SANTANA, Luiz Henrique Batista de; GARCIA, Domingos Sávio da Cunha; ARAÚJO, Leonardo Amorim de; LEAL, Carlos Artur Alevato; ASSUNÇÃO, Arthur Nascimento; ROSSETI, Gustavo José Santiago; MOREIRA, Silvana Rodrigues Pires; MACHADO, Lisleandra. Utilização Inovadora do *Machine Learning* para a Segmentação da Carteira de Clientes em uma Empresa de Hortifrutigranjeiros de Grande Porte.
DOI: <https://doi.org/10.21166/cpitt.v5i1.4918>. ISSN: 2763-5449

UTILIZAÇÃO INOVADORA DO *MACHINE LEARNING* PARA A SEGMENTAÇÃO DA CARTEIRA DE CLIENTES EM UMA EMPRESA DE HORTIFRUTIGRANJEIROS DE GRANDE PORTE[1]

INNOVATIVE USE OF MACHINE LEARNING TO SEGMENT THE CUSTOMER PORTFOLIO IN A LARGE FRUIT AND VEGETABLE WHOLESALE REDISTRIBUTOR COMPANY

Luiz Henrique Batista de Santana[1]
Domingos Sávio da Cunha Garcia[2]
Leonardo Amorim de Araújo[3]
Carlos Artur Alevato Leal[3]
Arthur Nascimento Assunção[3]
Gustavo José Santiago Rosseti[3]
Silvana Rodrigues Pires Moreira[3]
Lisleandra Machado[3]
[1]Escola Superior de Agricultura Luiz de Queiroz, Universidade de São Paulo
[2]Universidade do Estado de Mato Grosso, Campus Cáceres
[3]Instituto Federal do Sudeste de Minas Gerais, Campus Santos Dumont

RESUMO: Na gestão da cadeia de suprimentos, compreender os clientes e agir de modo a satisfazer suas necessidades e desejos proporciona diferenciais competitivos frente à concorrência. As estratégias de *marketing* tradicionais têm migrado do foco no produto para o foco no cliente, trazendo com essa evolução o conceito do *marketing* de relacionamento, um modelo de gestão que consiste na oferta de produtos e serviços, atendendo exatamente às necessidades individuais de cada cliente. Nesta pesquisa, buscou-se implementar algoritmos para a segmentação e classificação da carteira de clientes numa empresa de hortifrutigranjeiros, extraindo dos seus dados brutos informações de alta qualidade sobre as demandas dos clientes, construídas a partir da análise dos históricos de compras. O estudo foi orientado por uma metodologia de natureza aplicada, com objetivos descritivos e de abordagem quantitativa. Os resultados foram alcançados através da combinação de duas técnicas, obtendo-se na aplicação na primeira a redução dimensional de 64.135 (sessenta e quatro mil cento e trinta e cinco) observações para 1.082 (mil e oitenta e duas), sendo estas agrupadas por clientes, dando origem ao surgimento de 8 (oito) segmentos, que serviram de insumo para a aplicação da segunda técnica, desta vez, reduzindo os segmentos para 4 (quatro grupos), os quais, foram submetidos a uma análise descritiva, orientados pelas variáveis recência, frequência e valor, os pilares o modelo RFV, o que permitiu uma caracterização detalhada de cada grupo de consumidores.
Palavras-chave: *Marketing* de Relacionamento; Modelo RFV; *Machine Learning*; *Clustering*; *K-means*.

ABSTRACT: In supply chain management, understanding customers and acting to satisfy their needs and desires provides competitive advantages compared to the competition. Traditional marketing strategies have migrated from a focus on the product to a focus on the customer, bringing with this evolution the concept of relationship marketing, a management model that consists of offering products and services, meeting exactly the individual needs of each customer. In this research, we sought to implement algorithms for the segmentation and classification of the customer portfolio in a fruit and vegetable company, extracting high-quality

[1] O presente trabalho é fruto da Monografia de Conclusão do primeiro autor junto ao Curso de Especialização MBA em Digital Business da ESALQ-USP.

DOI: <https://doi.org/10.21166/cpitt.v5i1.4918>. ISSN: 2763-5449

information about customer demands from their raw data, constructed from the analysis of purchase historics. The study was guided by an applied methodology, with descriptive objectives and a quantitative approach. The results were achieved through the combination of two techniques, obtaining in the first application a dimensional reduction from 64,135 (sixty-four thousand, one hundred and thirty-five) observations to 1,082 (one thousand and eighty-two), these being grouped by clients , giving rise to the emergence of 8 (eight) segments, which served as input for the application of the second technique, this time, reducing the segments to 4 (four groups), which were subjected to a descriptive analysis, guided by the variables recency, frequency and value, the pillars of the RFV model, which allowed a detailed characterization of each consumer group.
Keywords: Relationship Marketing; RFV Model; Machine Learning; Clustering; K-means.

INTRODUÇÃO

As constantes mudanças no perfil dos clientes têm norteado diversos segmentos de negócios que compõem os elos da cadeia de suprimentos a direcionar os seus esforços e principalmente os seus investimentos, na busca por soluções que procurem entender o comportamento e os requisitos dos seus consumidores. As ações guiadas por essas premissas podem proporcionar às organizações diferenciais competitivos sobre os seus atuais concorrentes, bem como, sobre novos participantes do mercado. A inovação tecnológica aliada a uma abordagem empresarial orientada pela análise dos desejos, preferências, comportamentos e a lealdade dos clientes, podem ser o caminho para a obtenção de excelentes resultados (SWIFT, 2001).

