias de marketing. Este enfoque permite direcionar as campanhas de forma mais precisa e personalizar as ofertas de acordo com as características e comportamentos de cada cluster, contribuindo, assim, para a eficácia das ações de marketing e a melhoria da satisfação dos clientes.

4.3 ESCOLHA DAS ABORDAGENS A SEREM IMPLEMENTADAS

A escolha da abordagem foi fundamentada na combinação do método RFV com a aplicação de algoritmos de segmentação de clientes. Como evidenciado na Tabela 2, há uma gama variada de algoritmos disponíveis para a segmentação não supervisionada. No entanto, para o escopo deste estudo, optou-se pela utilização de

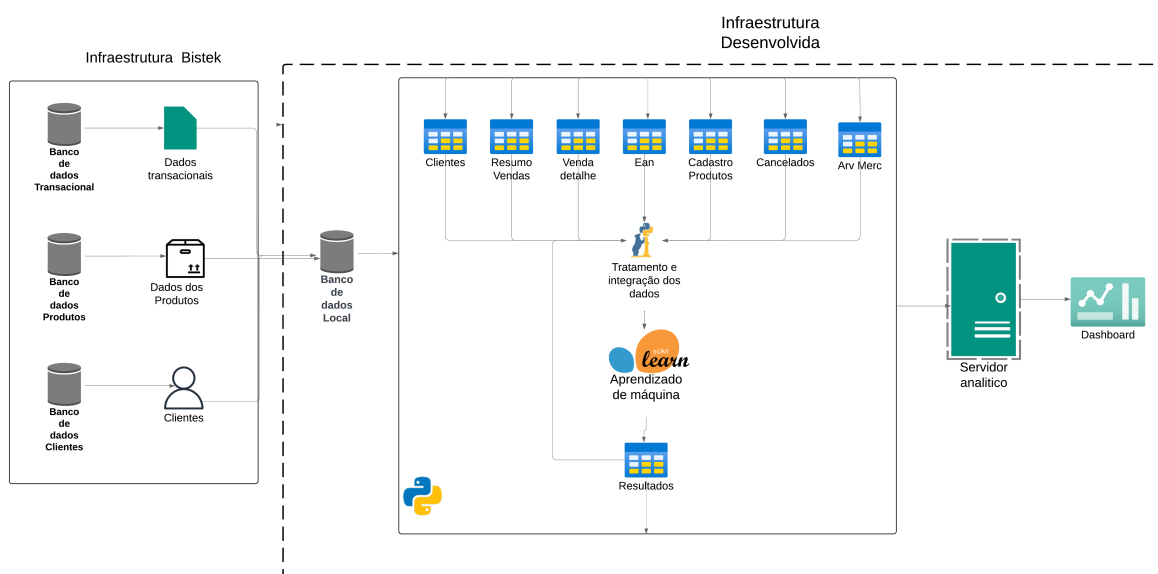
algoritmos de agrupamento baseados na distribuição, com destaque para o modelo de Mistura Gaussiana (*Gaussian Mixture Model*).

Esta escolha se justifica pelo potencial do *Gaussian Mixture Model* em lidar com a segmentação de clientes, uma vez que é capaz de modelar distribuições complexas de dados e identificar agrupamentos subjacentes. Além disso, a aplicação de métodos de agrupamento baseados na distribuição permite uma abordagem mais flexível para a identificação de clusters, que pode se ajustar aos padrões de comportamento dos clientes de forma mais precisa.

4.4 ARQUITETURA DO SISTEMA

Para cumprir os requisitos do projeto, foi levantado uma arquitetura conforme Figura 4.

Figura 4 – Arquitetura do Sistema



Fonte: Autor

1. Coleta de Dados de Segmentação de Clientes Clube

Os dados necessários para realizar a segmentação dos clientes clube são inicialmente obtidos a partir de fontes de dados transacionais e analíticas. Portanto, é necessário que esses dados sejam coletados dessas fontes e disponibilizados localmente para permitir um tratamento adequado e uma modelagem precisa. Normalmente, essa coleta de dados é executada por meio de um software que estabelece conexão com as fontes de dados e transfere as informações para um banco de dados local ou para arquivos.

2. Preparação e Modelagem dos Dados de Segmentação

Os dados extraídos inicialmente se encontram em estado bruto, carentes de regras de negócio e otimização para servir como entrada para algoritmos de segmentação de clientes clube. Nessa etapa, uma série de processos é executada com o propósito de gerar tabelas que possam ser facilmente utilizadas pelos algoritmos de segmentação. Vale destacar que a modelagem resultante dos dados frequentemente não segue uma forma normalizada, mas é configurada como uma única tabela com todas as características necessárias, podendo variar dependendo do algoritmo utilizado. Uma parte importante desse processo é o agrupamento de clientes, que tem como principal objetivo limitar a quantidade de dados que os algoritmos precisam processar, tornando o processo mais eficiente. Além disso, o agrupamento de clientes facilita a posterior aplicação de regras de negócio, como por exemplo os cálculos dos valores de RFV, que são métricas importantes da segmentação.

3. Algoritmos de Segmentação de Clientes Clube

Este estágio representa o núcleo do sistema de segmentação de clientes clube. Os resultados desse algoritmo consiste, efetivamente, em segmentações de grupos de clientes. A validação dessa segmentação partirá de regras de negócios e conhecimentos apriori da formação desses clusters.

4. Processo de Segmentação

Dado que múltiplas segmentações paralelas foram geradas, é necessário caracterizar essas segmentações. Esse processo pode ser feito analisando os dados já segmentados, através de uma análise exploratória pode-se averiguar quais são os valores médios de RFV desse cluster, dados como idade, sexo, etc.

5. Segmentação Final de Clientes Clube

Após a conclusão das segmentações, que foram desenvolvidas até esta fase e se baseiam em grupos de clientes, é imperativo tornar esses dados segmentados acessíveis às diversas áreas da empresa. Nesse estágio, a segmentação já está totalmente elaborada, restando apenas a etapa de disponibilização do acesso às informações segmentadas para as demais áreas da organização.

Essa fase de disponibilização é crucial, pois permite que as conclusões e os *insights* obtidos por meio da segmentação sejam utilizados de forma eficaz por diferentes departamentos, como marketing, vendas, atendimento ao cliente e planejamento estratégico.

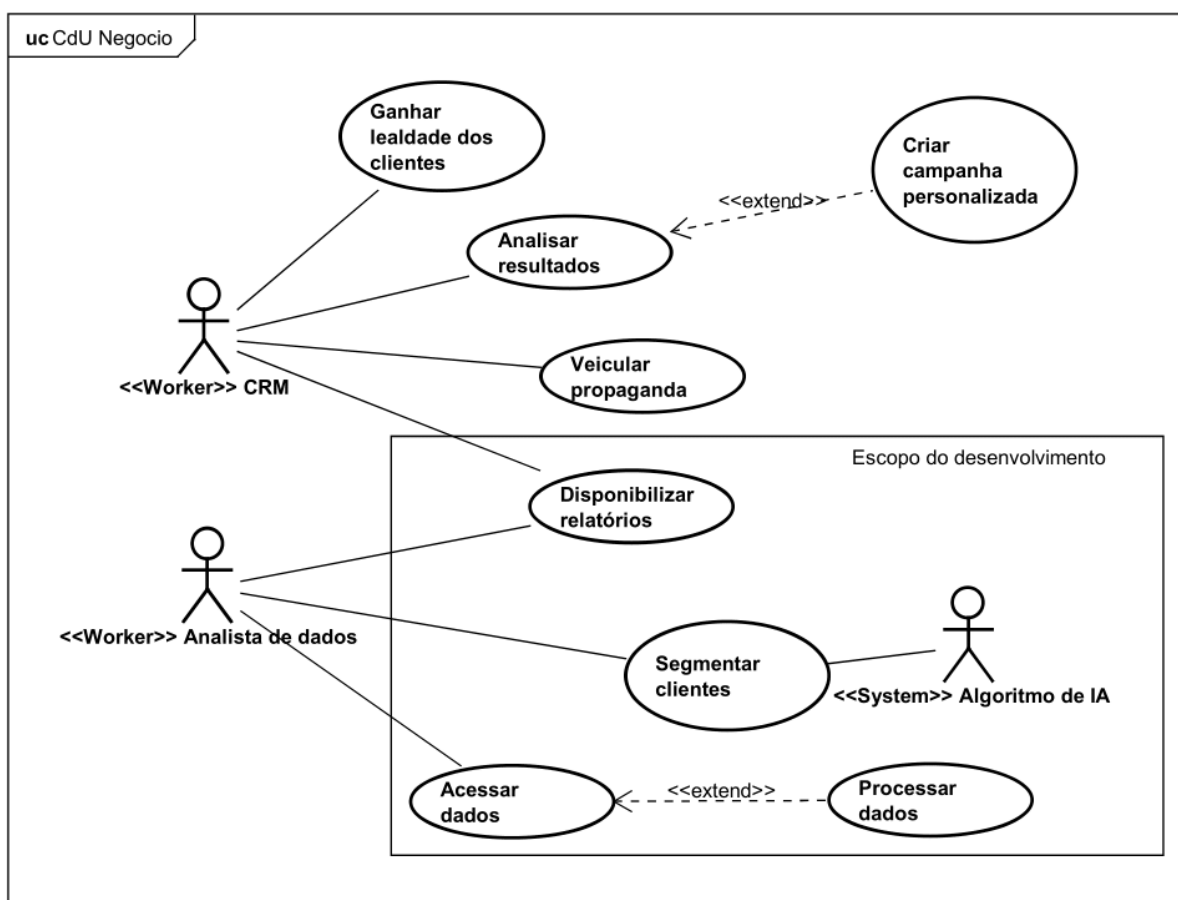
Conforme a Figura 5 pode-se verificar o diagrama de atores do sistema em que nota-se como o setor de CRM atua e como o escopo deste trabalho afeta esse

processo. Tem-se que o foco é trabalhar em cima dos dados disponíveis em conjunto com algoritmos de inteligência artificial e disponibilizar relatórios para que o setor de marketing consiga criar campanhas personalizadas e ganhar lealdade dos clientes.

Aprofundando sobre como esse processo vai ser elaborado, tem-se a Figura 6 em que mostra-se o fluxo dos dados no sistema, iniciando no banco de dados transacional da empresa e percorrendo o caminho até um servidor analítico onde o analista de marketing vai conseguir entrar para acessar os relatórios. Os dados começam sendo copiados no banco de dados transacional e partem para o servidor de homologação que atualmente serve como um ambiente de teste de novas atualizações e novos programas que vão se conectar ao banco, após passar por esse servidor os dados são exportados para a máquina local onde são processados e segmentados e servidos no servidor analítico para o CRM.

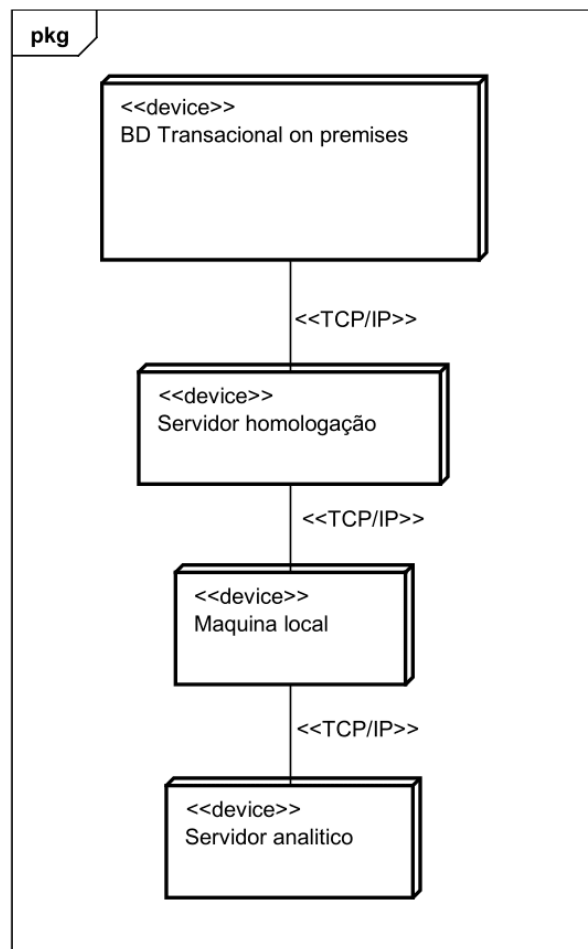
Adicionalmente, a arquitetura que abriga esses dados pode ser visualizada na Figura 7. Essa representação evidencia a relação entre as tabelas e a fonte de cada informação para o funcionamento do sistema.

Figura 5 – Diagrama de atores



Fonte: Autor

Figura 6 – Diagrama de implantação



Fonte: Autor

5 IMPLEMENTAÇÃO DO PROJETO E INFRAESTRUTURA TECNOLÓGICA

Este capítulo se concentra na implementação prática do projeto, onde os conceitos teóricos se transformam em ações concretas. Explora a infraestrutura tecnológica subjacente, destacando os aspectos técnicos da aquisição e preparação de dados, bem como a aplicação de algoritmos de inteligência artificial para segmentar os clientes.

Examinaremos as ferramentas, linguagens de programação e ambientes de desenvolvimento utilizados no projeto.

