CURSO DE CIÊNCIA DA COMPUTAÇÃO – TCC	
(X) PRÉ-PROJETO ( ) PROJETO	ANO/SEMESTRE: 2021/2

UTILIZAÇÃO DE CLUSTERIZAÇÃO PARA AUXÍLIO EM TOMADA DE DECISÃO A PARTIR DE DADOS DE VAREJO

Henrique José Wilbert

Prof. Aurélio Faustino Hoppe – Orientador

Prof. Christian Daniel Falaster – Coorientador

1 INTRODUÇÃO

Com a evolução da tecnologia de informação a partir dos anos 80 e início dos anos 90, várias grandes empresas adotaram sistemas de gerenciamento na forma de softwares Enterprise Resource Planning (ERP) (RASHID; HOSSAIN; PATRICK, 2001, p.2). Estes softwares auxiliam em suas rotinas à nível operacional, seja no controle do estoque, fiscal, financeiro, transacional e até recursos humanos. A partir disso, alcançou-se um patamar de eficiência nunca concebido, visto que registros antes realizados em papel e caneta, passaram a ser produzidos automaticamente. Ainda segundo os autores, em paralelo a informatização destes processos, houve também um crescimento da quantidade de dados armazenados referentes à produtos, clientes, transações, gastos e receitas.

Diante deste contexto, avançaram-se também as táticas de marketing direto, como por exemplo, o envio de catálogos por correio, até ofertas altamente objetivas, focando em indivíduos selecionados, cujas informações transacionais estavam presentes na base de dados. Percebeu-se que o foco das relações empresa-cliente não está em clientes novos, e sim em clientes já existentes nas bases de dados, visto que o custo para adquirir um cliente novo através de publicidade é muito maior que o custo de alimentar uma relação já existente (PETRISON; BLATTBERG; WANG, 1997, p. 119, tradução nossa).

Segundo Reinartz, Thomas e Kumar (2005, p.77), quando empresas tratam os gastos entre aquisição e retenção de clientes, destinar menos recursos para a retenção impactará em uma lucratividade menor à longo prazo, comparando-se a investimentos menores em aquisição de clientes. Ainda segundo os autores, no conceito de relações de retenção, atribui-se grande ênfase à lealdade e lucratividade de um cliente, sendo lealdade a tendência do cliente comprar e a lucratividade, a medida geral de quanto lucro um cliente traz à empresa através de suas compras.

De acordo Nguyen, Sherif e Newby (2007, p.114) com o avanço da gerência das relações com clientes foram abertas novas vias pelas quais sua lealdade e lucratividade pode ser cultivada, atraindo uma crescente demanda por parte de empresas, visto que a adoção destes meios permite que as organizações melhorarem seu serviço ao consumidor, consequentemente gerando renda. Com isso, diferentes ferramentas acabam sendo utilizadas, como sistemas de recomendação que, geralmente em ramos e-commerce, levam em conta várias características pertinentes ao comportamento do cliente, construindo um perfil próprio que será utilizado para realizar a recomendação de um produto que talvez seja de seu interesse. Outra ferramenta pertinente à lucros e lealdade é a segmentação, que visa separar uma única e confusa massa de clientes em segmentos homogêneos em termos de comportamento, permitindo o desenvolvimento de campanhas e estratégias de marketing especializadas à cada grupo de acordo com suas características (TSIPTSIS; CHORIANOPOULOS, 2009, p.4).

Em relação a segmentação de clientes, algumas métricas tornam-se relevantes nos contextos aos quais estão inseridas. Segundo Kumar (2008, p.29), o modelo Recency-Frequency-Monetary (RFM), é utilizado em empresas de venda por catálogo, enquanto empresas de high-tech tendem a usar Share of Wallet (SOW) para implementar suas estratégias de marketing. Já o modelo Past Customer Value (PCV), geralmente é utilizado em empresas de serviços financeiros. Dentre os modelos citados, o RFM é o que possui maior facilidade de aplicação em diversas áreas de comércio, varejo e supermercados, visto que são necessários apenas os dados transacionais dos clientes (vendas), dos quais são obtidos os atributos de recência (R), frequência (F) e monetário (M).

A partir desses dados, segundo Tsiptsis e Chorianopoulos (2009, p.335), é possível detectar bons clientes a partir das melhores pontuações de RFM. Se o cliente efetuou uma compra recentemente, seu atributo R será alto. Caso ele compre muitas vezes ao longo de um determinado período, seu atributo F será maior. E, por fim, caso seus gastos totais forem significativos, terá um atributo M alto. Ao categorizar o cliente dentro destas três características, é possível obter uma hierarquia de importância, tendo os clientes que possuem valores RFM altos no topo, e clientes que possuem valores baixos na base. Apesar destas vantagens, o modelo padrão original é um tanto quanto arbitrário, segmentando os clientes em quintis, cinco grupos com 20% dos clientes, não atentando-se

às nuances e todas a interpretações que a base de clientes pode possuir. Além disso, o método também pode produzir uma grande quantidade de grupos, que por muitas vezes, não representam significativamente os clientes de um estabelecimento, e caso o método de quintis seja utilizado, 125 grupos serão criados. Outro ponto a se observar é a variada gama de interpretações que os atributos RFM podem ter em relação aos tipos de atividades dos estabelecimentos, sendo necessário a adaptação do modelo para cada empresa.