Em sua definição tradicional, o ‘marketing’ é o “*conjunto de procedimentos e estratégias de otimização dos lucros que, através de pesquisas de mercado, busca adequar os produtos às necessidades dos consumidores*” MARKETING (2022). No entanto, o ‘marketing’ vem evoluindo e se modificando ao longo do tempo, o que não mudou, foram os objetivos das relações comerciais, estes, ainda buscam compreender os clientes e agir de modo a satisfazer suas necessidades e seus desejos, partindo de uma visão focada no produto para uma atenção maior no cliente Zenone (2019). Esta evolução trouxe à tona um novo conceito, o ‘marketing’ de relacionamento, uma estratégia de gestão que consiste na oferta de produtos e serviços atendendo às necessidades individuais de cada cliente. Swift (2001) afirmou que para satisfazer aos desejos ou necessidades dos clientes obtendo alto índice de retenção e alta lucratividade as empresas precisam do produto (ou serviço), para o cliente certo, pelo preço ideal, na hora oportuna e pelos canais corretos. A aplicação dos princípios do ‘marketing’ em práticas de relacionamentos na cadeia de suprimentos é sem dúvidas uma iniciativa que fortalece e incentiva tais práticas como um elemento estratégico e altamente lucrativo para aqueles que a adotarem.

Diferentes dos programas de computador que vieram antes deles, os algoritmos de *Machine Learning*, termo em inglês para a tecnologia conhecida no Brasil como aprendizado de máquina, permitem que um computador varie as suas respostas introduzindo um ciclo de *feedbacks* para as respostas boas e ruins. Os algoritmos de aprendizagem de máquina fundamentalmente aprendem com a experiência. A aprendizagem de máquina é uma ferramenta poderosa para reconhecer padrões e prever o comportamento com base em dados históricos. Padrões de reconhecimento podem significar desde reconhecimento de caracteres até à manutenção preditiva, recomendação de produtos a clientes com base em

compras anteriores. O objetivo da aprendizagem de máquina é produzir e apresentar dados, estes, a matéria prima para tomar melhores decisões (NORMAN, 2019).

Na gestão do relacionamento com o cliente, a análise de RFV (Recência, Frequência e Valor) apresenta-se como fundamental para o planejamento de ações de 'marketing' nas carteiras de clientes. São métricas usadas para quantificar o histórico de transações de cada cliente em um banco de dados. Define-se como recência como uma medida de quanto tempo se passou desde a última transação do cliente com a empresa; a frequência como uma medida de quão frequente um cliente efetua transações e o valor monetário como o gasto médio feito por transação. Estratégias baseadas em RFV buscam métricas ou regras para avaliar o comportamento e valor do cliente para a empresa. Este tipo de análise classifica os clientes em grupos de acordo com suas medidas, e relaciona essas classificações a comportamentos com a probabilidade de positivação em uma oportunidade de negócio (BLATTEBERGER et al., 2008).

O presente trabalho busca descrever o processo de implementação de algoritmos para a segmentação e classificação da carteira de clientes, para a empresa objeto desta pesquisa, como uma oportunidade de transformar os seus dados brutos em compreensão, ideias e conhecimento a partir da análise dos históricos de compras dos seus clientes. E os seus resultados podem responder a questionamentos relacionados a conversão de vendas e a fidelização dos clientes.

MATERIAIS E MÉTODOS

Situada na cidade de Aracaju, no estado de Sergipe, a empresa objeto desta pesquisa é um comércio atacadista distribuidor de produtos hortifrutigranjeiros. Apesar do termo significar uma variedade de gêneros alimentícios, ovos de granja em seus diversos tamanhos e cores são o carro chefe nas transações comerciais estabelecidas com os seus mais de quatro mil clientes cadastrados em seu banco de dados. Abad (2022) definiu que as empresas que atuam neste segmento de mercado, atendem principalmente ao pequeno comércio varejista independente, fazendo as vendas por meio de visitas de vendedores e a entrega dos itens adquiridos no estabelecimento do cliente.

Este trabalho pode ser definido como uma pesquisa de natureza aplicada, com objetivos descritivos e de abordagem quantitativa, utilizando como meio a implementação de algoritmos de *Machine Learning,* ou aprendizado de máquina*,* aplicado a um conjunto de dados no intuito de segmentar e classificar a carteira de clientes utilizada neste estudo. Em essência, esta tecnologia é fundamentada na utilização de modelos estatísticos aplicados e desenvolvidos para extrair informação de conjuntos de dados. Existem dois tipos de aprendizado de máquina, o supervisionado, que objetiva uma análise confirmatória, preditiva, de um dado fenômeno estudado e o não supervisionado, que objetiva uma análise exploratória, diagnóstica, de um dado fenômeno estudado (MORETTIN et al., 2020).

O desenvolvimento da pesquisa levou em consideração uma amostra de transações comerciais realizadas pela empresa, no período entre 01/08/2021 até 31/08/2022. A definição do período foi fundamentada na informação de que prazo médio de recebimento das vendas é de oito dias, pois, em virtude ao alto giro dos produtos comercializados, seus vendedores realizam visitas semanais aos clientes e na oportunidade, além de captarem novas vendas, realizam cobranças de vendas anteriores. Além desta definição, entendeu-se que a dinâmica do negócio gera diariamente um grande volume de dados, e que, dados históricos anteriores ao

DOI: <https://doi.org/10.21166/cpitt.v5i1.4918>. ISSN: 2763-5449

recorte da amostra, não possuem mais tanta relevância para a implantação da segmentação da carteira de clientes.

A fonte original dos dados disponibilizados pela empresa objeto deste estudo, está originalmente armazenada no formato padrão do sistema gerenciador de bancos de dados 'Microsoft SQL Server'. Os modelos das análises RFV (Recência, Frequência e Valor) e de agrupamentos foram desenvolvidos através da linguagem de programação estatística *R 4.2.1 R Core Team* (2022), em ambiente *RStudio Team* (2022), além disso, foram utilizadas nesta pesquisa, funções adicionais à linguaguem R padrão, os chamados pacotes da linguagem, algoritmos com diferentes finalidades que potencializaram a análise dos dados. O Quadro 1 a seguir exibe os pacotes utilizados, o que fazem e onde encontrar referências sobre eles.