O cerne da arquitetura foi apresentado anteriormente, neste capítulo será demonstrado em mais detalhes como foi executada a arquitetura e como está o processo de extração, transformação e entrega dos dados.

5.1 BASE DE DADOS

Para a implementação e testes da arquitetura proposta, tem-se a disposição uma valiosa base de dados generosamente concedida pela rede de supermercados Bistek. É importante ressaltar que qualquer informação de natureza pessoal ou que possa revelar segredos sensíveis da empresa será tratada com o mais absoluto sigilo, conforme estritamente solicitado pela própria organização.

É de primordial importância a etapa de levantamento das variáveis para o desenvolvimento do modelo. Desta forma, buscou-se trabalhar com as variáveis que caracterizam os clientes pertencentes a cada cluster. Durante a etapa de coleta de dados, as amostras foram tomadas diretamente do banco de dados *on premises* da empresa. Este que contém todos os dados, dos diferentes sistemas do Bistek, unificados.

Para efeitos de validação da solução será trabalhado com dados de vendas pertencentes as três lojas da ilha de Florianópolis, durante o período de 01/09/2022 a 31/12/2022.

Ainda foram extraídos os dados de todos os clientes clube da rede Bistek. Fez-se necessário extrair todos pois os clientes não são segmentados regionalmente, assim um cliente do Rio Grande do Sul pode aparecer na análise se ele tiver realizado uma compra em uma das lojas analisadas no período.

Além disso, retirou-se também as tabelas de cadastro dos produtos e classificação mercadológica.

A base representa a seguinte quantidade de dados:

- 865.000 compras
- Aproximadamente 52.000 clientes clube identificados

Como o banco de dados que continha essas informações é o banco de dados transacional da empresa, foi realizada uma exportação dos dados para outro banco de dados local, para efeitos de prova de conceito foi utilizado o banco de dados *Oracle Express Edition*.

5.2 FERRAMENTAS UTILIZADAS

A efetiva implementação da solução demandou a utilização de duas ferramentas-chave: um banco de dados para a armazenagem dos dados e uma ferramenta que permitisse a manipulação e aplicação de algoritmos de aprendizado de máquina sobre esses dados.

No que tange ao armazenamento dos dados, o procedimento iniciou-se com a extração em lotes dos dados do banco de dados transacional, cuidadosamente planejada para evitar qualquer impacto no funcionamento do sistema durante os momentos de menor demanda. Posteriormente, os dados foram replicados, mantendo o mesmo esquema de relacionamentos e tabelas, em um banco de dados local.

No que diz respeito ao tratamento, análise e utilização dos modelos de aprendizado de máquina, a linguagem de programação Python emergiu como a ferramenta central. Ela foi combinada com bibliotecas especializadas, como Pandas, Numpy e Scikit-learn.

O Pandas é uma biblioteca Python amplamente utilizada para análise e manipulação de dados. Ele fornece estruturas de dados de alto desempenho e fáceis de usar, como DataFrames e Series, que permitem a organização e a manipulação de dados de maneira eficiente.

Enquanto que o NumPy é uma biblioteca fundamental para computação científica em Python. Ela oferece suporte para arrays multidimensionais (conhecidos como ndarrays) e funções matemáticas que possibilitam a execução de cálculos de maneira eficiente.

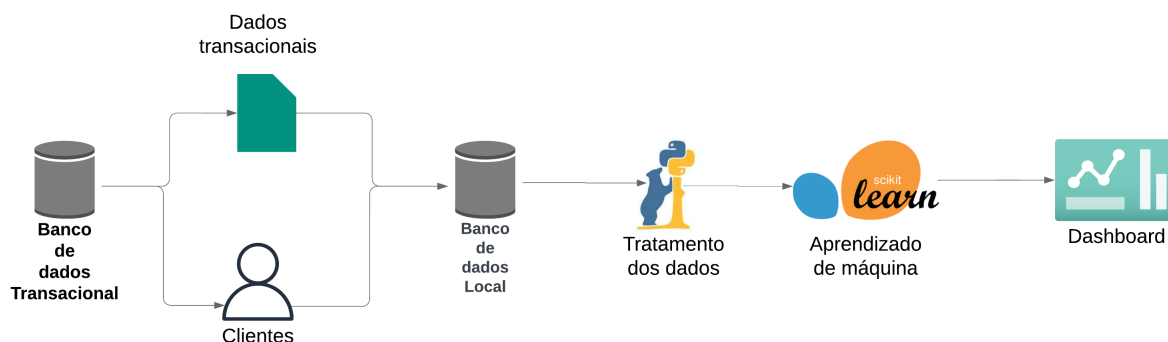
E por fim o Scikit-learn é uma das bibliotecas de aprendizado de máquina mais populares em Python. Ela oferece uma variedade de algoritmos para tarefas de aprendizado supervisionado e não supervisionado, bem como ferramentas para avaliação de modelos e pré-processamento de dados.

5.3 ARQUITETURA

De acordo com a arquitetura detalhada na seção 4.4, a implementação adotará a estrutura representada na Figura 8. Para efetuar o armazenamento dos dados, foi escolhido o banco de dados Oracle. A organização e relação das tabelas que compõem o banco de dados podem ser examinadas detalhadamente na Figura 7.

Cada etapa da arquitetura vai ser melhor detalhada a seguir.

Figura 8 – Arquitetura simplificada do Sistema



Fonte: Autor

5.3.1 Extração dos dados

Dentro da seção de extração e importação dos dados em outro banco, pode-se verificar a estrutura das tabelas na Figura 7. Neste processo não há alteração nos dados originais, apenas uma cópia fiel. Os dados foram extraídos e o esquema do banco foi replicado localmente, assim temos acesso irrestrito durante a etapa de desenvolvimento.

5.3.2 Seleção de Variáveis

Para o desenvolvimento do modelo de segmentação, foram selecionadas, inicialmente, possíveis variáveis que pudessem impactar positivamente para a predição, dentre o universo de dados que temos disponíveis.

Muitos dos dados presentes nas tabelas acabam sendo informações desnecessárias para a segmentação, como *flags* internas dos softwares, ou outras informações de integração.

Ao contrário das abordagens convencionais, este estudo adota uma metodologia inovadora que não se baseia na divisão dos clientes em quintis a partir de uma ordem predefinida e na atribuição de pontuações de 1 a 5 a cada dimensão RFV em cada quintil. Para uma caracterização mais precisa dos clusters, foram selecionadas variáveis que podem proporcionar uma representação mais abrangente dos indivíduos. Assim, optou-se por variáveis como sexo, recência, frequência e valor. Além de os de RFM outras variáveis vão ser adicionadas para complementar os resultados.

Essas variáveis possibilitam uma análise mais profunda do comportamento do cliente ao longo do tempo, permitindo-nos compreender não apenas a sua frequência e valor de compra, mas também a sua fidelidade e preferências.

5.3.3 Pré-processamento dos dados

Nesta seção, serão abordadas técnicas de pré-processamento de dados que desempenharam um papel crucial na preparação, organização e estruturação dos dados para otimizar o desempenho do algoritmo. A adequação dos dados ao formato requerido pela biblioteca GMM também é uma parte essencial desse processo.

5.3.3.1 Valores ausentes

A presença de valores ausentes representou um desafio significativo em relação a certos tipos de informações contidas na base de dados. Esse problema se tornou especialmente evidente ao considerar duas variáveis relevantes para o modelo: a idade dos clientes e a data de entrada no Clube Bistek. A identificação dessas lacunas levantou discussões internas com a equipe de suporte e manutenção da infraestrutura da empresa, resultando na identificação de possíveis causas.

Inicialmente, o Clube Bistek foi estabelecido há muito tempo, e, em seus primórdios, algumas informações não eram coletadas dos clientes durante o processo de cadastramento. Isso se deve, em parte, ao fato de que, na época, os clientes preenchiam formulários físicos nos quais essas informações eram frequentemente omitidas ou fornecidas de forma genérica. Posteriormente, à medida que as bases de dados foram integradas e o Clube evoluiu ao longo do tempo, esses dados jamais foram atualizados. Como resultado, parte da base de dados apresenta ausência de informações referentes a esses aspectos específicos.

Na análise realizada considerando os dados de vendas de 2022, cerca de 40% dos clientes analisados não possuem esses dados, assim para melhor aplicar o algoritmo foi estabelecido que essas variáveis seriam descartadas como candidatas para a segmentação.

5.3.3.2 Tratamento de Valores Ausentes

Uma etapa crucial na análise exploratória de dados é o pré-processamento, que engloba a limpeza e correção dos dados. Muitos problemas são recorrentes em diferentes conjuntos de dados, o que leva à aplicação de técnicas de análise exploratória semelhantes em vários cenários de desenvolvimento de modelos de Machine Learning. Neste estudo, identificamos a presença de valores ausentes em algumas variáveis dos dados brutos, e abordamos essa questão por meio de diversas técnicas.

Após a seleção das variáveis candidatas, procedemos ao tratamento dos valores nulos, concentrando nossos esforços nas variáveis que continham dados faltantes. Inicialmente, identificamos valores nulos nas seguintes variáveis:

- Idade

- Tempo de clube
- Sexo

Conforme abordado na seção 5.3.3.1, as variáveis Idade e Tempo de clube, acabaram com muitos dados faltantes, e mesmo abordando técnicas de predição nesses valores os resultados não foram satisfatórios. Por este motivo elas foram desconsideradas da análise.

Enquanto que a variável Sexo, além das categorias *Homem* e *Mulher*, contém as categorias *Não identificado* e *Nenhum*, que seria quando o cliente prefere não identificar. Para motivos de análise as categorias *Nenhum* e *Não identificado* foram unificadas em uma única categoria denominada *Indefinido*.

5.3.3.3 Criação de Novas Variáveis

A Engenharia de Recursos, também conhecida como *Feature Engineering*, é um processo fundamental que utiliza técnicas para criar novas variáveis a partir dos dados existentes. Esse procedimento visa aumentar o poder preditivo dos algoritmos de aprendizado de máquina. Ao aplicar a Engenharia de Recursos, é possível extrair informações valiosas das variáveis brutas ou não transformadas, melhorando assim a capacidade do modelo de fazer previsões precisas e significativas.

No conjunto inicial de variáveis, destacam-se os dados de RFV, juntamente com o gênero dos clientes. No entanto, para enriquecer a caracterização do comportamento dos clientes, optou-se por incorporar duas novas variáveis essenciais:

- **Total de Itens:** Essa variável representa a contagem cumulativa de itens registrados nos recibos de compra de cada cliente. A métrica considera tanto unidades individuais quanto peso em quilogramas, tratando-os como uma única dimensão. Isso significa que, por exemplo, se um cliente adquiriu 5 unidades de água e 0,574 kg de bananas em uma única compra, o total de itens para essa transação seria de 5,574.
- **Diversidade:** A segunda variável tem como objetivo discernir entre os clientes com maior diversidade em suas compras. Cada produto disponível em um supermercado é associado a uma categoria específica que o representa do ponto de vista mercadológico, assim foi estabelecido a quantidade de categorias distintas que o cliente comprou pelo menos um produto no período analisado. Essa variável busca capturar a variedade de categorias de produtos adquiridos por cada cliente, proporcionando *insights* valiosos sobre seus padrões de compra.

A inclusão dessas novas variáveis visa aprofundar a análise do comportamento dos clientes, fornecendo informações adicionais que podem ser fundamentais para a

segmentação eficaz e aprimoramento das estratégias de marketing. Ao considerar não apenas o RFV, mas também o total de itens e a diversidade das compras, é possível obter uma visão mais completa e detalhada do perfil de consumo de cada cliente, o que é fundamental para uma abordagem personalizada e direcionada.

5.3.4 Padronização dos Dados

A padronização é importante em algoritmos de aprendizado de máquina porque ajuda a normalizar as características ou variáveis no conjunto de dados. Isso garante que todas as variáveis estejam em uma escala semelhante, o que é crucial para o funcionamento eficaz de muitos algoritmos de aprendizado de máquina. E segundo (ELEN; AVUÇLU, 2021), essas são algumas razões pelas quais a padronização é importante:

1. **Iguala o intervalo:** Diferentes características em um conjunto de dados podem ter escalas e intervalos diferentes. A padronização transforma os dados de modo que cada característica tenha uma média zero e um desvio padrão de um. Isso iguala o intervalo das variáveis e evita que uma característica domine as outras.
2. **Auxilia na otimização baseada em gradientes:** Muitos algoritmos de aprendizado de máquina usam técnicas de otimização baseadas em gradientes para encontrar a solução ideal. A padronização ajuda esses algoritmos de otimização a convergirem mais rapidamente, fornecendo um espaço de entrada mais equilibrado e comportado.
3. **Melhora o desempenho do modelo:** A padronização pode melhorar o desempenho de modelos de aprendizado de máquina, especialmente aqueles sensíveis à escala das variáveis de entrada. Ela pode evitar que determinadas características tenham uma influência desproporcional nas previsões do modelo.
4. **Possibilita comparações significativas:** A padronização permite comparações significativas entre diferentes variáveis. Ela garante que as variáveis sejam medidas na mesma escala, facilitando a interpretação dos coeficientes ou pesos atribuídos a cada característica.

Em resumo, a padronização desempenha um papel crucial na preparação de dados para algoritmos de aprendizado de máquina, garantindo que as variáveis estejam em um formato consistente e comparável. Isso ajuda a eliminar viés e aprimorar a precisão e confiabilidade dos modelos.