Diante deste cenário, este trabalho propõe a criação de um artefato computacional que utilize o modelo RFM em conjunto com diferentes algoritmos de clusterização ao invés de quintis para segmentar clientes, extraindo de maneira automática as informações de bases de dados, sendo aplicado ao contexto de diversas empresas de varejo, atacado e comércio, visando adequar-se dinamicamente à suas eventuais diferenças de comportamento nos clientes.

1.1 OBJETIVOS

O objetivo deste trabalho é desenvolver um artefato computacional de auxílio à segmentação de clientes a partir de múltiplas bases de dados utilizando o modelo RFM.

Os objetivos específicos são:

a) implementar e testar diferentes algoritmos de clusterização;
b) disponibilizar um mecanismo de visualização dos agrupamentos;
c) avaliar a qualidade em relação ao agrupamento dos clientes.

2 TRABALHOS CORRELATOS

Neste capítulo serão apresentados os trabalhos similares ao proposto. Na seção 2.1 é apresentado o trabalho de Gustriansyah, Suhandi e Antony (2020), que consiste na aplicação do modelo *Recency Frequency Monetary* (RFM) para clusterização de produtos utilizando K-means, tendo como foco otimizar o número de clusters através de índices de validação. A seção 2.2 descreve o modelo de segmentação denominado *Length, Recency, Frequency, Monetary and Periodicity* (LRFMP)proposto por Peker, Kocyigit e Eren (2017), ao qual considera a longevidade e a periodicidade. Por final, na seção 2.3 detalha-se o modelo R+FM, sendo uma versão modificada do modelo original que foi aplicada em clientes de uma empresa de e-commerce (TAVAKOLI *et al*., 2018).

2.1 CLUSTERING OPTIMIZATION IN RFM ANALYSIS BASED ON K-MEANS

Gustriansyah, Suhandi e Antony (2020) utilizaram o algoritmo de clusterização K-means para agrupar 2.043 produtos de uma farmácia visando otimizar o manuseio de estoque. Foram utilizadas três características de acordo com o modelo *Recency Frequency Monetary* (RFM) para a separação dos produtos, levando em consideração dados transacionais capturados num período de um ano. O atributo recência classificou os produtos através da última venda realizada num intervalo de 1 a 364 dias. A frequência estabelece a quantidade de transações em que o produto ocorreu, variando num intervalo de 1 a 14.872. Já o atributo monetário, refere-se ao valor total proveniente das vendas acumuladas do produto, sendo definido num intervalo entre 1250 e 1.151.952.500 Rupias Indonésias (Rp.).

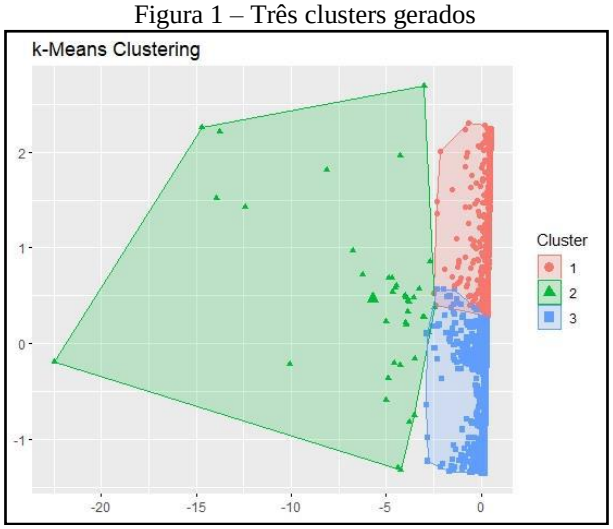
Após a atribuição de valores RFM aos produtos, foram utilizados oito índices de validação do melhor número de *clusters*: *Elbow Method* (EM), que calcula a variação intra-cluster conforme são aumentados os clusters e conclui que o melhor número é aquele que está no cotovelo (elbow) da curva. *Silhouette* Index (SI) que resulta em uma nota de -1 a 1 que indica a quão adequada é a classificação de um objeto dentro de um cluster em comparação aos outros. *Calinski-Harabasz Index* (CHI) que também mede a adequação da quantidade de clusters levando em conta a dispersão entre e intra clusters. *Davies-Bouldin Index* (DBI) que calcula as similaridades entre clusters levando em conta as distâncias e tamanhos dos clusters, quanto menor este índice melhor a separação entre os clusters. *Ratkowski Index* (RI) que é baseado na média da soma dos quadrados dos dados entre clusters e a soma total dos quadrados de cada dado dentro de um cluster, dentre as quantidades calculadas escolhe-se a que obtém um maior índice. *Hubert Index* (HI) que é um método visual que indica a quantidade preferida através de um pico no gráfico e é calculado pelo coeficiente de correlação entre matrizes de distância. *Ball-Hall Index* (BHI) definido pela média da distância dos itens com os respectivos centroides do cluster, onde no gráfico o ponto de quantidade de clusters com maior diferença do anterior é sugerido. *Krzanowski-Lai Index* (KLI), que propõe índices internos definidos pelas diferenças entre matrizes de dispersão, e aponta a melhor quantidade de clusters pelo maior número gerado ao realizar a equação com quantidade k. Constatou-se que a maioria deles indicou que o melhor número de clusters seria 3, com base nas condições de interpretação de cada índice explicadas anteriormente.