Quadro 1 – Relação dos pacotes que adicionaram funcionalidades à linguagem R padrão

Pacote	O que faz?	Referências
DBI	Uma definição de interface de banco de dados para comunicação entre R e sistemas de gerenciamento de banco de dados relacionais.	https://CRAN.R-project.org/package=DBI
dplyr	Manipulação de dados: agregar, sumarizar, filtrar, ordenar, criar variáveis, joins, dentre outras.	https://CRAN.R-project.org/package=dplyr
odbc	Conecta a bancos de dados compatíveis com ODBC, usando a interface DBI.	https://CRAN.R-project.org/package=odbc
dbplyr	Permite trabalhar com tabelas de banco de dados remotas como se fossem quadros de dados na memória.	https://CRAN.R-project.org/package=dbplyr
lubridate	Funções para trabalhar com data-hora e intervalos de tempo: análise rápida e amigável de dados de data-hora, extração e atualização de componentes de uma data-hora (anos, meses, dias, horas, minutos e segundos), manipulação algébrica em objetos de data-hora e intervalo de tempo.	https://CRAN.R-project.org/package=lubridate
tidyverse	Manipulação, exploração e visualização de dados além de compartilharem uma filosofia de design comum.	https://CRAN.R-project.org/package=tidyverse
cluster	Localiza agrupamentos nos dados.	https://CRAN.R-project.org/package=cluster
dendexend	Oferece um conjunto de funções para estender objetos 'dendrogramas' em R, permitindo visualizar e comparar árvores de agrupamentos hierárquicos.	https://CRAN.R-project.org/package=dendextend
factoextra	Fornece algumas funções fáceis de usar para extrair e visualizar a saída de análises de dados multivariados.	https://CRAN.R-project.org/package=factoextra
Fpc	Vários métodos para clustering e validação de cluster. Agrupamento de pontos fixos.	https://CRAN.R-project.org/package=fpc
gridExtra	Fornece várias funções de nível de usuário para trabalhar com gráficos de grade, para principalmente para organizar vários gráficos baseados em grade em uma página e desenhar tabelas.	https://CRAN.R-project.org/package=gridExtra
Rfm	Segmentação de clientes a partir do método RFV (Recência, Frequência e Valor). Gera pontuação RFM a partir de dados de nível de transação e cliente.	https://CRAN.R-project.org/package=rfm
Summarytools	Uma coleção de funções que resumem dados numéricos e categóricos de forma organizada e rápida.	https://CRAN.R-project.org/package=summarytools

Fonte: Elaborado pelos autores (2023)

DOI: <https://doi.org/10.21166/cpitt.v5i1.4918>. ISSN: 2763-5449

Para a análise dos dados, resultados desta pesquisa, foram utilizadas algumas das principais medidas resumo quem compõem a estatística descritiva. Favero e Belfiori (2021) afirmaram que este ramo da ciência estatística permite uma melhor compreensão de um conjunto de dados, descrevendo e sintetizando as características observadas, sem tirar quaisquer conclusões ou inferências sobre a população estudada.

RESULTADOS E DISCUSSÃO

O período equivalente a 395 (trezentos e noventa e cinco) dias, serviu como referência para a extração de um conjunto de dados organizados e transformados, construído a partir das tabelas de origem armazenadas no sistema gerenciador de bancos de dados transacional da empresa.

Foram 64.135 (sessenta e quatro mil cento e trinta e cinco) observações de vendas, convertidas, sumarizadas e agrupadas pela variável código_cliente através da instrução "*group_by*" do pacote "*dplyr*", dando surgimento a primeira ABT (*Analytical Base Table* ou Tabela Base Analítica) com 1.082 (mil e oitenta e duas) observações agrupadas por clientes, esta, um novo conjunto de dados formatado com as métricas de entrada para a realização da análise RFV (Recência, Frequência e Valor).

A Tabela 1 a seguir apresenta um recorte com as treze primeiras linhas do conjunto de dados estruturado com as variáveis código_cliente, recência, frequência, valor e última_compra.

Tabela 1 – Estrutura da ABT cliente, recência, frequência, valor e data da última compra.

código_cliente	recência	frequência	Valor (R$)	última_compra
000002:01	3	90	542,33	29/08/2022
000004:01	2	213	372,07	30/08/2022
000004:02	2	160	1.167,43	30/08/2022
000004:03	1	118	972,62	31/08/2022
000004:04	1	100	280,23	31/08/2022
000004:05	5	114	1.974,11	27/08/2022
000004:06	2	167	1.789,76	30/08/2022
000004:08	2	93	578,83	30/08/2022
000004:10	8	40	270,19	24/08/2022
000006:01	3	65	169,09	29/08/2022
000008:01	1	38	428,50	31/08/2022
000013:01	269	21	74,86	06/12/2021
000020:01	27	21	1.196,90	05/08/2022

Fonte: Elaborado pelos autores com dados originais da pesquisa (2023)

Nela observamos que o cliente nº 000002:01 comprou há 3 (três) dias e acumula um total de 90 (noventa) compras realizadas no período com um volume médio de transações no valor de R$ 542,33 (quinhentos e quarenta e dois reais e trinta e três centavos) e sua última compra realizada em 29/08/2022. O cliente nº 000004:01, por sua vez, comprou há 2 (dois) dias, somando 213 (duzentas e treze) compras no período com um volume médio de transações no valor de R$ 372,07 (trezentos e setenta e dois reais e sete centavos) e realizou a sua última compra em 30/08/2022.

A primeira ABT serviu de insumo para a realização da análise RFV (Recência, Frequência e Valor), neste etapa foram geradas as pontuações RFV, e a partir destas, foram extraídos os respectivos segmentos. Como já citado, a análise RFV

DOI: <https://doi.org/10.21166/cpitt.v5i1.4918>. ISSN: 2763-5449

(Recência, Frequência e Valor) é um método que usa três métricas para classificar os clientes de acordo com seu perfil de consumo, e tem como objetivo realizar ações com cada cliente conforme a sua classificação. Inicialmente, dividimos o conjunto de dados em quintis (cinco partes), e de acordo com a distribuição dos valores de recência, frequência e valor monetário, atribuímos a cada quintil, de cada uma das métricas, uma pontuação de 1 a 5. As cinco categorias em três variáveis poderiam criar até 125 (5x5x5) pontuações RFV (Recência, Frequência e Valor) de clientes, após a aplicação do modelo da análise RFV (Recência, Frequência e Valor) foram gerados 109 (cento e nove) pontuações diferentes de clientes.