Devido à natureza de algumas das variáveis selecionadas, que podem ser consideradas categóricas, como gênero, recência e frequência, mesmo que estas duas últimas sejam valores numéricos inteiros, podem ser divididas em categorias. Para lidar

com esses tipos de dados no algoritmo, aplicou-se a técnica conhecida como *one-hot encoding*.

Essa abordagem envolve a transformação de cada um dos possíveis valores de uma categoria em colunas de valores binários. Assim, se um cliente se encaixa em um valor específico dessa categoria, ele terá o valor 1 nessa coluna correspondente e 0 nas outras colunas da mesma categoria, indicando sua associação com esse valor em particular.

Essa técnica é fundamental para representar variáveis categóricas de uma maneira que o algoritmo de aprendizado de máquina possa interpretar de maneira adequada, garantindo que as categorias sejam tratadas de maneira justa e eficaz durante o processo de modelagem e análise. Dessa forma, a codificação one-hot permite que o algoritmo reconheça e utilize essas variáveis categóricas de maneira apropriada em seu funcionamento.

E para as variáveis contínuas foi aplicada a transformação de escores z . Trata-se de uma estratégia de normalização de dados que resolve o problema dos valores atípicos. Nessa técnica, os valores são normalizados com base na média e no desvio padrão dos dados.

A essência dessa técnica reside na transformação dos dados, convertendo os valores para uma escala comum em que o valor médio seja igual a zero e o desvio padrão seja igual a um. Tecnicamente, mede-se o número de desvios padrão abaixo ou acima da média. A padronização ou normalização com escore Z não é afetada por valores atípicos, uma vez que não há um intervalo predefinido para as características transformadas. Na equação 1 tem-se que μ é a média, σ é o desvio padrão e x o valor a ser transformado.

$$Z = \frac{x - \mu}{\sigma} \quad (1)$$

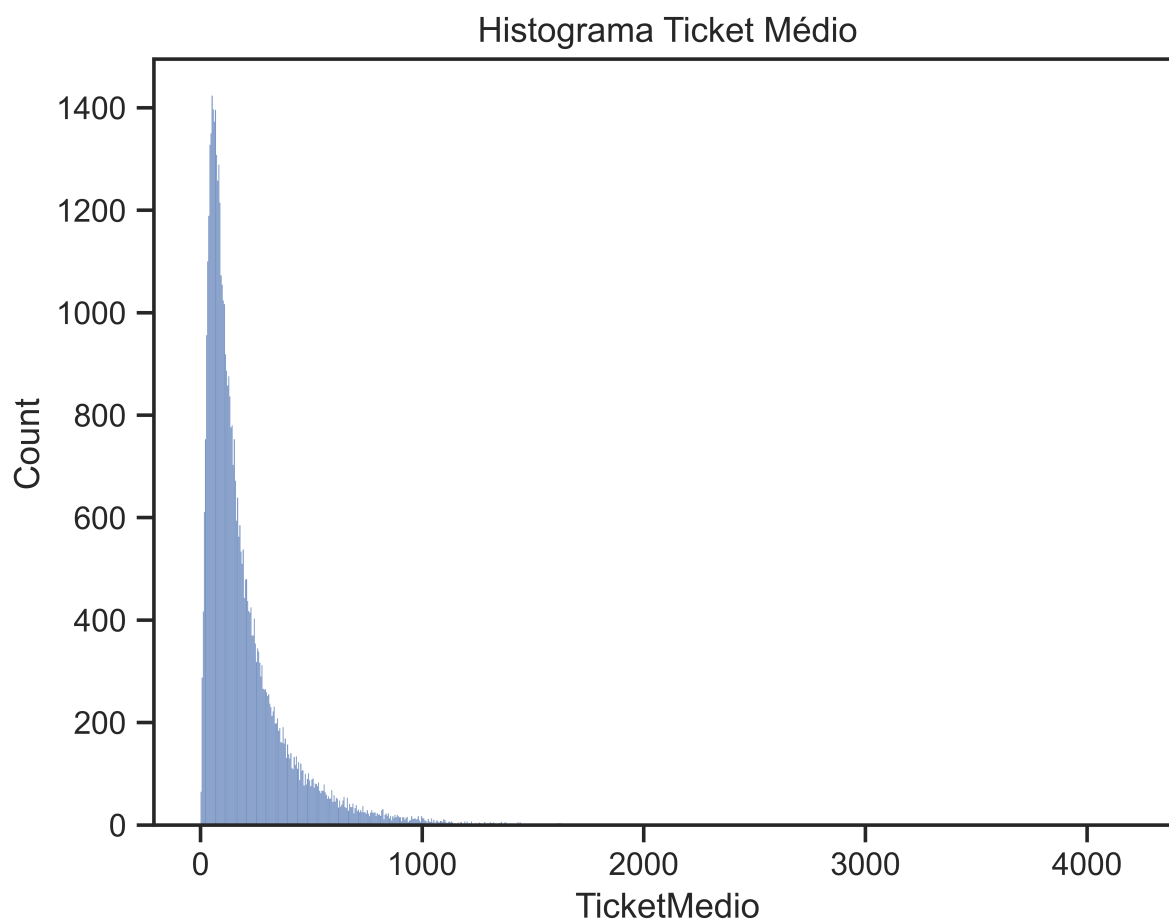
Para poder deixar os dados com uma distribuição normal e ainda reescalar os dados para que estejam centrados em zero, antes de se aplicar a equação 1, foi aplicada a função logarítmica nos dados, pois conforme Figura 9, nota-se que os dados não possuem uma distribuição normal antes do escore z .

5.4 ALGORITMO DE APRENDIZADO DE MÁQUINA

Para resolver este problema de segmentação foi empregado o algoritmo de mistura Gaussiana.

O Modelo de Mistura Gaussiana (Gaussian Mixture Model - GMM) é um algoritmo amplamente utilizado em Aprendizado de Máquina para modelar a distribuição de dados em termos de várias distribuições gaussianas (também conhecidas como distribuições normais). Ele é um modelo probabilístico que assume que os dados são

Figura 9 – Distribuição do Ticket Médio pré processamento dos dados



Fonte: Autor

gerados por uma combinação de várias distribuições gaussianas subjacentes.

A ideia principal por trás do GMM é que um conjunto de dados pode ser representado como uma mistura de várias distribuições gaussianas, cada uma representando um cluster ou componente do conjunto de dados. Cada componente gaussiano é caracterizado por seu próprio conjunto de parâmetros, incluindo média e variância. Portanto, um GMM pode ser usado para modelar dados complexos que não se ajustam bem a uma única distribuição gaussiana.

5.4.1 Experimentos

Nesta seção, irá ser abordado os experimentos realizados, onde cada um deles empregou uma maneira distinta de estabelecer as variáveis *Total de itens* e *Diversidade*, e como isso afetou os resultados. Além disso, apresentaremos um resumo dos métodos de transformação de dados utilizados para chegar ao modelo de segmentação. Também serão detalhadas as variáveis selecionadas.

O desenvolvimento dos experimentos teve início com a etapa de seleção das variáveis (conforme abordado na Subseção 5.3.2). Uma vez escolhidas as primeiras variáveis de interesse, procedeu-se ao tratamento de valores ausentes, conforme mencionado na Subseção 5.3.3.1. Em seguida, com o intuito de identificar variáveis adicionais com potencial relevância para as previsões, foram criados novos atributos a partir dos já existentes (conforme explicado na Subseção 5.3.3.3). Por fim, foram aplicadas transformações a variáveis que não estavam no formato necessário para o algoritmo GMM.

Em um primeiro momento, considerando a possível multicolinearidade entre as variáveis *Ticket Médio* e *Total de itens*, pensou-se em vez de utilizar a quantidade de itens comprados, empregar a quantidade de descontos presentes nas compras.

Atualmente no clube Bistek o cliente pode, através do seu aplicativo de *smartphone* realizar o login com o seu *Cadastro de Pessoa Física* (CPF) e ativar promoções especiais para o seu cadastro, estas que caso não sejam ativadas não garantem o desconto no momento da venda. Como essa informação está disponível para ser utilizada, em um primeiro momento foi realizado um experimento utilizando a quantidade de vezes que o cliente ativou uma dessas promoções.

E para a variável *Diversidade* foi calculada conforme a Subseção 5.3.3.3.

Com base nos dados disponíveis, e utilizando o algoritmo GMM, foi possível identificar 3 clusters distintos, conforme Figura 10.

O primeiro cluster, identificado como *Rancho*, se caracteriza por apresentar um *Ticket Médio* superior em relação aos outros clusters. Além disso, a recência média desse cluster, considerando a análise até o dia 31/12, é de aproximadamente 11 dias. Esta recência pode ser influenciada por compras substanciais durante o período natalino, uma vez que historicamente, a véspera de Natal é associada a um alto volume de vendas na rede. Adicionalmente, o cluster de *Rancho* demonstra uma frequência de compra ligeiramente mais baixa, corroborando com a caracterização mencionada na Seção 4.2.

Os clientes pertencentes a este cluster parecem não ser tão suscetíveis a campanhas de marketing no aplicativo do clube, como indicado pela taxa relativamente mais baixa de ativações de descontos no atributo *Total de descontos*. Ademais, apresentam uma variabilidade elevada no atributo *Diversidade*.

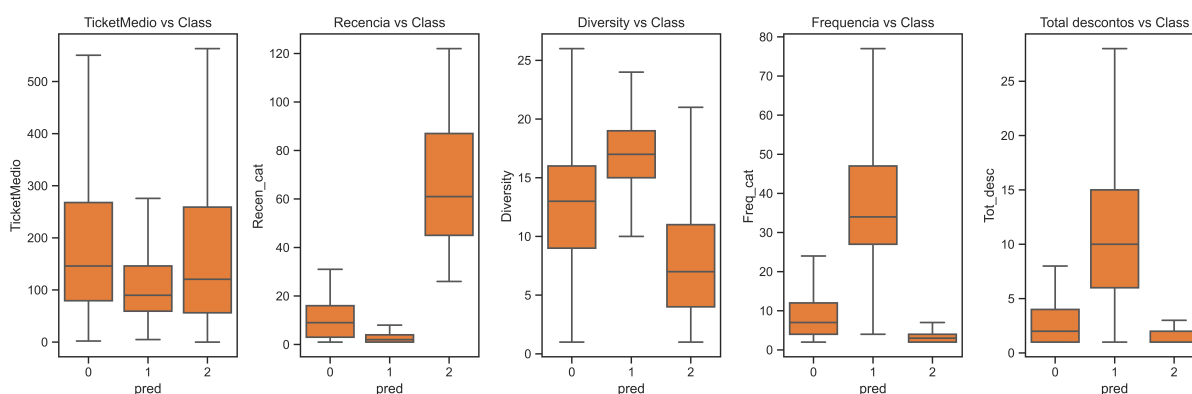
Quanto ao cluster denominado *Cestinha*, na Figura 10 pode ser identificado como o cluster 1, conforme descrito na Seção 4.2, seus membros exibem um *Ticket Médio* inferior, porém, ostentam uma alta frequência de compras. Em média, esses clientes realizaram 40 compras no período de análise de quatro meses, o que equivale a 10 compras mensais. Adicionalmente, apresentam uma recência extremamente baixa e são sensíveis a campanhas de marketing, conforme evidenciado pelo atributo *Total de Descontos*. Esses clientes também se destacam por adquirirem produtos de diversos

setores ao longo do tempo, exibindo uma diversidade média superior à do cluster *Rancho*.

Um ponto interessante de se notar nesse experimento é o cluster de *Outros*, ou cluster 2 na Figura 10. Nessa segmentação esse cluster apresentou clientes com uma variabilidade alta de *Ticket Médio*, contendo clientes com *Tickets Médios* altos, mas que visitam pouco o Bistek e fez muito tempo desde a última compra.

Esse cluster é um segmento dos clientes interessante para campanhas de recuperação de público, já que se tem a informação de que podem gastar muito no mercado, mas deixaram de visitar o Bistek.