Segundo Gustriansyah, Suhandi e Antony (2020) nos testes para verificar os clusters gerados, utilizou-se a equação de variância (R). Sendo "R" o valor da divisão entre a distância média dos dados cluster (distância intra-

cluster) pela distância média dos dados em outros clusters (inter-cluster). O valor médio alcançado para R foi de 0.19113, sendo que quanto mais próximo de zero, maior a similaridade entre os membros dentro de cada cluster.

Gustriansyah, Suhandi e Antony (2020) utilizaram o software R Programming para gerar a clusterização, resultando na visualização demonstrada na Figura 1, tendo dois clusters com densidade maior e um cluster mais disperso. Também se observa que o cluster em verde possui uma maior variância entre os próprios dados, enquanto os outros dois clusters possuem uma menor diferença interna.

Figura 1 – Três clusters gerados



Fonte: Gustriansyah, Suhandi e Antony (2020).

Segundo Gustriansyah, Suhandi e Antony (2020) também foram adquiridos os valores médios RFM para cada cluster, conforme apresenta a Tabela 1, sendo que o cluster número 3 possui a maior média dos três atributos, e o cluster 1 possui a menor média dos três. É possível identificar um intervalo entre os valores médios de cada atributo, indicando uma diferença significativa inter-cluster.

Tabela 1 – Valores médios de RFM de cada cluster

Cluster	Recency	Frequency	Monetary (in thousands)
1	75.8167	3,436.744	3,089,608
2	224.3947	13,013.333	76,920,847
3	331.9681	107.418	286,927,000

Fonte: Gustriansyah, Suhandi e Antony (2020).

Gustriansyah, Suhandi e Antony (2020) concluem que o método gerou clusters com alta similaridade em relação aos dados existentes, apresentado uma segmentação mais objetiva quando comparado ao modelo RFM tradicional no qual os dados são divididos igualmente em cinco segmentos (20% dos dados para a cada segmento). Além disso, os autores também sugerem como extensões a utilização de outros métodos para a comparação, como *Particle Swarm Optimization* (PSO), que é um método computacional que otimiza soluções para uma equação de uma certa medida de qualidade, medioide (centroides que são parte do conjunto de dados) ou *maximizing-expectancy*, que é um método iterativo para encontrar estimativas de parâmetros para modelos estatísticos com variáveis não observadas.

2.2 LRFMP MODEL FOR CUSTOMER SEGMENTATION IN THE GROCERY RETAIL INDUSTRY: A CASE STUDY

Peker, Kocyigit e Eren (2017) propuseram o modelo *Length*, *Recency*, *Frequency*, *Monetary* (LRFM) denominado *Length*, *Recency*, *Frequency*, *Monetary and Periodicity* (LRFMP) para classificar dados reais de 16.024 clientes de mercados de uma franquia na Turquia. Para isso, utilizou-se o algoritmo K-means para segmentar os clientes e três índices de validação de clusters para a otimização das suas quantidades, *Silhouette* Index (SI), *Calinski-Harabasz Index* (CHI) e *Davies-Bouldin Index* (DBI). Após a segmentação dos dados, verificou-se estratégias de gerenciamento e relações com os clientes para aumentar a lucratividade, como tratamento preferencial para clientes importantes, implementação de cartões fidelidade para aumentar a frequência de compra de clientes não costumam comprar com frequência, promoções voltadas para clientes incertos com sua escolha de local de compra, dentre outras estratégias.

Primeiramente, Peker, Kocyigit e Eren (2017) adaptaram o modelo LRFM, incluindo o parâmetro *periodicity* (periodicidade), pois a análise dos dados foi realizada a partir do histórico de compras em supermercados, que são estabelecimentos com alto número de visitas, tornando importante a regularidade nos padrões de visita e compra. Peker, Kocyigit e Eren (2017) definem a periodicidade como a regularidade das visitas de um determinado cliente. Sendo atribuída como o desvio padrão dos tempos inter-visita do cliente (quantia de dias entre duas visitas consecutivas). Se um cliente possui valores baixos de periodicidade, significa que este realiza visitas ou compras em intervalos fixos, podendo caracterizá-lo como cliente regular. Além disso, os autores também modificaram o atributo de recência, transformando-o na média das diferenças entre a data das três últimas compras e a data atual, ao invés da simples diferença entre a data da última compra e a data atual estabelecida no modelo RFM padrão.