A Tabela 2 a seguir apresenta um recorte com as treze primeiras linhas do conjunto de dados com as pontuações, resultados da análise RFV (Recência, Frequência e Valor). Nesta segunda ABT manteve-se a estrutura da primeira, com as variáveis código_cliente, recência, frequência, valor, e foram adicionadas a estas, as variáveis R_score, F_score, V_score, correspondentes pontuações individuais de cada métrica, e RFV_score que concatena em um único valor as três pontuações individuais.

Tabela 2 – Estrutura da ABT cliente, recência, frequência, valor e data da última compra.

código_cliente	recência	frequência	Valor (R$)	R_score	F_score	V_score	RFV_score
000002:01	3	90	542,33	4	5	4	454
000004:01	2	213	372,07	5	5	4	554
000004:02	2	160	1.167,43	5	5	5	555
000004:03	1	118	972,62	5	5	5	555
000004:04	1	100	280,23	5	5	3	553
000004:05	5	114	1.974,11	3	5	5	355
000004:06	2	167	1.789,76	5	5	5	555
000004:08	2	93	578,83	5	5	4	554
000004:10	8	40	270,19	3	3	3	333
000006:01	3	65	169,09	4	5	2	452
000008:01	1	38	428,50	5	3	4	534
000013:01	269	21	74,86	1	2	1	121
000020:01	27	21	1.196,90	2	2	5	225

Fonte: Elaborado pelos autores com dados originais da pesquisa (2023)

Nela observamos que o cliente nº 000002:01 obteve a pontuação quatro para a recência, cinco para a frequência e quatro para o valor, tendo a sua pontuação geral ou RFV_score de 454. O cliente nº 000004:01, por sua vez, obteve a pontuação cinco para a recência, cinco para a frequência, e quatro para o valor, atingindo com isso um RFV_*score* de 554.

É importante destacar que ao longo desta pesquisa, foram encontradas na literatura sobre o tema, diversas maneiras para realizar uma análise RFV (Recência, Frequência e Valor). O que nos fez pensar que, não existe uma receita única e correta quando o assunto é esse tipo de análise. E que, é possível encontrar diferentes abordagens, cada uma com os seus respectivos pontos positivos e negativos. As etapas abaixo explicam como a pontuação RFV (Recência, Frequência e Valor) foi calculada para cada cliente neste trabalho, elas foram fundamentadas na metodologia disponível no pacote “rfm”, que divide em 5 categorias cada uma das variáveis, como seguem:

A pontuação de recência foi atribuída a cada cliente com base no valor de

DOI: <https://doi.org/10.21166/cpitt.v5i1.4918>. ISSN: 2763-5449

referência da variável recência. A pontuação foi gerada agrupando os valores de recência em 5 categorias numa escala de valores categóricos que vão de 1 a 5. Ou seja, clientes com baixa recência receberam a pontuação mais alta 5 e aqueles com alta recência, receberam a pontuação mais baixa 1.

A Tabela 3 apresenta os limites de valores mínimos e máximos de recência e sua respectiva categoria.

Tabela 3 – Escala e limites de valores para a pontuação da recência.

Categorias	Mínimo (em dias)	Máximo (em dias)
5	1	3
4	3	4
3	4	11
2	11	155
1	155	396

Fonte: Elaborado pelos autores com dados originais da pesquisa (2023)

A pontuação da frequência foi atribuída de maneira semelhante, clientes com alta frequência de compras receberam pontuação mais alta ou 5, e aqueles com frequências mais baixas receberam a pontuação 1.

A Tabela 4 apresenta os limites de valores mínimos e máximos de frequência e sua respectiva categoria.

Tabela 4 – Escala e limites de valores para a pontuação da frequência.

Categorias	Mínimo (em transações)	Máximo (em transações)
1	1	12
2	12	31
3	31	50
4	50	62
5	62	289

Fonte: Elaborado pelos autores com dados originais da pesquisa (2023)

A pontuação do valor foi atribuída com base no ticket médio de compras gerada pelo cliente no período supracitado, os clientes com maior valor de ticket médio receberam uma pontuação mais alta, enquanto aqueles com menor ticket médio receberam uma pontuação de 1.

A Tabela 5 apresenta os limites de valores mínimos e máximos de ticket médio e sua respectiva categoria.

Tabela 5 – Escala e limites de valores para a pontuação do valor.

Categorias	Mínimo (em R$)	Máximo (em R$)
1	39,00	141,00
2	141,00	177,09
3	177,09	331,16
4	331,16	758,11
5	758,11	14.404,16

Fonte: Elaborado pelos autores com dados originais da pesquisa (2023)

Para exemplificar, utilizando a Tabela 1 como a referência, nela observamos que o cliente nº 000013:01 comprou pela última vez há 269 dias, obteve a pontuação 1 para a recência, que na sua escala e limites de valores para a pontuação desta variável, está entre o período de 155 e 396 dias. Este mesmo cliente realizou 21

DOI: <https://doi.org/10.21166/cpitt.v5i1.4918>. ISSN: 2763-5449

transações, obteve a pontuação 2 para frequência, que na escala e limites de valores para a pontuação desta variável, está entre 12 e 31. Finalmente, gastou o equivalente a R$ 74,86 (setenta e quatro reais e oitenta e seis centavos), obteve a pontuação 1 para valor, que na escala e limites de valores para a pontuação desta variável, está entre R$ 39,00 (trinta e nove reais) e R$ 141,00 (cento e quarenta e um reais).