Figura 10 – Histogramas dos atributos dos clusters



Fonte: Autor

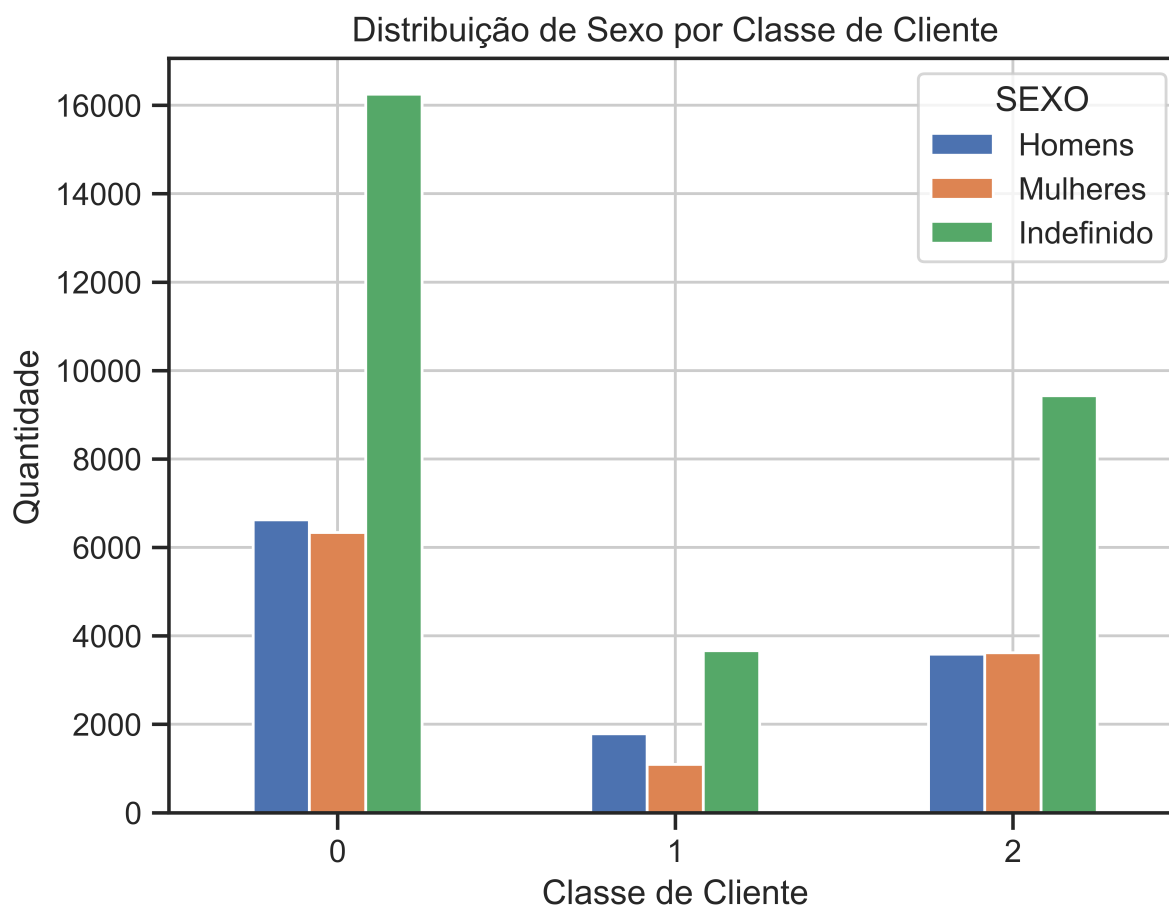
Além disso, ao analisar a Figura 11, é possível observar a distribuição de gênero por cluster.

Nesse contexto, o atributo de gênero não apresenta uma variação substancial na segmentação, com apenas uma ligeira tendência de contar com mais homens do que mulheres nos clusters "Rancho" e "Cestinha". Contudo, é notável a presença de um problema relacionado à ausência de dados, o que é evidenciado pela categoria "Indefinido". Este grupo inclui tanto clientes que optaram por não especificar seu gênero quanto aqueles para os quais os dados de gênero não estavam disponíveis.

E ainda, pela Figura 12, nota-se a quantidade de clientes identificados em cada segmento, em que mais da metade dos clientes foi segmentado como cluster 0, ou *Rancho*.

Para o segundo experimento, procedemos a uma modificação no método de cálculo da variável denominada *Diversidade*, que agora é definida como a média total do número de seções nas compras efetuadas por um cliente específico. Especificamente, consideramos um cliente fictício que realizou duas compras, uma em 07 de outubro de 2022 em uma das lojas de Florianópolis e outra em 11 de dezembro de 2022. Nessas compras, foram adquiridos 7 e 10 itens de diferentes seções, respectivamente. Com

Figura 11 – Distribuição do sexo por cluster



Fonte: Autor

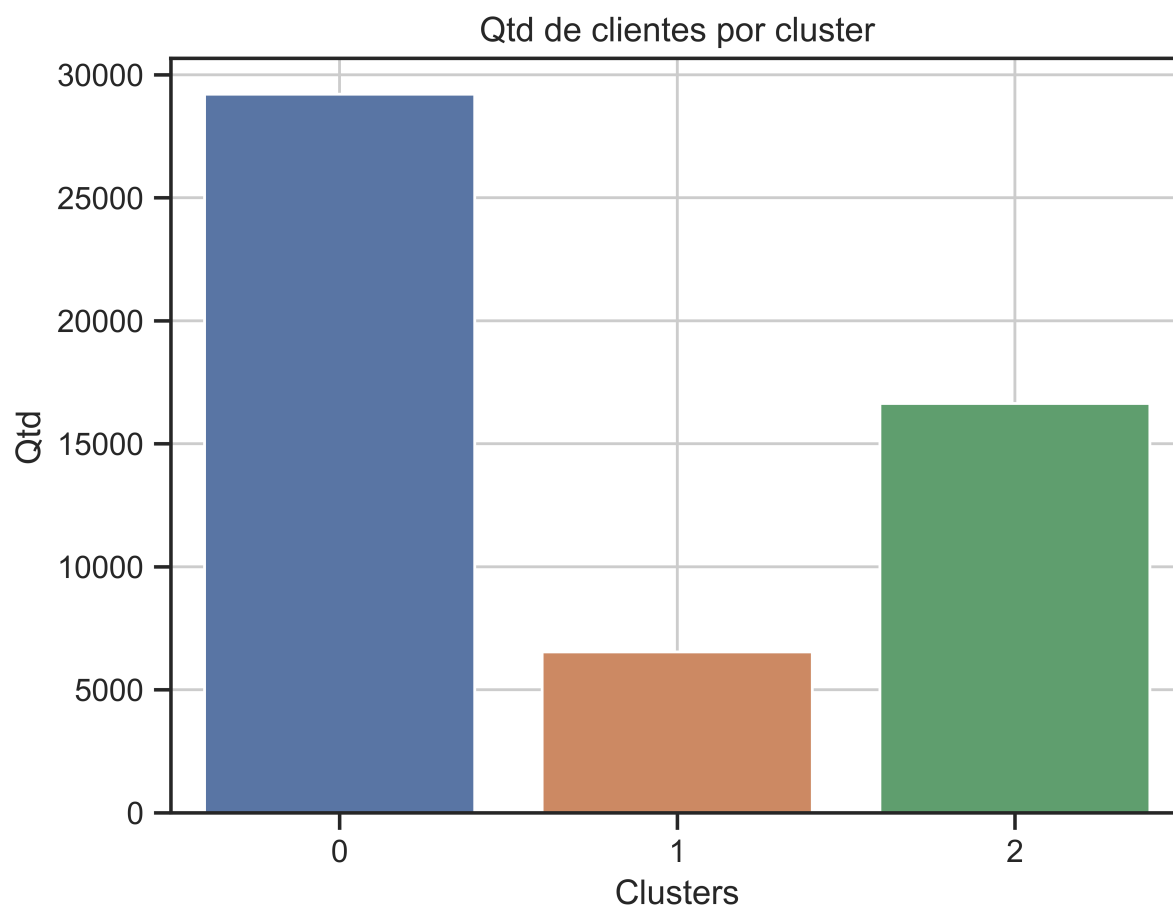
base nesses dados, o cálculo resulta em uma Diversidade de 8.5. Pode-se verificar a nota distribuição dessa variável na Figura 13.

Com base nos dados disponíveis anteriormente e essa alteração proposta, foi possível identificar 3 clusters distintos, conforme Figura 14.

Observa-se que, em comparação com a primeira segmentação, a segunda apresenta clusters mais definidos, notadamente no que diz respeito ao atributo Ticket Médio. O primeiro cluster, denominado Rancho, distingue-se por um "Ticket Médio" superior em relação aos demais clusters, tendo agora um valor médio significativamente superior ao anterior, passando de R\$200 para R\$409.25. No entanto, é importante notar o aumento da recência em todos os clusters, embora possa ser influenciado pelo período natalino. Além disso, o cluster "Rancho"(ou cluster 0) mantém suas características previamente observadas, mantendo uma frequência média de compra, uma recência ligeiramente mais alta, mas compensada por um alto Ticket Médio.

Outra alteração notável diz respeito à sensibilidade desse cluster em relação a campanhas de marketing. Isso é evidenciado pelo aumento no atributo "Quantidade

Figura 12 – Quantidade de clientes por cluster



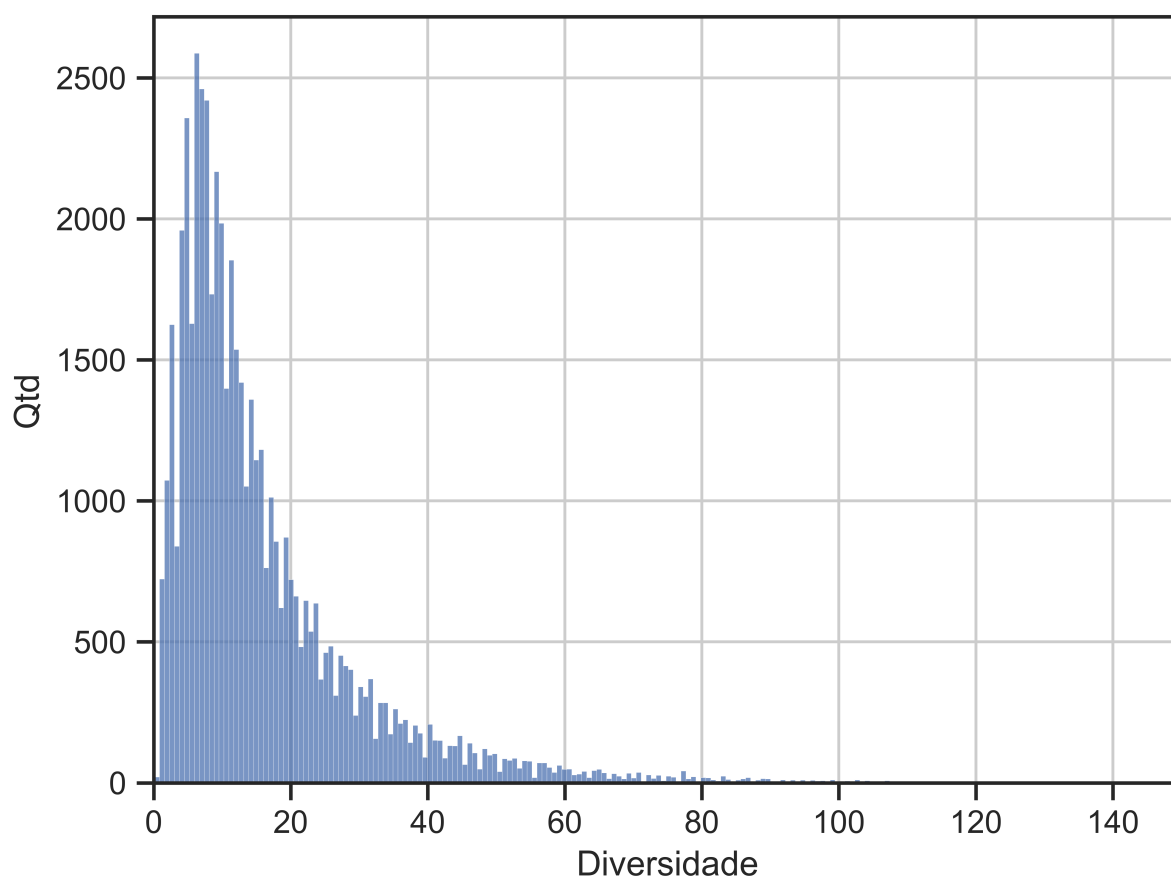
Fonte: Autor

de Descontos" nesse cluster, que reflete uma maior adesão às promoções oferecidas pelo clube.

Analisando os outros dois clusters, percebe-se uma mudança na relação do Ticket Médio. Anteriormente, o cluster Outros possuía um Ticket Médio mais elevado, mas agora essa tendência foi revertida a favor do cluster Cestinha, que apresenta um valor médio ligeiramente superior de R\$114.01 em comparação com R\$76.04, além de uma menor variabilidade de valores. O cluster Cestinha (ou cluster 1) continua a manter suas características de alta frequência, baixa recência e um alto nível de aproveitamento das promoções oferecidas aos membros do clube. Uma alteração notável é o atributo Diversidade, que agora apresenta uma média menor em comparação com o cluster Rancho.

Outro aspecto a ser destacado é o cluster denominado Outros (cluster 2), que atualmente é composto por clientes caracterizados por um baixo "Ticket Médio", alta recência e baixa frequência de compras. Esses clientes praticamente não utilizam o aplicativo do clube e realizam compras esporádicas. Isso se evidencia ao observar

Figura 13 – Distribuição da variável Diversidade



Fonte: Autor

o atributo de Diversidade, que apresenta uma média bastante reduzida, em torno de 6.504.