Após adquirir os atributos LRFMP dos dados transacionais dos clientes, Peker, Kocyigit e Eren (2017) aplicaram um método de normalização simples nos dados, considerando o intervalo de 0 e 1. Esta normalização foi feita pois os valores LRFMP variam em relação ao intervalo e escala, fato que poderia afetar negativamente a análise dos clusters.

Antes de aplicar a clusterização, Peker, Kocyigit e Eren (2017) utilizaram três índices para validação da quantidade possível de clusters: *Silhouette* Index (SI) que resulta em uma nota de -1 a 1 que indica o quão adequada é a classificação de um objeto dentro de um cluster em comparação aos outros, quanto maior o valor, melhor. *Calinski-Harabasz Index* (CHI) que mede a adequação da quantidade de clusters levando em conta a dispersão entre e intra clusters, um valor alto é preferido. *Davies-Bouldin Index* (DBI) que calcula as similaridades entre clusters levando em conta as distâncias e tamanhos dos clusters, quanto menor este índice melhor será a separação entre os clusters. A partir deles, Peker, Kocyigit e Eren (2017) executaram o algoritmo K-means variando o k de 2 a 9, e os resultados destas iterações foram avaliadas utilizando os três índices. Com base nos resultados, decidiu-se utilizar um número de 5 clusters, pois 2 dos 3 índices sugerem 5 como sendo a quantidade ideal.

Peker, Kocyigit e Eren (2017) utilizaram uma base de dados de uma franquia de mercados que possui mais de dez lojas na cidade de Antália na Turquia. Os dados são compostos por cerca de dois milhões de transações de 16.024 clientes num período de dois anos. Foram removidos os clientes com menos que três compras. Além disso, os autores removeram dados duplicados, transações com valores faltantes assim como, também agregaram as compras dentro de um mesmo dia. Depois dessas operações, a quantidade de clientes caiu para 10.471, sendo aplicado na sequência o K-means. A Tabela 2 demonstra a quantidade de clientes nos clusters, os valores médios de LRFMP para cada cluster. Já na última coluna, aplicou-se uma técnica no qual o atributo do cluster recebe uma seta para cima (↑) caso o seu valor for maior que a média do atributo dos outros clusters, e uma seta para baixo (↓), caso seu valor for menor que a média.

Tabela 2 – Valores médios dos clusters

Cluster	Sample size	Average L	Average R	Average F	Average M	Average P	LRFMP Scores
1	538	633.29	39.67	175.24	24.32	4.99	$L\uparrow R\downarrow F\uparrow M\downarrow P\downarrow$
2	4,681	564.50	90.19	33.44	31.73	31.49	$L\uparrow R\downarrow F\uparrow M\downarrow P\downarrow$
3	1,091	482.17	301.18	5.33	34.21	159.32	$L\uparrow R\uparrow F\downarrow M\downarrow P\uparrow$
4	818	374.01	220.24	11.85	104.18	45.81	$L\downarrow R\uparrow F\downarrow M\uparrow P\uparrow$
5	3,343	173.70	399.79	10.34	30.14	27.85	$L\downarrow R\uparrow F\downarrow M\downarrow P\downarrow$
Average		419.81	218.59	28.74	36.76	43.41	

Fonte: Peker, Kocyigit e Eren (2017).

A partir destes resultados, Peker, Kocyigit e Eren (2017) descreveram as características dos grupos. O grupo 1 representa clientes leais de alta contribuição que, apesar de comporem a menor parcela dos clientes (5,14%), possuem a maior contribuição total entre os grupos. Também é possível observar, que este grupo possui a menor periodicidade média de todos, caracterizando estes clientes como regulares. O grupo 2, representando a maior parcela dos clientes (44,70%) foi classificado como clientes leais de baixa contribuição pois apesar de visitar mais frequentemente as lojas, não possuem tanta contribuição quanto o grupo 1. O grupo 3, com tamanho de 10,42%, foi classificado como clientes incertos, pois possui o atributo de longevidade alto e recência também alta, significando que são clientes com longa história de compra, porém sem muitas compras recentes, vale notar que este grupo possui o maior valor de periodicidade de todos os grupos, caracterizando-o como um grupo de clientes sem rotina de compra definida. O grupo 4 e 5 foram classificados como clientes perdidos, visto que possuem poucas compras recentes, baixa frequência, e baixa longevidade, denotando um cliente que tem uma pouca interação com a franquia. O grupo 4, contendo uma pequena parcela de 7,81% dos clientes, gasta consideravelmente mais, logo foi classificado como contribuição alta, e o 5, cuja parcela é 31,93%, classificado como contribuição baixa.