Em seguida, com base na pontuação individual apurada para a recência, frequência e valores, os clientes foram enquadrados em seus respectivos segmentos. Na Tabela 6 são apresentados os nomes dos segmentos, as suas descrições e as escalas de pontuação para R, F e V, que serviram de referência para aplicar o modelo e delimitar a segmentação, inserindo cada cliente em seu segmento baseado em sua correspondente pontuação.

Tabela 6 – Segmentação de clientes por recência, frequência e pontuações monetárias

Segmento	Descrição	R	F	V
Campeões	Clientes que compraram mais recentemente, com mais frequência e que mais gastaram	4 - 5	4 - 5	4 - 5
Clientes fiéis	Gastou um bom dinheiro	2 - 5	3 - 5	3 - 5
Potencial leal	Clientes recentes, gastaram uma boa quantia, compraram mais de uma vez	3 - 5	1 - 3	1 - 3
Novos clientes	Comprou mais recentemente, mas não com frequência	4 - 5	<=1	<=1
Promissor	Compradores recentes, mas não gastaram muito	3 - 4	<= 1	<= 1
Precisa atenção	Recência, frequência e valores monetários acima da média	2 - 3	2 - 3	2 - 3
Prestes a perder	Recência, frequência e valores monetários abaixo da média	2 - 3	<= 2	<= 2
Em risco	Gastou muito dinheiro, comprou com frequência, mas há muito tempo	<= 2	2 - 5	2 - 5
Não posso perdê-los	Fez grandes compras e muitas vezes, mas há muito tempo	<= 1	4 - 5	4 - 5
Perdido	Menor recência, frequência e pontuações monetárias	<= 2	<= 2	<= 2

Fonte: Adaptado de Hebbali (2022)

Em forma de resumo, o resultado da segmentação pode ser observado através da Tabela 7 a seguir onde são apresentadas as descrições dos segmentos, a quantidade de clientes por segmento, a quantidade de pedidos feitos pelos clientes do segmento e o valor médio total do segmento.

Tabela 7 – Segmentação de clientes por recência, frequência e pontuações monetárias

Segmento	Nº de Clientes	Nº de Pedidos	Valor Total (R$)
Campeões	209	16763	242.096,82
Clientes fiéis	231	13463	175.954,97
Potencial leal	211	6827	136.513,67
Perdido	120	710	13.642,29
Em risco	70	1797	59.890,24
Promissor	144	4578	20.618,16
Precisa de atenção	44	1110	8.012,06
Prestes a perder	53	664	5.748,66

Fonte: Elaborado pelos autores com dados da pesquisa, a partir de Hebbali (2022)

Como se vê na Tabela 7, os segmentos: “não posso perdê-los” e “novos

DOI: <https://doi.org/10.21166/cpitt.v5i1.4918>. ISSN: 2763-5449

clientes" não tiveram observações dentro das escalas de pontuação para R, F e V e consequentemente não estavam presentes no resultado.

Numa visão detalhada do resultado final da análise RFV (Recência, Frequência e Valor), a Tabela 8 a seguir exibe um recorte com as treze primeiras linhas da terceira ABT estruturada com o código_cliente, as suas medidas originais de transação RFV (Recência, Frequência e Valor), a pontuação RFV (Recência, Frequência e Valor) auferida na amostra de dados e qual o segmento foi inserido o cliente a partir da análise.

Tabela 8 – Resultado da segmentação dos clientes da empresa analisada.

código_cliente	recência	frequência	Valor(R$)	RFV_score	segmento
000002:01	3	90	542,33	454	Campeões
000004:01	2	213	372,07	554	Campeões
000004:02	2	160	1.167,43	555	Campeões
000004:03	1	118	972,62	555	Campeões
000004:04	1	100	280,23	553	Clientes fiéis
000004:05	5	114	1.974,11	355	Clientes fiéis
000004:06	2	167	1.789,76	555	Campeões
000004:08	2	93	578,83	554	Campeões
000004:10	8	40	270,18	333	Clientes fiéis
000006:01	3	65	169,09	452	Outros
000008:01	1	38	428,50	534	Clientes fiéis
000013:01	269	21	74,86	121	Perdido
000020:01	27	21	1.196,90	225	Em risco

Fonte: Elaborado pelo autor com dados da pesquisa (2023)

Uma vez que os clientes estejam classificados em segmentados específicos, pode-se identificar os melhores e de alto potencial, tomar ações apropriadas para melhorar a experiência e as ofertas direcionadas para cada segmento. Oferecendo mais clareza sobre o que esperar de suas interações futuras, considerando para isso a pontuação RFV (Recência, Frequência e Valor) atualizada dos clientes.

Apurada a pontuação RFV (Recência, Frequência e Valor), definidos os oito segmentos da carteira de clientes e sob o argumento de que seria menos oneroso direcionar recursos no planejamento de ações direcionadas para uma quantidade menor de segmentos, surgiu um questionamento por parte da empresa objeto desta pesquisa, seria possível reduzir o número de segmentos pela metade?

Tendo o número quatro como referência para reduzir ainda mais a dimensionalidade do conjunto de dados e sob o argumento de que o custo para ações direcionadas em 4 grupos é menor e ainda cabe no orçamento de despesas da área comercial do período.

Após um aprofundamento teórico nas técnicas exploratórias aplicadas quando há a intenção de verificar a existência de comportamentos semelhantes entre observações, identificou-se que, a análise de agrupamentos ou 'clusters' é um conjunto de técnicas exploratórias do tipo não supervisionado, que tem por objetivo ordenar e alocar as observações de um conjunto de dados em grupos. Esta técnica agrupa um conjunto de dados de clientes para que, os clientes dentro dos seus grupos sejam semelhantes entre si e coletivamente diferentes dos clientes em outros grupos Fávero e Belfiore (2021).

A semelhança está em termos das variáveis de agrupamento, que podem ser psicográficos, demográficos ou medidas de transação, como a análise de RFV (Recência, Frequência e Valor monetário). Os agrupamentos geralmente têm

DOI: <https://doi.org/10.21166/cpitt.v5i1.4918>. ISSN: 2763-5449

interpretações ricas com fortes implicações para as quais os clientes devem ser direcionados com uma oferta específica ou comercializados em um determinado contexto (BLATTEBERGER et al., 2008).