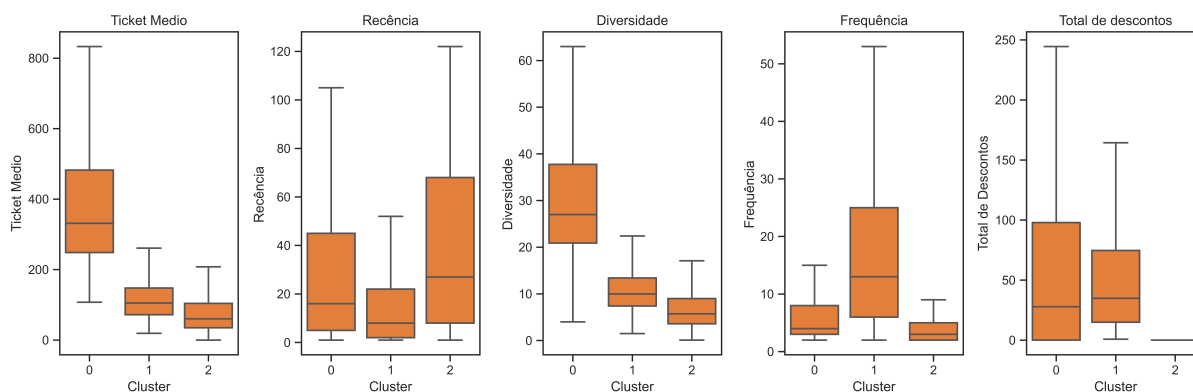
Além disso, outro ponto de destaque é a quase ausência de descontos nesse cluster, o que pode sugerir que esses clientes desconhecem esse benefício oferecido ou possam ter migrado para outros supermercados.

Repara-se também que agora teve uma leve mudança da distribuição de gênero pelos clusters, conforme Figura 15, em que mulheres passaram a ser uma leve maioria no cluster rancho, e agora há uma distribuição mais uniforme de clientes com gêneros identificados entre os clusters.

Outro fator que também apresentou um comportamento mais uniforme foi a quantidade de clientes por cluster, conforme Figura 16, em que os cluster ficaram melhor balanceados.

A Figura 14 evidencia, em termos de requisitos funcionais, a segmentação distintiva em três clusters. Cada atributo demonstra uma segmentação, especialmente notável quando se realiza uma análise RFV, na qual as características conhecidas a

Figura 14 – Histogramas dos atributos dos novos clusters



Fonte: Autor

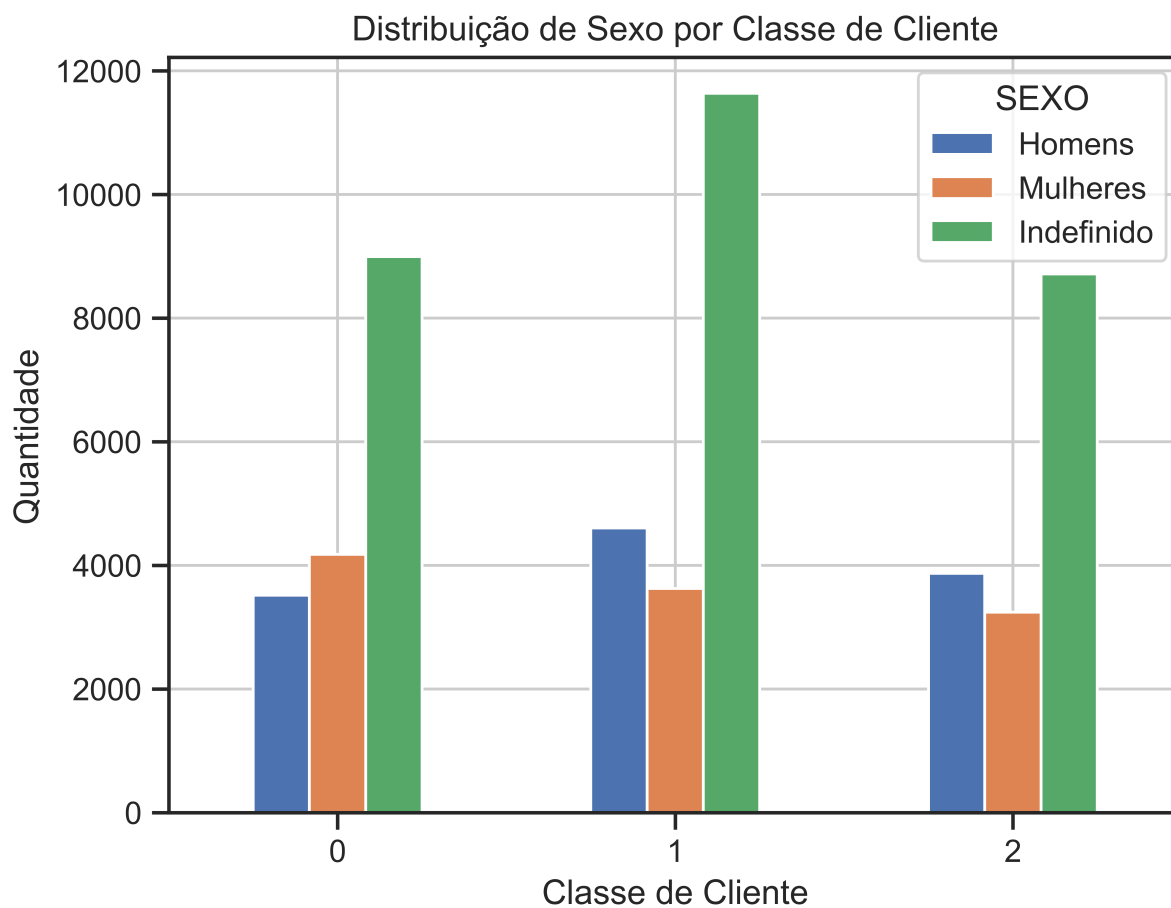
priori de cada cluster se alinham às descobertas proporcionadas pelo algoritmo.

Em relação aos outros requisitos funcionais do projeto, foi elaborado um dashboard usando a biblioteca *Streamlit* em Python, visando a disponibilização dos dados e uma visualização rápida. Esse dashboard foi concebido para transformar a visualização em uma página web hospedada dentro do servidor analítico. Para garantir a fluidez da página, todo o processamento e segmentação dos dados são previamente realizados, sendo exportados para um arquivo CSV que é acessado pela página web, onde as visualizações são processadas.

Os requisitos não funcionais, particularmente em relação à disponibilização remota e ao controle de acesso ao sistema, foram satisfatoriamente alcançados, uma vez que há capacidade para controlar o acesso ao servidor analítico. Contudo, para expandir a cobertura a um maior número de lojas, seria necessária uma infraestrutura mais robusta. As limitações da máquina local tornam-se evidentes, demonstrando saturação mesmo ao lidar apenas com três lojas. Embora a segmentação por loja seja um objetivo alcançável com a infraestrutura atual, priorizou-se a disponibilização do sistema sobre essa funcionalidade específica.

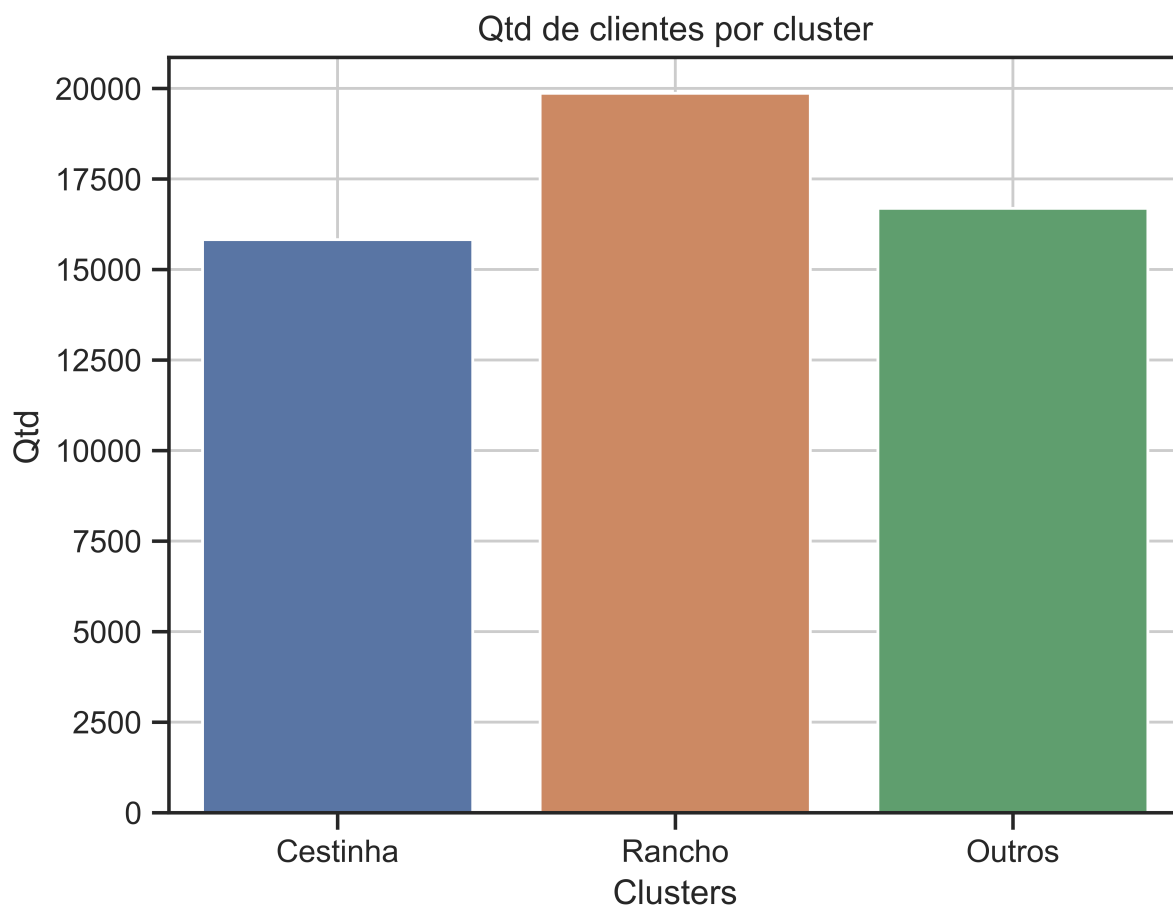
Um outro requisito não funcional relevante para ampliar a análise seria a identificação de outros clusters, como, por exemplo, clientes aficionados por vinho. Contudo, para realizar essa análise, seriam necessários novos atributos, como os gastos por cliente nessa seção específica ou a comparação do gasto médio por cliente com o gasto geral dos clientes em uma categoria específica. Devido ao considerável esforço demandado, esse requisito foi desconsiderado em termos de prioridade.

Figura 15 – Distribuição de gênero por cluster



Fonte: Autor

Figura 16 – Quantidade de clientes por cluster



Fonte: Autor

6 ANÁLISE DOS RESULTADOS E IMPACTO DA SOLUÇÃO PROPOSTA

Tão crucial quanto o processo de desenvolvimento do sistema é a avaliação meticulosa de seus resultados. Será conduzida uma avaliação para determinar se a arquitetura proposta e o sistema desenvolvido atenderam a todos os requisitos iniciais estabelecidos. Ao término desse processo, será realizada uma análise dos resultados sob duas perspectivas: a primeira relacionada ao desempenho em si e a segunda direcionada aos aspectos de negócios.

O desenvolvimento deste projeto foi impulsionado pela participação crucial da inteligência artificial, especialmente no campo do aprendizado de máquina, dentro da ampla área da IA. Foi utilizado o algoritmo de misturas gaussianas, detalhado na seção 3.4.1.4, para identificar os padrões intrínsecos e subjacentes nos dados do sistema CRM. Este algoritmo foi alimentado com os registros de compras de cada cliente durante o período analisado, considerando aqueles que realizaram pelo menos uma compra em uma das três lojas entre 01/09/2022 e 31/12/2022.

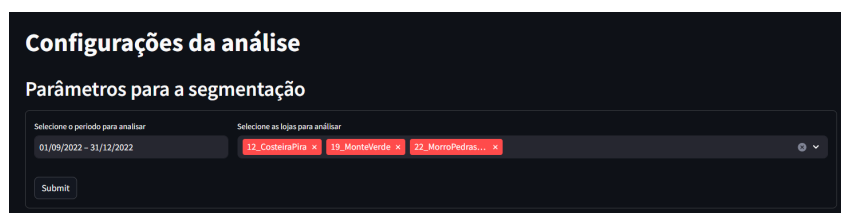
Cada tabela presente na Figura 4 contribuiu para buscar informações relevantes de cada cliente. Foi realizada a junção de todas as tabelas para criar conjuntos de dados únicos, conectando cada cliente a suas compras e itens adquiridos. Em seguida, aplicou-se o processo de tratamento de dados descrito na seção 5.3, no qual ocorreu a seleção das variáveis importantes e a mineração de dados para filtrar informações pertinentes ao modelo. Isso possibilitou a identificação de novos atributos para os clientes a partir dos dados existentes. Aqui, a mineração de dados desempenhou um papel essencial ao descobrir novos atributos, como diversidade de compras, total de descontos e quantidade total de itens adquiridos.

Após a preparação da base de dados, realizou-se a padronização dos dados, conforme 5.3.4, para melhorar o desempenho do modelo. Em seguida, o modelo foi alimentado com esses dados para aprender e identificar os clusters aos quais cada cliente pertence.

Essa aplicação permitiu não apenas a identificação, mas também a compreensão e aprendizado desses padrões, essenciais para a classificação eficiente dos clientes. Com essa abordagem, a segmentação deixou de ser um processo manual, baseado principalmente no conhecimento do analista, para se tornar um procedimento orientado por dados. Para ilustrar essa transição, podemos observar a diferença entre o processo de segmentação inicial, conforme mostrado na Figura 3, e o novo processo simplificado, representado na Figura 17. Essa transformação significativa só foi viável graças ao papel fundamental desempenhado pela inteligência artificial no refinamento e na automação do processo de segmentação.