A partir desta classificação, Peker, Kocyigit e Eren (2017) estabeleceram estratégias para cada grupo de clientes, como tratamento especial (vagas de estacionamento preferencial, presentes de aniversário, filas preferenciais) para clientes do grupo 1, de maneira a não perder a relação leal com a loja. Para os grupos 3, 4 e 5 cuja frequência é baixa, foi sugerida a adoção de programas de cartão fidelidade para aumentar a frequência deles. Para clientes incertos como no grupo 3, aplicou-se descontos e promoções, de maneira a incentivar os clientes, supostamente sensíveis aos preços, a voltar sua atenção à franquia. Para clientes perdidos, sugeriu-se uma análise mais profunda sobre o motivo da perda, como análise de feedback, inferência de motivos, dentre outros.

Por fim, Peker, Kocyigit e Eren (2017) concluem que o estudo contribuiu com a proposta de um novo modelo RFM, que possibilita uma análise mais profunda que o seu modelo original, visto que as características utilizadas e modificadas permitem uma melhor definição do comportamento de cada cliente. Outra contribuição foi a adição do atribuito periodicidade (P) no modelo que, ao contrário do modelo RFM padrão, permite identificar se os clientes de um grupo variam em sua rotina de compras. Outra melhoria apontada é a modificação do atributo de recência, que uma vez calculado como uma média, permite uma caracterização mais precisa que o atributo R comumente utilizado. Uma das limitações destacadas pelos autores é a localidade do estudo, pois foi realizado somente com dados originários de uma cidade, sendo que o comportamento de clientes pode variar de acordo com as diferentes localidades onde é feita a análise. A partir disso, sugere-se uma análise mais ampla contemplando outros locais. Outra sugestão feita por Peker, Kocyigit e Eren (2017) é a adição de novos atributos ao modelo, como a quantidade de produtos comprados, quantidade de produtos perecíveis e não perecíveis comprados, a fim de promover uma interpretação mais profunda do comportamento.

2.3 CUSTOMER SEGMENTATION AND STRATEGY DEVELOPMENT BASED ON USER BEHAVIOR ANALYSIS, RFM MODEL AND DATA MINING TECHNIQUES: A CASE STUDY

Tavakoli *et al*. (2018) desenvolveram o modelo RFM, denominado "R+FM", sendo utilizado em conjunto com o algoritmo de clusterização K-means para segmentar 3 milhões de clientes da maior empresa de *E-commerce* do Oriente Médio. Além disso, o modelo de segmentação foi comparado com o utilizado pela empresa, sendo aplicado em uma campanha de Short Message Service (SMS) focada em aumentar os ganhos de cada segmento.

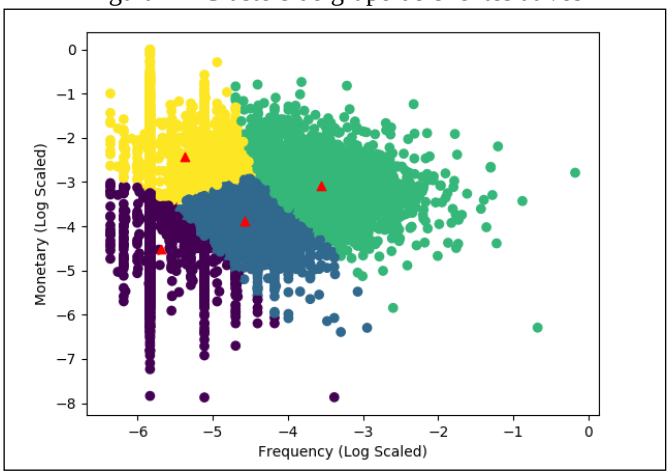
Tavakoli *et al*. (2018) defendem a utilização de um novo modelo de caracterização de clientes, argumentando que o modelo ideal necessita adaptar-se às mudanças de comportamento dos clientes, possuir certa independência de supervisão, levando em consideração a similaridade dos comportamentos dos clientes e a relação entre os atributos Frequência e Monetário. Para isso, construiu-se uma variante do modelo Recency, Monetary, Frequency (RFM) denominado de R+FM, que possui o atributo de recência separado dos demais, utilizando uma segmentação à parte do modelo FM. Os autores separaram os clientes em 3 grupos: os que compraram recentemente (cuja última compra foi dentro de 90 dias), denominados de ativos, clientes que compraram em um passado recente (cuja última compra foi entre 90 e 365 dias), denominados de expirando e por fim, os clientes que não compraram por um longo tempo (cuja última compra foi a mais de 365 dias), denominados de expirados. Para o atributo de frequência, Tavakoli *et al*. (2018) atentaram-se especialmente com a data da primeira compra, pois acreditam que a frequência tem uma importância maior conforme sua recência. Logo, definiram a frequência como a quantidade de compras dividida pela quantidade de dias desde a primeira compra, utilizando também uma função exponencial de decaimento, que efetivamente atribui um peso maior para anos mais recentes, sendo cada ano duas vezes mais pesado que o ano anterior. Como atributo monetário, estabeleceu-se a média dos valores das compras de um cliente, visto que um valor de soma total de compras, segundo os autores, estaria encorajando duas vezes os clientes.