Fávero e Belfiore (2021) ainda apresentam que para analisar, interpretar e comparar os resultados na aplicação de uma análise de agrupamentos é necessário previamente escolher determinada medida de distância (dissimilaridade) ou semelhança (similaridade), esta escolha, servirá de base para que as observações sejam consideradas menos ou mais próximas, e determinado esquema de aglomeração, que deverá ser definido entre os métodos hierárquico e não hierárquicos.

Para este trabalho, foi estabelecido como esquema de aglomeração o procedimento '*k-means*' ou k-médias, um dos métodos não hierárquicos de agrupamento de dados que se refere a processos em que são definidos os centros de aglomeração e a partir dos quais são alocadas as observações pela proximidade a eles. Devido à simplicidade do algoritmo e à velocidade de seleção do centro do agrupamento (centroide), o procedimento é um dos métodos de agrupamento mais populares. O procedimento 'k-means' ou k-médias, geralmente aplica a fórmula de distância euclidiana para determinar a medida de distância dos dados em um grupo (FÁVERO; BELFIORE, 2021).

Diante do contexto, o conjunto de dados obtido no resultado da primeira ABT, Quadro 1, foi submetido previamente a uma análise exploratória para a sua aplicação no modelo de análise de agrupamentos. Inicialmente foram selecionadas as variáveis com as métricas de entrada para a realização de uma análise de agrupamentos, foram mantidas apenas as variáveis do tipo numéricas: recência, frequência e valor, as demais, foram removidas por não serem necessárias ao objetivo e/ou do tipo categóricas, estas, incompatíveis na condição atual com os requisitos do algoritmo '*k-means*' ou k-médias.

A Tabela 9 apresenta um recorte com as treze primeiras linhas da quarta ABT, nela, constam as variáveis de interesse para aplicação do modelo de análise de agrupamentos.

Tabela 9 – Variáveis originais RFV (Recência, Frequência e Valor).

Recência	Frequência	Valor (em R$)
3	90	542,33
2	213	372,07
2	160	1.167,43
1	118	972,62
1	100	280,23
5	114	1.974,11
2	167	1.789,76
2	93	578,83
8	40	270,19
3	65	169,09
1	38	428,50
269	21	74,86
27	21	1.196,90

Fonte: Elaborado pelos autores com dados originais da pesquisa (2023)

Em seguida, foi aplicado um procedimento para a padronização das escalas, uma vez que, as variáveis selecionadas se apresentaram em escalas de medidas distintas. Recência é medida em dias, frequência em número de transações e valor

DOI: <https://doi.org/10.21166/cpitt.v5i1.4918>. ISSN: 2763-5449

em moeda corrente. Este procedimento conhecido como *Zscores* transforma a variável original em uma nova variável com média zero e variância 1 e sem alterar a sua forma de distribuição (FÁVERO e BELFIORE, 2021). De forma que, na Tabela 10 pode-se observar um recorte com as treze primeiras linhas da quinta ABT, esta, pronta para a aplicação do modelo de análise de agrupamentos.

Tabela 10 – Variáveis RFV (Recência, Frequência e Valor) padronizadas**.**

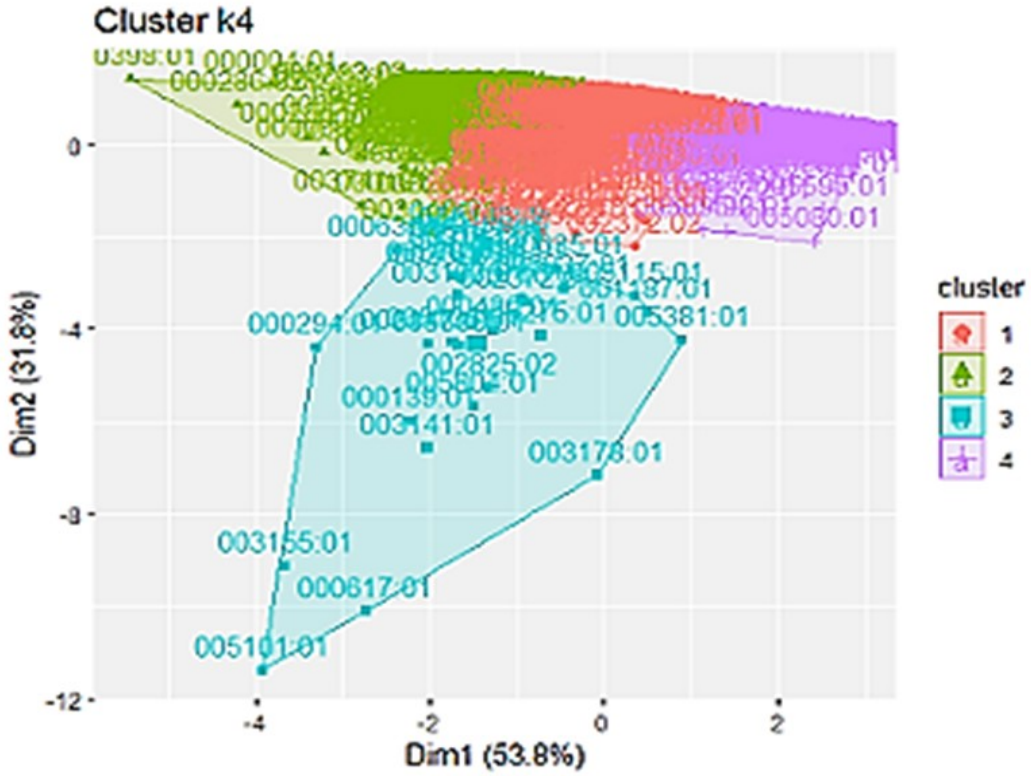
Recência	Frequência	Valor
-0.59	1.42	-0.06
-0.60	5.08	-0.21
-0.60	3.50	0.48
-0.61	2.25	0.31
-0.61	1.72	-0.29
-0.57	2.13	1.19
-0.60	3.71	1.03
-0.60	1.51	-0.03
-0.55	-0.07	-0.30
-0.59	0.67	-0.39
-0.61	-0.13	-0.16
1.78	-0.64	-0.47
-0.38	-0.64	0.51

Fonte: Elaborado pelos autores com dados originais da pesquisa (2023)

Executado o modelo de agrupamentos através do procedimento *k-means*, onde “k” foi o equivalente ao número de grupos, previamente definido como quatro.