Este capítulo consiste em apresentar e explicar o resultado final dos testes realizados. O resultado final consiste na escolha das variáveis de entrada e no método de

Figura 17 – Dashboard desenvolvida



Fonte: Autor

pré-processamento escolhido. Também, será exposto os valores finais dos hiperparâmetros escolhidos da biblioteca GMM. A análise dos experimentos foi feita a partir de conhecimento tácito previamente colhido da área de negócios e da validação desses resultados com os clusters apriori definidos.

Os resultados serão apresentados, primeiramente, mostrando-se as variáveis selecionadas. Depois, a escolha do método de pré-processamento. Após, a seleção dos hiperparâmetros e resultados das métricas de avaliação.

6.1 MODELO DE PREDIÇÃO FINAL

O modelo de clusterização visa conseguir distinguir entre o universo de clientes clube, aqueles que se enquadram em três perfis distintos, estes que foram definidos apriori, majoritariamente com base na metodologia RFV. É um algoritmo desenvolvido utilizando a biblioteca *sklearn*, principalmente *sklearn mixture Gaussian Mixture* que possui como entradas atributos dos clientes identificados nas vendas em complemento com dados cadastrais.

As variáveis independentes que apresentaram uma maior relevância com a variável dependente podem ser visualizadas na tabela 5.

Tabela 5 – Variáveis do modelo

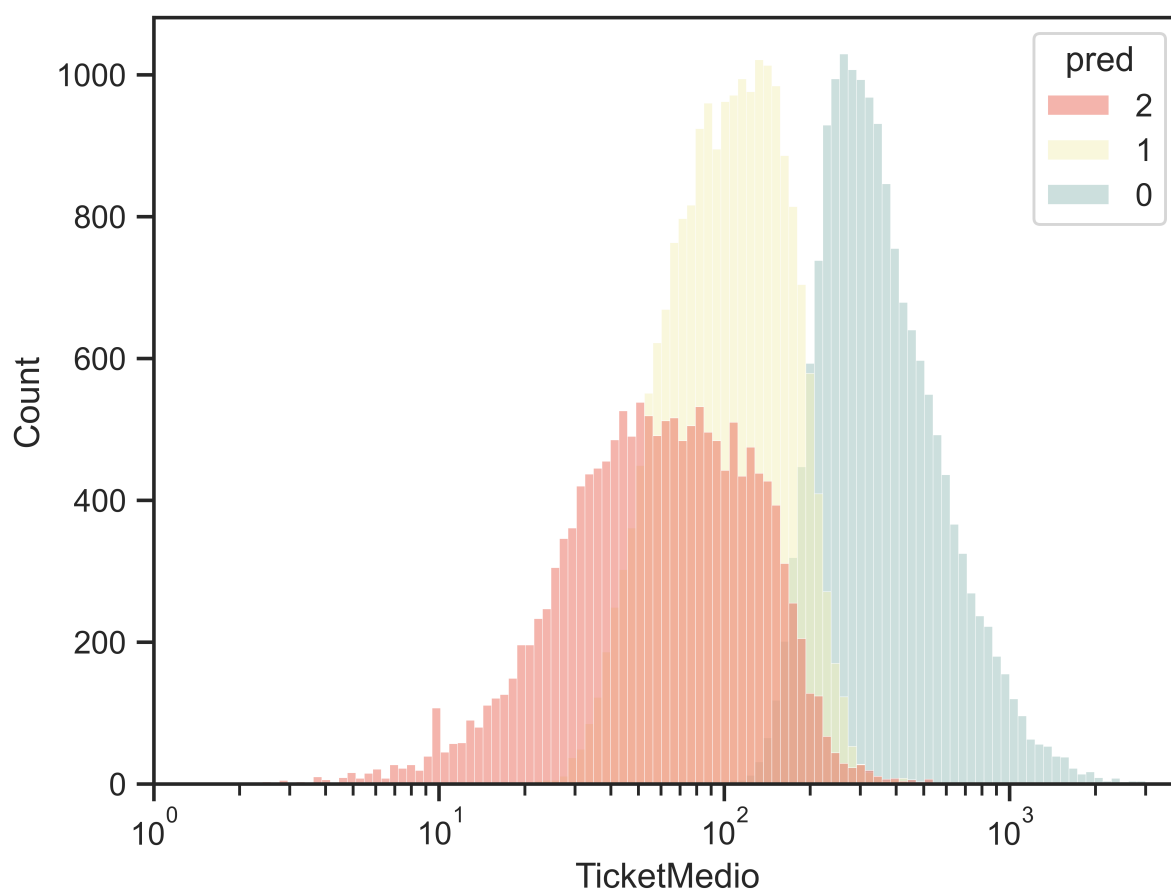
Variáveis selecionadas
Ticket Médio
Frequência
Recência
Diversidade
Total de descontos
Sexo

Para validar a segmentação atribuiu-se maior prioridade à análise da métrica Ticket Médio. Essa escolha decorre da relevância do Ticket Médio como um dos principais fatores de distinção entre os clusters que foram pré-definidos. Esperava-se que

essa métrica revelasse uma segmentação significativa entre os grupos, e o algoritmo de Misturas Gaussianas foi empregado com sucesso para atingir esse objetivo.

Na Figura 18, é possível observar como o algoritmo identificou três distribuições gaussianas distintas em relação ao Ticket Médio. Ressalta-se que o eixo x da figura está em escala logarítmica, o que foi adotado para aprimorar a visualização dos dados. Essa abordagem permite compreender, de forma mais aprofundada, como o Ticket Médio varia entre os diferentes clusters de clientes.

Figura 18 – Histograma do Ticket Médio por Cluster

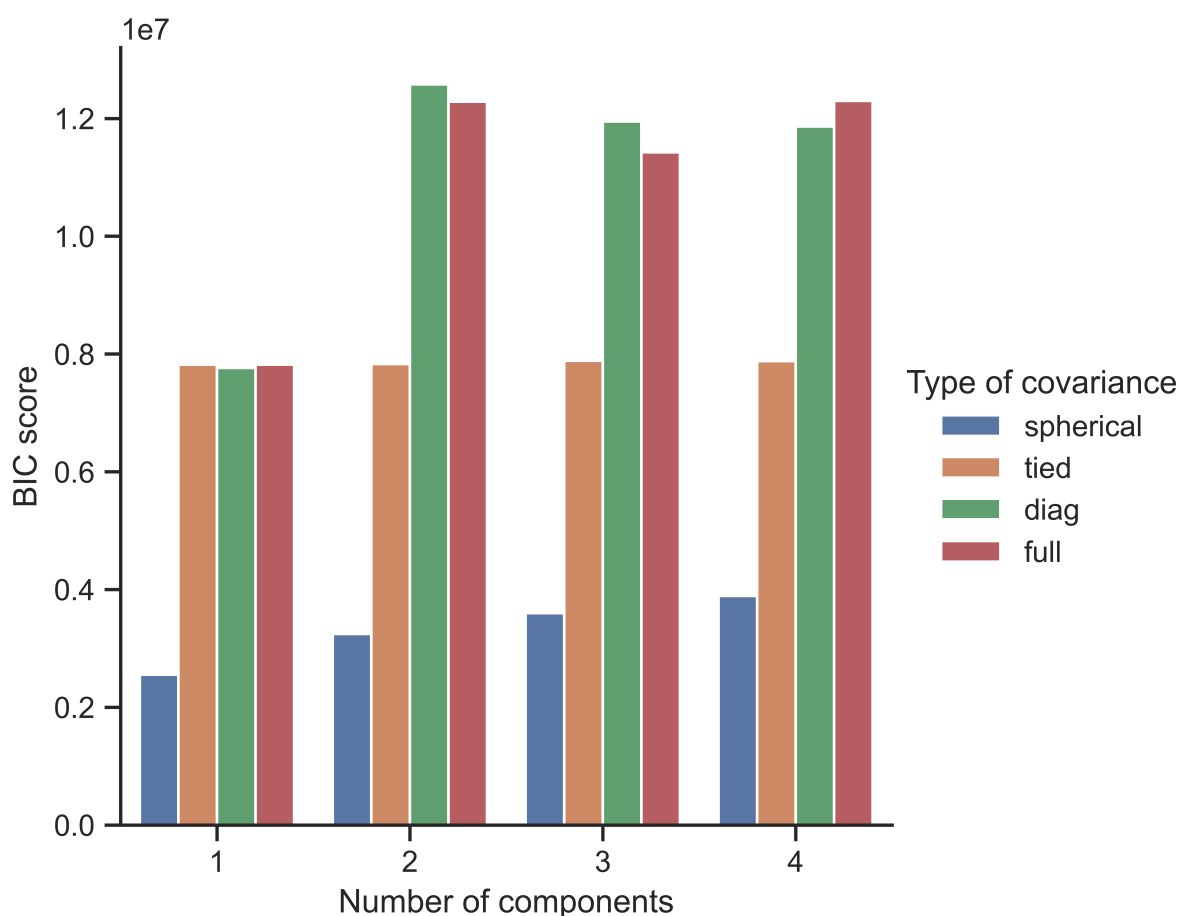


Fonte: Autor

Outro elemento crítico do projeto é a decisão de não realizar a divisão do conjunto de dados em conjuntos de teste e validação. Essa escolha decorre do fato de se tratar de uma aplicação de aprendizado não supervisionado, na qual a validação de problemas de *overfitting* e a avaliação da performance do algoritmo não são diretamente aplicáveis. Em virtude dessa natureza do problema, optou-se por empregar todos os dados disponíveis no processo de clusterização. Isso significa que, neste estudo, não é reservada uma parcela dos dados para fins de validação ou teste, em contrapartida, utiliza-se a totalidade do conjunto de dados na análise.

Com relação aos hiper-parâmetros utilizados no modelo, como apriori já se sabe o número de clusters que deve-se encontrar, o segundo hiper-parâmetro a ser definido é referente à Inicialização dos Componentes, este hiper-parâmetro define a estratégia de inicialização dos parâmetros do modelo, como as médias, covariâncias e pesos dos componentes gaussianos, como a biblioteca sklearn já utiliza o k-means como algoritmo padrão ele foi utilizado para a inicialização, e por último foi definido o parâmetro do tipo de Covariância, este hiper-parâmetro determina o tipo de matriz de covariância a ser usada nos componentes gaussianos. Pode ser "full"(matriz de covariância completa), "tied"(uma única matriz de covariância compartilhada entre todos os componentes), "diag"(matriz de covariância diagonal) ou "spherical"(matriz de covariância esférica). Para determinar esse último hiper-parâmetro foi utilizado o critério de informação Bayesiano (Bayesian information criterion) ou BIC, conforme figura 19

Figura 19 – Critério de informação Bayesiano



Fonte: Autor

Ao escolher entre vários modelos, geralmente prefere-se aqueles com valores mais baixos do critério de informação bayesiano. O BIC é uma função crescente da

variância do erro σ_{ϵ}^2 e uma função crescente de k . Ou seja, a variação não explicada na variável dependente e o número de variáveis explicativas aumentam o valor do BIC. E analisando a Figura 19 nota-se que não há muita diferença entre os valores de BIC para 2 ou 3 clusters, e ainda que o tipo de covariância com o menor valor é a spherical. Logo seguiu-se com o hiper-parâmetro spherical para o modelo.

6.2 VALIDAÇÃO DOS RESULTADOS

Para validar os cluster obtidos foram elaboradas validações, para que em conjunto com a área de negócios, que auxiliassem a melhor visualizar as características dos cluster e como se comportam entre si. Para tal validação foram realizados pontos de controle durante o projeto e apresentadas as evoluções do sistema e o resultado final. Nos momentos de contatos era detalhado todo o desenvolvimento presente do modelo e da *dashboard* que servirá para visualização dos resultados e feito uma análise crítica em conjunto com os stakeholders sobre os pontos de melhoria presentes e resultados obtidos. Durante o desenvolvimento, foram adotadas iterações semanais com a supervisora do projeto, e em cada marco do projeto foram realizadas apresentações à gestora da área de CRM para validar os resultados, nesses encontros era validado o desenvolvimento e se necessário o caminho era pivotado. E os resultados finais foram apresentados e validados no mesmo processo com ambas as gestoras. Outro ponto importante do projeto é a redução do tempo despendido realizando as análises pelo time de CRM, já que na nova ferramenta os dados já estão pré-processados o que possibilita que os dados sejam apenas acessados e não ocorra a necessidade de maior processamento computacional, resultando em uma diminuição significativa no consumo de recursos computacionais e tempo de processamento. Para efeitos de comparação, para realizar uma análise na região da grande Florianópolis, englobando São José e Palhoça, o tempo despendido apenas para gerar os dados girava em torno de 2h, agora na nova ferramenta pode ser gerado em 15 minutos.

Conforme Figura 20, observa-se como cada cluster possui sua própria característica quando comparado com os outros, para poder realizar essa comparação os atributos foram normalizados em um intervalo de 0 a 5, sendo que foi atribuído 0 para o cluster que tiver o menor valor desse atributo e 5 para o cluster com maior valor. Por exemplo, apesar de os clusters Cestinha e Outros possuírem um valor parecido de Ticket Médio, eles se destoam entre si na recência e frequência, já que apresentam uma relação praticamente oposta nesses atributos. Enquanto os clientes do grupo Cestinha possuem uma frequência de 5 e uma recência de 0, os clientes do cluster Outros possuem uma frequência de 0 e uma recência de 5.