Após a definição do modelo, Tavakoli *et al*. (2018), balancearam a relação entre frequência e monetário, criando a quarta característica que é definida pela combinação linear dos dois atributos, que nada mais é que a soma de cada atributo ponderada pelo peso de cada um. No tratamento dos dados, utilizou-se a técnica de remoção de *outliers* (clientes que não se encaixam no padrão normal) que não se encontram dentro dos intervalos interquartis, que são os intervalos que possuem os dados que pertencem à tendência média do conjunto de dados em geral. Também foram escalados os atributos de frequência e monetário para que seus intervalos sejam iguais, sendo aplicada a normalização *min-max*, que transforma os valores para estarem dentro do intervalo entre 1 e 0. Como os dados monetários e de frequência tratados possuem uma característica de cauda longa, fenômeno estatístico onde os dados são distribuídos de forma decrescente, foi aplicada uma transformação logarítmica para normalizar a distribuição, visto que a quantidade de valores baixos é muito alta, podendo atrapalhar a análise.

Como existem duas segmentações (R e FM), Tavakoli *et al*. (2018) estabeleceram segmentos FM para cada segmento R, resultando nos seguintes grupos: para clientes ativos existem os grupos de alto valor, médio valor com alto monetário, médio valor com alta frequência e baixo valor. Para clientes que estão expirando existem os grupos de alto, médio e baixo valores, sendo aplicado também para clientes expirados. Estes grupos foram definidos por Tavakoli *et al*. (2018) com ajuda da empresa de *E-commerce* Digikala. A partir disso, aplicou-se o K-means com k=4 para o grupo de clientes ativos, k=3 para o grupo de clientes expirando e k=3 para o grupo de

clientes expirados, resultando em um total de 10 clusters. Na Figura 2 são identificados os clusters gerados somente a partir do grupo de clientes ativos, organizados em um gráfico de valor monetário por frequência. Sendo o cluster de cor verde composto pelos clientes de alto valor, o cluster de cor amarela composto pelos clientes de médio valor com alto monetário, o cluster de cor azul composto pelos clientes de médio valor com alta frequência, e por fim, o cluster de cor roxa composto pelos clientes de baixo valor.

Figura 2 – Clusters do grupo de clientes ativos



Fonte: Tavakoli *et al*. (2018).

Após a geração dos grupos, Tavakoli *et al*. (2018) discorrem sobre possíveis estratégias para cada segmento. Sugerindo um maior foco em clientes ativos com valor médio e baixo, bem como a manutenção de clientes já valiosos. Os autores também enfatizam a importância em recuperar os clientes do grupo expirando, cuja chance de retorno não é tão baixa quanto o grupo expirado, que por si só requer uma estratégia especial de reengajamento dos clientes à empresa.

Além da elaboração de estratégias, Tavakoli *et al*. (2018) implementaram uma campanha de SMS focada somente no segmento de clientes ativos (recência abaixo de 90 dias), pois a empresa já tinha realizado outras campanhas em clientes ativos anteriormente. Nesta campanha cada cliente foi presenteado com um *Voucher* condizente com o segmento ao qual o cliente pertencia. Para clientes ativos com valor alto foi oferecido um desconto de 10 por cento com um desconto máximo de até $20, com o objetivo de manter a lealdade destes clientes. Para clientes ativos de valor médio com alto valor monetário, foi oferecido um desconto de 10 por cento com um desconto máximo de até $10, o valor foi menor pois o objetivo era aumentar frequência de compra destes clientes, que em tese já gastam bastante. Para clientes ativos de valor médio com alta frequência foi oferecido um desconto de 10 por cento com um desconto máximo de até $20 para vendas que custem mais que $50 (que é o valor médio gasto por este segmento), incentivando assim uma compra de maior valor. Por fim, para clientes ativos de baixo valor, foi oferecido um desconto de 10 por cento com um desconto máximo de até $20, com o objetivo de converter estes clientes em mais leais.

Para a análise da campanha, Tavakoli *et al*. (2018) selecionaram aleatoriamente 20% dos clientes de cada segmento para compor um grupo de controle, cujos *Vouchers* não foram enviados. Este grupo de controle foi comparado com os outros grupos da campanha para obter um valor de referência do aumento do valor monetário após sua conclusão. Os resultados alcançados por Tavakoli *et al*. (2018) podem ser observados no Quadro 1, ao qual percebe-se um aumento de $14,30 na média monetária dos clientes ativos com alta frequência, mais do que o aumento de $3,20 sofrido pelo grupo de controle. É possível também observar que a média monetária de todos os grupos alvo da campanha aumentou consideravelmente, enquanto o grupo de controle aumentou pouco ou até diminuiu, indicando uma efetividade no objetivo da campanha.

Quadro 1 – Dados da média monetária do grupo de controle e grupo de campanha

Segment		Average Monetary (USD)			
Recency	Monetary and Frequency	Control Users		Campaign Users	
		Before Campaign	After Campaign	Before Campaign	After Campaign
Active	High Value	74.2	73.8	88.2	89.2
	Medium Value with High Monetary	100.2	97.6	104.6	105.2
	Medium Value with High Frequency	32	35.2	35.4	49.7
	Low Value	50.7	53.2	56.4	65.2

Fonte: Tavakoli *et al*. (2018).