Na Figura 1 é possível verificar de maneira visual o resultado do agrupamento dos clientes e de que forma esses grupos foram formados

Figura 1 – Visualização do resultado do agrupamento dos clientes da empresa analisada.



Fonte: Elaborado pelos autores com dados originais da pesquisa (2023).

DOI: <https://doi.org/10.21166/cpitt.v5i1.4918>. ISSN: 2763-5449

Uma vez estabelecidos os agrupamentos dos clientes, o seu resultado é um novo conjunto de dados, este, apresentado na Tabela 11, e nela, pôde-se observar um recorte da sexta ABT com a variável indexadora código_cliente e o seu respectivo grupo.

Tabela 11 – Agrupamentos dos clientes**.**

código_cliente	grupo
000002:01	2
000004:01	2
000004:02	2
000004:03	2
000004:04	2
000004:05	2
000004:06	2
000004:08	2
000004:10	1
000006:01	1
000008:01	1
000013:01	4
000020:01	1

Fonte: Elaborado pelos autores com dados originais da pesquisa (2023)

Finalizando, foram unificados em um único conjunto de dados os resultados da Tabela 11 com os da Tabela 9, dando origem a Tabela 12, nela, há um recorte com as treze primeiras linhas desta sétima ABT estruturada com as variáveis código_cliente, recência, frequência, valor, RFV_score, segmento e grupo

Tabela 12 – Resultado da análise segmentada combinada com a análise de agrupamentos.

código_cliente	recência	frequência	Valor(R$)	RFV_score	segmento	grupo
000002:01	3	90	542,33	454	Campeões	2
000004:01	2	213	372,07	554	Campeões	2
000004:02	2	160	1.167,43	555	Campeões	2
000004:03	1	118	972,62	555	Campeões	2
000004:04	1	100	280,23	553	Clientes fiéis	2
000004:05	5	114	1.974,11	355	Clientes fiéis	2
000004:06	2	167	1.789,76	555	Campeões	2
000004:08	2	93	578,83	554	Campeões	2
000004:10	8	40	270,18	333	Clientes fiéis	1
000006:01	3	65	169,09	452	Outros	1
000008:01	1	38	428,50	534	Clientes fiéis	1
000013:01	269	21	74,86	121	Perdido	4
000020:01	27	21	1.196,90	225	Em risco	1

Fonte: Elaborado pelo autor com dados da pesquisa (2023)

A sétima ABT, serviu de insumo para a realização de uma análise descritiva dos dados, afinal, o seu resultado indicou apenas a qual grupo pertence cada cliente e essa informação isoladamente não permitiu uma caracterização detalhada dos grupos, informações relevantes do tipo, qual dos grupos tem melhor desempenho monetário ou qual deles possui os clientes menos rentáveis, dentre outros indicadores. Sendo assim, fez-se necessário uma análise individual usando as medidas resumo da estatística sob o prisma das variáveis recência, frequência e

DOI: <https://doi.org/10.21166/cpitt.v5i1.4918>. ISSN: 2763-5449

valor, conforme apresentado na Tabela 13.

Tabela 13 – Resumo dos grupos medidos pela média aritmética simples.

grupo	clientes	recência	frequência	valor(R$)
1	657	16,8	40,4	458,00
2	153	3,57	104	699,00
3	32	34,6	43,50	5.831,00
4	240	260	8,74	284,00

Fonte: Elaborado pelos autores com dados originais da pesquisa (2023)

Na Tabela 13, observa-se que o grupo 1 é formado por 60,72% dos clientes, é o grupo com a maior quantidade de clientes. Ele é composto pelos clientes que, em média, compraram há 16,8 dias, realizaram 40,4 transações de compras e gastaram R$ 458,00.

De maneira semelhante, o grupo 2 possui a melhor recência 3,57 dias, é formado por 153 clientes, o equivalente a 14,14% do todo, estes, realizaram a maior número de compras, foram 104 e obtiveram o segundo melhor valor monetário R$ 699,00.

Por sua vez, o grupo 3 é o menor e o mais importante dos grupos, reuniu o equivalente a 2.96% dos clientes, foram 32 no total. Esse grupo possui os clientes que mais gastaram, foram R$ 5.831,00, estes, diretamente proporcionais a sua frequência de 43,50 compras no período, com a recência de 34,6 dias.

Por fim, o grupo 4 é o que mais precisa de atenção, foi o segundo em número de clientes, acumulou 240 no total, o equivalente a 22,18%, no entanto, obteve a pior recência, a pior frequência e o pior valor monetário.

CONSIDERAÇÕES FINAIS

Este trabalho teve como objetivo apresentar pesquisa voltada a implementar algoritmos de ‘*Machine Learning’* para a segmentação e classificação de clientes, orientados pelas variáveis recência, frequência e valor, que formam o modelo RFV (Recência, Frequência e Valor), bastante utilizado em ‘marketing’ para oportunizar novos negócios e a melhoria contínua do relacionamento com cliente.

Foi utilizada a combinação de duas técnicas para esta finalidade, na primeira, a análise RFV (Recência, Frequência e Valor) reduziu a dimensionalidade das 64.135 (sessenta e quatro mil cento e trinta e cinco) observações para 1.082 (mil e oitenta e duas), estas, foram agrupadas por clientes, em seguida, foram reduzidas para 8 (oito) segmentos com características distintas.