Outro dado que pode ser utilizado para misturas gaussianas para a validação dos clusters é o coeficiente de silhueta que avalia quão bem cada ponto de dados se ajusta ao seu próprio cluster em comparação com os clusters vizinhos mais próximos.

Figura 20 – Características dos clusters comparadas entre si

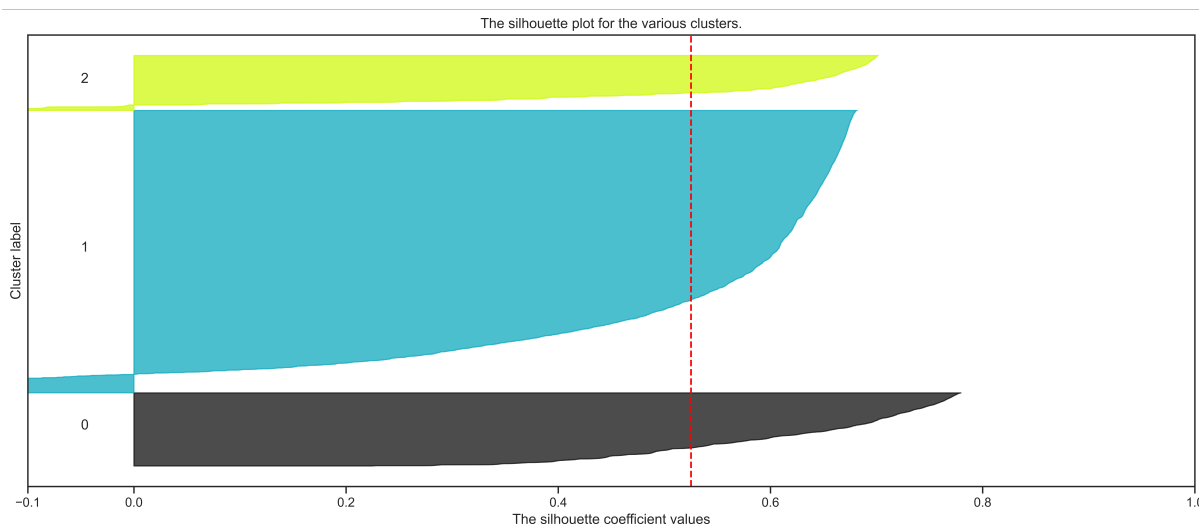


Fonte: Autor

Ele combina a noção de separação entre clusters com a coesão interna de cada cluster, proporcionando uma avaliação abrangente da qualidade da clusterização.

O resultado desse cálculo varia de -1 a 1. Valores próximos de 1 indicam clusters densos e bem distintos, valores próximos de 0 sugerem que os clusters estão sobrepostos, e valores negativos indicam uma clusterização inadequada.

Figura 21 – Coeficiente de silhueta para 3 clusters



Fonte: Autor

Pode-se notar que os cluster 1 e 2 podem possuir clientes em clusters inadequados, mas não houve sobreposição dos clusters identificados.

Um dos pontos cruciais deste projeto reside na validação dos resultados pelos *stakeholders*. No âmbito do CRM, almeja-se uma abordagem mais estratégica, visando a geração de análises inovadoras para impulsionar a expansão do negócio.

Até o momento, essas análises têm sido conduzidas principalmente com base no conhecimento tácito dos analistas. No entanto, com a implementação deste novo sistema, busca-se embasar o processo de tomada de decisão em dados concretos, especialmente ao abordar questões relacionadas à migração e ao comportamento dos clientes, convertendo essas informações em dados tangíveis e mais precisos.

Outro ponto crucial a considerar é o tempo atualmente consumido para realizar análises. Esse processo demanda um esforço considerável da equipe, desviando o foco de análises estratégicas para um trabalho repetitivo e demorado. Com a introdução do sistema, essas análises serão significativamente mais ágeis, reduzindo substancialmente o tempo necessário para sua execução. Essa mudança permitirá uma transição para uma tomada de decisão mais estratégica, liberando os analistas para se concentrarem em análises de maior valor estratégico, em vez de tarefas operacionais demoradas.

Com a vasta gama de dados disponíveis, é possível realizar análises detalhadas dos diferentes segmentos de clientes. Podemos identificar quais deles estão inativos e compreender o perfil desses clientes em relação a diversos parâmetros, como sexo, frequência de compra, histórico de descontos aproveitados e outros. Esse tipo de análise permite recuperar uma parcela desses clientes inativos, que possuem um custo de aquisição menor do que aquele cliente que ainda não conhece o Bistek, e são despendidos esforços e ações para atraí-lo para uma loja.

Podemos aprofundar a análise dos clientes já ativos na base, elaborando estratégias específicas para potencializar o faturamento desse grupo. Isso implica não apenas em aumentar a frequência das visitas à loja, mas também em incrementar o valor médio gasto por visita, ou seja, o ticket médio.

O foco está em maximizar o potencial de gastos desses clientes já engajados, desenvolvendo iniciativas que os incentivem a retornar mais vezes à loja ou a realizarem compras mais substanciais, contribuindo, assim, para o crescimento do faturamento geral da empresa.

7 CONCLUSÕES E SUGESTÕES PARA TRABALHOS FUTUROS

Este estudo emerge da necessidade de fundamentar estratégias de marketing direcionadas aos distintos segmentos de clientes associados ao Clube Bistek. Dada a vasta diversidade de participantes no programa, a segmentação se revela como um diferencial para a concepção de estratégias assertivas e de alta precisão. O propósito central desta pesquisa consistiu em identificar os principais segmentos de clientes presentes na rede, adotando critérios como as métricas de Frequência, Recência e Valor.

A abordagem metodológica para alcançar tal identificação envolveu o uso do algoritmo de misturas Gaussianas, escolhido devido à sua adequação à distribuição dos dados observados. Os resultados obtidos foram explorados e discutidos ao longo deste trabalho, demonstrando-se como eficazes e satisfatórios para a caracterização dos grupos identificados.

A análise detalhada dos segmentos de clientes revelou *insights* valiosos que podem ser estrategicamente aplicados na customização e otimização das estratégias de marketing direcionadas ao Clube Bistek.

O método de clusterização adotados demanda a pré-definição do número de clusters, esse que foi escolhido em alinhamentos realizados com as áreas de negócio da empresa, em que foram mapeados os três principais clusters, foco deste trabalho. Essa quantidade de clusters apresentou uma diferenciação satisfatória e condizentes com as expectativas previamente estabelecidas. Isto é, em cada um dos dois clusters que haviam características marcantes, essas características foram identificadas.

Os resultados obtidos foram então analisados com intuito de aprofundar o conhecimento acerca do conjunto de clientes que compõem cada cluster encontrado. Pode-se ter uma visualização dos clusters encontrados na Figura 22.

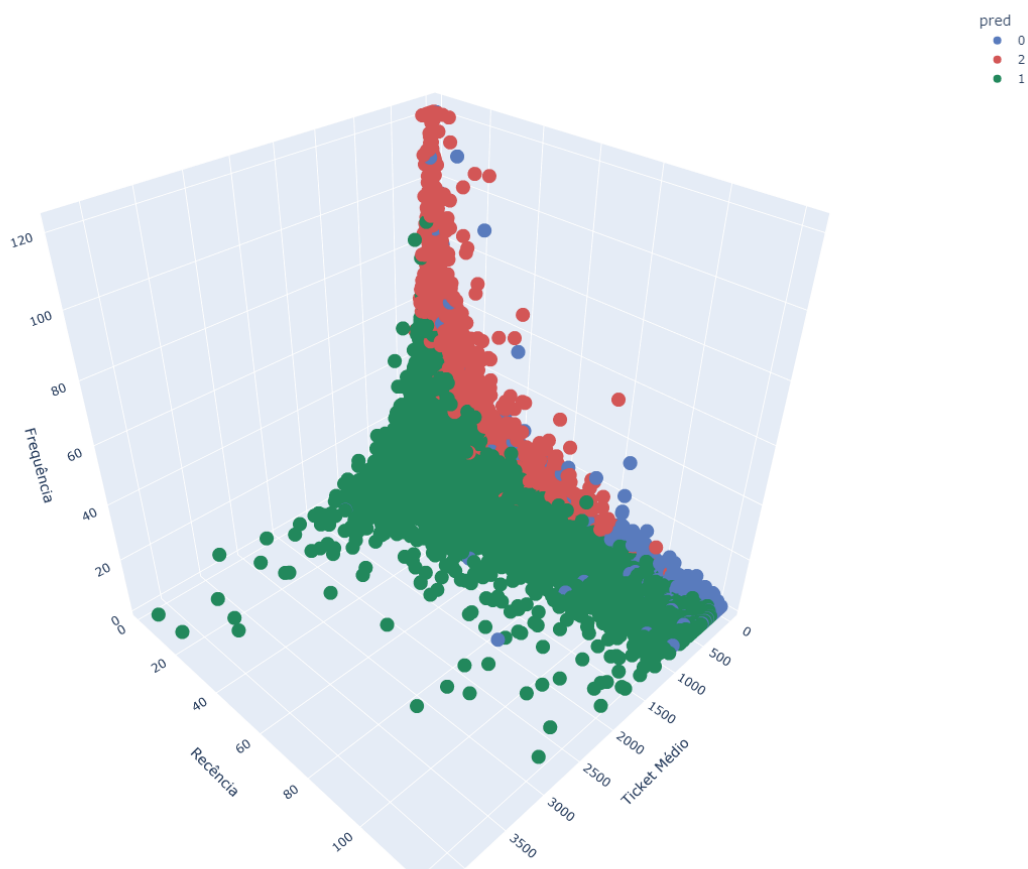
7.1 TRABALHOS FUTUROS

Esta seção do trabalho é dedicada a explicar as possíveis melhorias do sistema de segmentação, para que ele obtenha uma maior acurácia e consiga se adequar melhor aos requisitos traçados.

7.1.1 Aumento das classes identificadas

Neste estudo, a segmentação se concentrou em apenas três grupos, utilizando padrões globais de compras dos clientes como critério. Contudo, uma abordagem mais abrangente e refinada poderia ser explorada em trabalhos futuros. A consideração de uma variedade mais extensa de atributos, ou a análise mais detalhada dos dados, como a divisão do ticket médio por seções específicas, ofereceria uma perspectiva

Figura 22 – Visualização da Segmentação



Fonte: Autor

mais precisa.

Ao invés de avaliar apenas o ticket médio geral do cliente, essa abordagem mais granular permitiria identificar em quais seções específicas o cliente está adquirindo mais itens ou apresentando um ticket médio mais elevado. Essa granularidade aprimorada proporcionaria uma segmentação mais minuciosa, possibilitando a classificação dos clientes em uma quantidade maior de grupos distintos. Como resultado, as campanhas de marketing poderiam ser ainda mais personalizadas, adaptando-se de maneira mais precisa aos padrões de compra específicos de cada segmento de clientes.

Dessa forma, a expansão dos atributos considerados e a maior granularidade dos dados não apenas enriqueceriam a análise de segmentação, mas também ofereceriam oportunidades para otimizar ainda mais a eficácia das estratégias de marketing, proporcionando uma abordagem mais personalizada e alinhada aos comportamentos de compra individualizados de cada cliente.

7.1.2 Aumento do conjunto de dados

Este estudo foi conduzido, considerando apenas as lojas de uma região geográfica específica, embora a rede de supermercados esteja presente em dois estados brasileiros. A escolha por esse escopo restrito foi motivada pela necessidade de estabelecer uma prova de conceito efetiva. Entretanto, é importante salientar que a ampliação do conjunto de dados, incorporando mais lojas, pode proporcionar valiosos *insights* e estratégias adicionais para a equipe de marketing.

Outra limitação do sistema é referente ao período analisado, já que atualmente os limites de tempo dos dados são estáticos, sendo possível apenas reduzir o escopo da análise e não abranger um conjunto maior de dados. Esse é um ponto importante de ser considerado, já que o período analisado é afetado por datas importantes para o varejo alimentar e que podem alterar os resultados se comparados a outros meses do ano.

A inclusão de novas lojas em diferentes regiões geográficas pode enriquecer significativamente a compreensão dos padrões de compras, possibilitando uma segmentação mais refinada das campanhas. Especificamente, ao considerar clientes que mudam de estado ou ao levar em conta as mesorregiões dos estados, os padrões de compra e os perfis dos grupos identificados podem apresentar variações notáveis.

Assim, a expansão do escopo geográfico deste estudo ofereceria à equipe de marketing uma visão mais abrangente das nuances regionais, permitindo ajustes mais precisos nas estratégias de segmentação.

7.1.3 Integração do Treinamento Online para Atualização Dinâmica dos Algoritmos

A atual configuração do sistema não incorpora a funcionalidade de atualizar o algoritmo com base na variação de novas informações. Em outras palavras, toda vez que novos dados são adicionados ao banco de dados, a necessidade de retreinar o sistema é inerente. Essa abordagem implica que a adaptação do algoritmo às mudanças nas vendas ocorre de maneira mais lenta.