Tavakoli *et al*. (2018) concluem que houve uma melhora no desempenho da campanha lançada em comparação com as anteriores, indicando ainda, que elas obtinham uma taxa de compra de 0,1 por cento, sendo

que a campanha lançada para validação do modelo obteve uma taxa de 1 por cento, cerca de dez vezes mais efetivo. Os autores justificam esta melhora ao processo de segmentação do modelo, resultando em clusters mais significativos, facilitando a aplicação de vouchers específicos. Por fim, Tavakoli *et al*. (2018) sugerem o melhoramento da definição do atributo de recência de forma que seja mais útil ao time de marketing. Também recomendam calcular o Customer Lifetime Value (CLV), que atribui o valor vitalício à cada segmento e cliente, de forma a quantificar o valor que um cliente pode proporcionar à empresa.

3 PROPOSTA DO PROTÓTIPO

Esse capítulo visa apresentar a justificativa para a elaboração deste trabalho, os requisitos que serão seguidos e a metodologia que será utilizada. Será apresentada também uma breve revisão bibliográfica das principais áreas de estudo que serão exploradas, bem como os principais termos utilizados.

3.1 JUSTIFICATIVA

No Quadro 2 é apresentado um comparativo entre os trabalhos correlatos. As linhas representam as características relevantes e as colunas representam os trabalhos.

Quadro 2 – Comparativo entre os trabalhos correlatos

Correlatos / Características	Gustriansyah, Suhandi e Antony (2020)	Peker, Kocyigit e Eren (2017)	Tavakoli *et al*. (2018)
Alvo da clusterização	Produtos	Clientes	Clientes
Modelo utilizado	RFM	LRFMP	R+FM
Objetivo da segmentação	Gerenciamento de estoque	Gerenciamento das relações com cliente	Gerenciamento das relações com cliente
Algoritmo de clusterização utilizado	K-means	K-means	K-means
Foco metodológico	Otimização de k com diferentes métricas	Formulação de um modelo novo e análise dos resultados	Formulação de um modelo novo e campanha de ofertas
Número de dados (clientes/produtos)	2.043	16.024	~3.000.000
Quantidade de índices para validação de k	8	3	-
Quantidade de clusters gerados	3	5	10
Inferências sobre os dados	-	Sim	Sim

Fonte: elaborado pelo autor.

A partir do Quadro 2, pode-se observar que Gustriansyah, Suhandi e Antony (2020) clusterizaram produtos de uma base de dados utilizando o modelo RFM padrão. Já Peker, Kocyigit e Eren (2017) optaram pelo desenvolvimento de um modelo novo, considerando a periodicidade (LRFMP). Tavakoli *et al*. (2018) também desenvolveram um novo modelo, ao qual a característica recência foi modificada e separada (R+FM).

Gustriansyah, Suhandi e Antony (2020) tinham como objetivo melhorar o gerenciamento de estoque, prezando por uma segmentação mais conclusiva sobre os produtos, visto que o modelo RFM padrão define segmentos arbitrariamente sem adequar-se às peculiaridades dos dados, enquanto o modelo aplicado através de k-means alcançou uma segmentação com dados altamente similares em cada cluster. Por outro lado, Peker, Kocyigit e Eren (2017) e Tavakoli *et al*. (2018) objetivavam o gerenciamento das relações com os clientes através de estratégias focadas em segmentos, visando aumentar a renda que eles fornecem à empresa. Todos os autores utilizaram o algoritmo K-means, por ser confiável e amplamente difundido. Vale ressaltar que no trabalho de Gustriansyah, Suhandi e Antony (2020), o algoritmo teve um foco metodológico maior, visto que foram utilizados 8 índices de validação para k clusters, visando otimizar a organização dos segmentos.

A quantidade de dados segmentados variou bastante entre os três trabalhos devido aos diferentes contextos de aplicação. Gustriansyah, Suhandi e Antony (2020) tinham 2.043 produtos na base de dados para segmentar, resultando em 3 clusters. Já Peker, Kocyigit e Eren (2017) possuíam o registro de 16.024 clientes de uma rede de padarias, sendo especificados 5 segmentos, obtidos através de uma análise por três índices de validação (Silhouette, Calinski-Harabasz e Davies-Bouldin). Por fim, Tavakoli *et al*. (2018) agruparam dados de 3 milhões de clientes pertencentes à base de dados de um E-commerce do Oriente Médio, resultando em 10 clusters, sendo 3 pertencentes à característica de recência, e os outros 7 distribuídos entre as características de frequência e monetária. Ressalta que Tavakoli *et al*. (2018) testaram o modelo em produção, montando uma campanha que focava no segmento de clientes ativos, visando primariamente aumentar os lucros da empresa, utilizando também um grupo de controle e comparação de renda antes e depois da campanha.