Na segunda técnica, a análise de agrupamentos, emergiu de um questionamento feito pela empresa objeto desta pesquisa, que diante das suas limitações de recursos, demandou através das mesmas variáveis a redução para 4 (quatro) grupos, estes, criados sob a premissa de que seus membros fossem semelhantes dentro de si e coletivamente diferentes dos membros em outros grupos, permitindo a compreensão e conhecimento dos clientes a partir da análise dos seus históricos de compras e, desta forma, poder traçar estratégias mais assertivas e direcionadas ao perfil de cada grupo de consumidores.

Trabalhos futuros poderão aproveitar o ‘*output’* desta pesquisa, compatibilizar as variáveis categóricas, através de uma terceira técnica, a análise de correspondência múltipla, capturando as respectivas coordenadas, estas, variáveis numéricas e adicioná-las para um novo processamento da análise de agrupamentos

para investigar qual das técnicas mais se aproxima da solução ideal para o problema de negócio.

REFERÊNCIAS

ABAD. Associação Brasileira de Atacadistas e Distribuidores de Produtos Industrializados. **Modelos de Negócio**. Disponível em: <https://abad.com.br/servicos/dados-do-setor/modelos-de-negocio/>. Acesso em: 17 abr. 2022.

BLATTEBERGER, ROBERT C.; KIM, BYUNG D.; NESLIN, SCOTT A. **Database Marketing Analyzing and Managing Customers**. New York: Springer, 2008.

FÁVERO, L. P.; BELFIORE, P. 2021. **Manual de análise de dados**: estatística e modelagem multivariada com Excel, SPSS e Stata. 1ed. Elsevier, Rio de Janeiro, RJ, Brasil.

HAIR, J. F., AANDERSON, R. E., TATHAM, R. L., BLACK, W. C. **Análise multivariada de dados**. Porto Alegre: 6ed. Bookman, 2009.

HEBBALI, A. **rfm: Recency, Frequency and Monetary Value Analysis**. 2022. Disponível em: <https://github.com/rsquaredacademy/rfm, https://rfm.rsquaredacademy.com/>. Acesso em 18 jul. 2022.

DICIO, Dicionário Online de Português. **Marketing.** Porto: 7Graus, 2020. Disponível em: <https://www.dicio.com.br/marketing/>. Acesso em 17 abr. 2022.

MORETTIN, P. A.; SINGER, J. M. **Introdução à Ciência de Dados**: fundamentos e Aplicações. São Paulo: USP, 2020. Disponível em: <https://www.ime.usp.br/~jmsinger/MAE0217/cdados2020mai14.pdf>. Acesso em 02 ago. 2022.

NORMAN, Alan T. **Aprendizagem de Máquina Em Ação**: Uma Obra Para o Leigo, Guia Passo a Passo Para Novatos. eBook: Tektime, 2019.

PINHO, Anderson Guimarães de. Análise RFV do Cliente por Algoritmos Genéticos na Otimização de Estratégias de Marketing. In: **Revista Pensamento Contemporâneo em Administração**. Rio de Janeiro, 2009.

R CORE TEAM. **R: A language and environment for statistical computing**. 2021. R Foundation for Statistical Computing, Vienna, Áustria. Disponível em <https://www.R-project.org/>. Acesso em: 17 abril 2022.

RSTUDIO TEAM. **RStudio**: Integrated Development for R. RStudio. 2022. PBC, Boston, MA Disponível em <http://www.rstudio.com/>. Acesso em: 17 abril 2022.

SWIFT, Ronald. **CRM, customer relationship management**: O revolucionário marketing de relacionamentos com o cliente. Rio de Janeiro: Elsevier, 2001.

WHEELAN, C. Estatística: o que é, para que serve, como funciona. 1 ed. Zahar, Rio de Janeiro. 2016.

DOI: <https://doi.org/10.21166/cpitt.v5i1.4918>. ISSN: 2763-5449

ZENONE, Luiz Claudio. **CRM (Customer Relationship Management)**: marketing de relacionamento, fidelização de clientes e pós-venda. São Paulo: Almedina, 2019.

SOBRE OS AUTORES

Luiz Henrique Batista de Santana. (*Especialista em Data Science e Analytics, ESALQ-USP, 2022*), henriquemago@hotmail.com.

Domingos Sávio da Cunha Garcia. (*Doutor em Histórica Econômica, Unicamp, 2005)*. Professor da Universidade do Estado de Mato Grosso, Cáceres, domingos.garcia@unemat.br.

Leonardo Amorim de Araújo. (*Doutor em Engenharia de Transportes, UFRJ, 2003)*. Professor do Instituto Federal Sudeste de Minas Gerais, Santos Dumont, leonardo.araujo@ifsudestemg.edu.br.

Carlos Artur Alevato Leal. (*Doutor em Engenharia Mecânica, UFMG, 2020)*. Professor do Instituto Federal Sudeste de Minas Gerais, Santos Dumont, artur.leal@ifsudestemg.edu.br.

Arthur Nascimento Assunção. (*Mestre em Ciência da Computação, UFOP, 2016)*. Professor do Instituto Federal Sudeste de Minas Gerais, Santos Dumont, arthur.assuncao@ifsudestemg.edu.br.

Gustavo José Santiago Rosseti. (*Doutor em Engenharia Elétrica, UFJF, 2015)*. Professor do Instituto Federal Sudeste de Minas Gerais, Santos Dumont, gustavo.rosseti@ifsudestemg.edu.br.

Silvana Rodrigues Pires Moreira. (*Doutora em Bioquímica Agrícola, UFV, 2013)*. Professora do Instituto Federal Sudeste de Minas Gerais, Santos Dumont, silvana.moreira@ifsudestemg.edu.br.

Lisleandra Machado. (*Doutora em Engenharia de Produção, UNIMEP, 2022)*. Coordenadora do Curso de Engenharia Ferroviária e Metroviária, Professora do Instituto Federal Sudeste de Minas Gerais, Santos Dumont, lisleandra.machado@ifsudestemg.edu.br.