Uma implementação vantajosa e estratégica seria a introdução de uma etapa de treinamento online, utilizando exclusivamente os dados adicionais. Essa abordagem permitiria que o algoritmo se ajustasse dinamicamente e de forma mais ágil às alterações nas tendências de vendas. Ao adotar o treinamento online, o sistema se beneficiaria da capacidade contínua de aprendizado, atualizando-se de maneira incremental e garantindo que as recomendações permaneçam alinhadas com as dinâmicas em constante evolução do comportamento do cliente e das preferências de compra.

Essa estratégia não apenas aprimoraria a agilidade do sistema em se adaptar a novos padrões de consumo, mas também contribuiria para uma resposta mais rápida

e precisa às mudanças no mercado.

Outro ponto que aflige o sistema e a regionalização do comportamento dos clientes, observa-se já hoje na rede uma mudança no comportamento de compras de clientes de regiões geográficas distintas, fazendo com que para adição de mais lojas da rede seja necessário um retreinamento do modelo, ou até a separação em dois modelos distintos.

7.1.4 Ajuste dos atributos

Outro fator de relevância destacado no estudo foram os atributos, tais como a Diversidade e o Total de Descontos. Notou-se que a forma como esses atributos foram calculados teve um impacto significativo nos resultados obtidos. Diante disso, uma oportunidade para futuras investigações surge, direcionada à realização de um novo estudo a fim de verificar possíveis melhorias e otimizações nos cálculos desses atributos.

É importante considerar que, dependendo da abordagem utilizada nos cálculos, os resultados podem variar consideravelmente. Neste contexto, é especialmente relevante explorar estratégias que minimizem problemas potenciais, como a multicolinearidade, que podem surgir na análise de atributos complexos como Diversidade e Total de Descontos.

REFERÊNCIAS

- ABRAS. **Um setor forte na economia brasileira**. [S.l.: s.n.], 2022. Disponível em: <https://www.abras.com.br/economia-e-pesquisa/ranking-abras/dados-gerais>. Acesso em: 14 de novembro 2023.
- AGRAWAL, Rakesh; GEHRKE, Johannes; GUNOPULOS, Dimitrios; RAGHAVAN, Prabhakar. Automatic subspace clustering of high dimensional data for data mining applications. *In: PROCEEDINGS of the 1998 ACM SIGMOD international conference on Management of data*. [S.l.: s.n.], 1998. P. 94–105.
- BARBARÁ, Daniel; CHEN, Ping. Using the fractal dimension to cluster datasets. *In: PROCEEDINGS of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*. [S.l.: s.n.], 2000. P. 260–264.
- BERNUZZI, Gabriel Marques. Segmentação de mercado em academias esportivas: uma revisão da literatura. Universidade Estadual Paulista (Unesp), 2022.
- BEZDEK, James C; EHRLICH, Robert; FULL, William. FCM: The fuzzy c-means clustering algorithm. **Computers & geosciences**, Elsevier, v. 10, n. 2-3, p. 191–203, 1984.
- CHAUBEY, Dhani; SUBRAMANIAN, Kalpathy. The Complexity of Market Segmentation Process. v. 7, p. 25–30, mai. 2020.
- CHEIDA, Marcel José. O poder, o monopólio e a produção exponencial de informação. **Cadernos de Fé e Cultura**, v. 2, n. 1, p. 31–47, maio 2017. DOI: 10.24220/cfc.v2i1.3941. Disponível em: <https://periodicos.puc-campinas.edu.br/cadernos/article/view/3941>.
- CHURCHILL, Gilbert A.; PETER, J. Paul. **Marketing: Criando Valor para o Cliente**. 2ª edição. São Paulo: Editora Saraiva, 2005.
- COOIL, Bruce; AKSOY, Lerzan; KEININGHAM, Timothy. Approaches to Customer Segmentation. **Journal of Relationship Marketing**, v. 6, p. 9–39, jan. 2007. DOI: 10.1300/J366v06n03_02.
- DAVE, Rajesh N; BHASWAN, Kurra. Adaptive fuzzy c-shells clustering and detection of ellipses. **IEEE Transactions on Neural Networks**, IEEE, v. 3, n. 5, p. 643–662, 1992.

DOGAN, Onur; AYÇIN, Ejder; BULUT, Zeki. CUSTOMER SEGMENTATION BY USING RFM MODEL AND CLUSTERING METHODS: A CASE STUDY IN RETAIL INDUSTRY. **International Journal of Contemporary Economics and Administrative Sciences**, v. 8, p. 1–19, jul. 2018.

ELEN, Abdullah; AVUÇLU, Emre. Standardized Variable Distances: A distance-based machine learning method. **Applied Soft Computing**, Elsevier, v. 98, p. 106855, 2021.

FARBOODI, Maryam; VELDKAMP, Laura. **A model of the data economy**. [S.l.], 2021.

FISHER, Douglas H. Knowledge acquisition via incremental conceptual clustering. **Machine learning**, Springer, v. 2, p. 139–172, 1987.

GIANESI, I G N; CORRÊA, Henrique Luiz. **Administracao estrategica de servicos: operacoes para a satisfacao do cliente**. [S.l.]: Atlas, 1994.

GUHA, Sudipto; RASTOGI, Rajeev; SHIM, Kyuseok. CURE: An efficient clustering algorithm for large databases. **ACM Sigmod record**, ACM New York, NY, USA, v. 27, n. 2, p. 73–84, 1998.

GUHA, Sudipto; RASTOGI, Rajeev; SHIM, Kyuseok. ROCK: A robust clustering algorithm for categorical attributes. **Information systems**, Elsevier, v. 25, n. 5, p. 345–366, 2000.

HALKIDI, Maria; BATISTAKIS, Yannis; VAZIRGIANNIS, Michalis. On clustering validation techniques. **Journal of intelligent information systems**, Springer, v. 17, p. 107–145, 2001.

JACE MCLEAN. **Data Never Sleeps 10**. [S.l.: s.n.], 2022. Disponível em: <https://www.domo.com/blog/data-never-sleeps-hits-double-digits/>. Acesso em: 24 de outubro 2023.

JAIN, Anil K; MURTY, M Narasimha; FLYNN, Patrick J. Data clustering: a review. **ACM computing surveys (CSUR)**, Acm New York, NY, USA, v. 31, n. 3, p. 264–323, 1999.

KARYPIS, George; HAN, Eui-Hong; KUMAR, Vipin. Chameleon: Hierarchical clustering using dynamic modeling. **computer**, IEEE, v. 32, n. 8, p. 68–75, 1999.

KOTLER, Philip. **Administração de Marketing: análise, planejamento, implementação e controle**. 5. ed. São Paulo: Atlas, 1998.

KOTLER, Philip; KELLER, Kevin Lane; YAMAMOTO, Sonia Midori; BARRETO, Iná Futino; CRESCITELLI, Edson. **Administração de Marketing**. 15th. São Paulo: Pearson Education do Brasil, 2018. ISBN 9788550813504.

LOBO, Luiz Carlos. **Inteligência artificial, o Futuro da Medicina e a Educação Médica**. v. 42. [S.l.]: SciELO Brasil, 2018. P. 3–8.

MANYIKA, JAMES ET AL. **Big data: The next frontier for innovation, competition, and productivity**. [S.l.: s.n.], 2011. Disponível em:

<https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/big-data-the-next-frontier-for-innovation>. Acesso em: 24 de outubro 2023.

MITCHELL, T.M. **Machine Learning**. [S.l.]: McGraw-Hill, 1997. (McGraw-Hill International Editions). ISBN 9780071154673. Disponível em:

<https://books.google.com.br/books?id=EoYBngEACAAJ>.

MONARD, Maria Carolina; BARANAUSKAS, José Augusto. Indução de regras e árvores de decisão. **Sistemas Inteligentes-fundamentos e aplicações**, sn, v. 1, p. 115–139, 2003.

NIMBALKAR, Divya; SHAH, Paulami. Data mining using RFM Analysis, p. 5, dez. 2013. DOI: 10.13140/RG.2.2.24229.04328.

RASMUSSEN, Carl. The infinite Gaussian mixture model. **Advances in neural information processing systems**, v. 12, 1999.

RICHERS, R. **Marketing: uma visão brasileira**. [S.l.]: Negócio Editora, 2000. ISBN 9788586014505. Disponível em:

<https://books.google.com.br/books?id=FOSUDfnYAKIC>.

RUSSELL, S.; NORVIG, P. **Inteligência artificial**. [S.l.]: Elsevier, 2004. ISBN 9788535211771. Disponível em:

<https://books.google.com.br/books?id=wBMvAAAACAAJ>.

SANTOS, Angela Maria Medeiros Martins; COSTA, Cláudia Soares. Características gerais do varejo no Brasil. Banco Nacional de Desenvolvimento Econômico e Social, 1997.

SCIKIT LEARN ORG. **Avaliação da performance do clustering**. 2023.

SHARAN, Roded; SHAMIR, Ron. CLICK: a clustering algorithm with applications to gene expression analysis. *In*: MARYLAND, MD, 307. PROC Int Conf Intell Syst Mol Biol. [S.l.: s.n.], 2000. P. 16.

TAMASO ROBERTO E FURTADO, Bruno. **O que esperar para o varejo alimentar em 2022 e nos próximos anos**. v. 42. [S.l.]: SciELO Brasil, 2018. P. 3–8.

TYNAN, A. Caroline; DRAYTON, Jennifer. Market segmentation. **Journal of Marketing Management**, v. 2, n. 3, p. 301–335, 1987. DOI: 10.1080/0267257x.1987.9964020. Disponível em: <https://app.dimensions.ai/details/publication/pub.1001045427>.

WANG, Kaijun; WANG, Baijie; PENG, Liuqing. CVAP: validation for cluster analyses. **Data Science Journal**, CODATA, v. 8, p. 88–93, 2009.

WANG, Wei; YANG, Jiong; MUNTZ, Richard *et al.* STING: A statistical information grid approach to spatial data mining. *In*: VLDB. [S.l.: s.n.], 1997. P. 186–195.

WEI, Jo-Ting; LIN, Shih-Yen; WU, Hsin-Hung. A review of the application of RFM model. **African Journal of Business Management December Special Review**, v. 4, p. 4199–4206, jan. 2010.

WEINSTEIN, Art. **Market Segmentation**. New York: McGraw-Hill, 1995.

WIND, Yoram. Issues and Advances in Segmentation Research. **Journal of Marketing Research**, v. 15, n. 3, p. 317–337, 1978. DOI: 10.1177/002224377801500302. eprint: <https://doi.org/10.1177/002224377801500302>. Disponível em: <https://doi.org/10.1177/002224377801500302>.

WITTEN, Daniela; JAMES, Gareth. **An introduction to statistical learning with applications in R**. [S.l.]: springer publication, 2013.

XU, Dongkuan; TIAN, Yingjie. A comprehensive survey of clustering algorithms. **Annals of Data Science**, Springer, v. 2, p. 165–193, 2015.

XU, Xiaowei; ESTER, Martin; KRIEGEL, H-P; SANDER, Jörg. A distribution-based clustering algorithm for mining in large spatial databases. *In: IEEE. PROCEEDINGS 14th International Conference on Data Engineering*. [S.l.: s.n.], 1998. P. 324–331.

YAGER, Ronald R; FILEV, Dimitar P. Approximate clustering via the mountain method. **IEEE Transactions on systems, man, and Cybernetics**, IEEE, v. 24, n. 8, p. 1279–1284, 1994.

ZHANG, Tian; RAMAKRISHNAN, Raghu; LIVNY, Miron. BIRCH: an efficient data clustering method for very large databases. **ACM sigmod record**, ACM New York, NY, USA, v. 25, n. 2, p. 103–114, 1996.

APÊNDICE A – CÓDIGO MISTURAS GAUSSIANAS

```
1 gm = GaussianMixture(n_components=3, covariance_type='spherical',  
    random_state=42, n_init=5).fit(X)  
2 pred = gm.predict(X)  
3 X['pred'] = pred
```

Listing A.1 – Código algoritmo de misturas gaussianas

APÊNDICE B – CÓDIGO PARA AJUSTE DAS VARIÁVEIS

```

1 categories = [[0,1,2], [i for i in recen_r], [i for i in freq_r]]
2 onehotencoder = OneHotEncoder(sparse=False, handle_unknown='ignore',
   categories=categories)
3 transformed_data = onehotencoder.fit_transform(X[['SEXO', 'Recen_cat', '
   Freq_cat']])
4 encoded_data = pd.DataFrame(transformed_data, index=X.index)
5 encoded_data.columns = onehotencoder.get_feature_names_out()
6 X = pd.concat([X, encoded_data], axis=1)
7 X.drop(columns=['SEXO', 'Recen_cat', 'Freq_cat'], inplace=True)

```

Listing B.1 – Código para transformação de variáveis categóricas

```

1
2 data.TicketMedio = np.log(data.TicketMedio)
3 data.Diversity = np.log(data.Diversity)
4 data.Tot_itens = np.log(data.Tot_itens)
5
6 s = StandardScaler()
7 data.SEXO = data.SEXO.replace({"M":0, "F":1, "Indefinido":2})
8 data['TicketMedio'] = s.fit_transform(data[['TicketMedio']])
9 data['Diversity'] = s.fit_transform(data[['Diversity']])
10 data['Tot_itens'] = s.fit_transform(data[['Tot_itens']])

```

Listing B.2 – Código para transformação de variáveis contínuas