Gustriansyah, Suhandi e Antony (2020) demonstraram a possibilidade da aplicação de RFM fora do uso convencional de segmentação de clientes, e adquiriram clusters com uma variância média de 0.19113. Além disso, os autores sugeriram outras formas de comparação de dados, como Particle Swarm Optimization (PSO), medioides ou até *maximizing-expectancy*. Peker, Kocyigit e Eren (2017) segmentaram clientes de uma rede de mercados na Turquia em "clientes leais de alta contribuição", "clientes leais de baixa contribuição", "clientes incertos", "clientes perdidos de alto gasto" e "clientes perdidos de baixo gasto". Desta maneira, os autores providenciaram visões e estratégias (promoções, ofertas, regalias) de aumento de renda sobre os comportamentos dos clientes, porém limitaram-se a aplicar em um segmento específico de mercado. Por fim, Tavakoli *et al*. (2018) agruparam clientes de uma empresa de E-commerce com base em sua recência, resultando em clientes "Ativos", "Expirando" e "Expirados", e destes segmentos, sucessivamente separados em grupos de "Alto", "Médio" e "Baixo" valores, validando posteriormente a segmentação através de uma campanha de ofertas para os clientes do grupo "Ativos".

Todos os trabalhos aqui citados procuraram implementar o modelo RFM num contexto de clusterização por K-means, alterando o modelo e o manejo dos dados de acordo com cada categoria, seja ele produto ou cliente, varejo ou mercado. Com isso, criaram-se atributos e foram modificados alguns já existentes para atender às especificidades de cada contexto, visto que todos os trabalhos focaram em uma só base de dados, inevitavelmente adequando-se às mesmas.

Desta forma, este trabalho demonstra ser relevante, pois almeja aplicar o modelo RFM em conjunto com vários algoritmos de clusterização em forma de um artefato computacional que se adeque à vários contextos (mercado, comércio, varejo etc.), utilizando várias bases de dados reais para testar a validade dos algoritmos utilizados. Vislumbra-se utilizar três índices para validação da qualidade dos clusters (Silhouette, Calinski-Harabasz e Davies-Bouldin). Além disso, deseja-se obter clusters significativos e coerentes com cada segmento de mercado aplicado. Outra contribuição deste trabalho seria no âmbito comercial, com a geração de informações sobre as similaridades de clientes de cada segmento de mercado, podendo auxiliar gestores e administradores de empresas a obter uma visão crítica sobre os comportamentos de clientes ao longo das diferentes bases de dados, podendo também denotar características comuns a todos. Outra relevância seria a utilização deste trabalho em ambiente acadêmico, visto que serão aplicados diferentes algoritmos de clusterização, podendo providenciar informações sobre seus desempenhos e qualidade de agrupamento, além de ser aplicados processos de obtenção, limpeza e transformação de dados.

3.2 REQUISITOS PRINCIPAIS DO PROBLEMA A SER TRABALHADO

O artefato computacional a ser desenvolvido deverá:

a) adquirir os dados transacionais de clientes a partir de um banco de dados (Requisito Funcional - RF);
b) extrair dos clientes as características (recência, frequência e monetária) utilizadas no modelo RFM (RF);
c) filtrar os clientes sem quantidade de compras relevantes (RF);
d) normalizar os dados para evitar disparidades nas escalas dos dados, principalmente no atributo monetário (RF);
e) aplicar três índices de validação (Silhouette, Calinski-Harabasz e Davies-Bouldin) para verificar a qualidade dos clusters (RF);
f) apresentar em um gráfico 3D os clientes, com sua localização definida pela pontuação do cliente nas características RFM (RF);
g) utilizar algoritmos de clusterização tais como K-means, mean-shift e DBSCAN (RF);
h) utilizar a linguagem Python para o desenvolvimento (Requisito Não Funcional - RNF);
i) utilizar o ambiente de desenvolvimento Jupyter Notebook (RNF);
j) utilizar o banco de dados PostgreSQL para ler os dados das bases utilizadas (RNF).

3.3 METODOLOGIA

O trabalho será desenvolvido observando as seguintes etapas:

a) levantamento bibliográfico: pesquisar trabalhos relacionados e estudar sobre o modelo RFM e suas aplicações, algoritmos de clusterização, métodos de tratamento de dados e índices de validação;
b) seleção de bases de dados: obter bases de dados de usuários cedidas pela empresa Intelidata Informática, desenvolvedora de software de gestão comercial. Serão selecionadas conforme sua adequação ao objetivo do trabalho, variando em tamanho e segmento de mercado;
c) definição das características do modelo RFM: definir os atributos utilizados para caracterizar os clientes no modelo RFM;
d) definição de métricas do modelo RFM: definir as métricas para mensuração e atribuição de pontuação de cada característica no modelo RFM;

e) definição dos algoritmos de clusterização: pesquisar e escolher o algoritmo de clusterização que realizará o agrupamento das características RFM;
f) implementação: implementar o artefato computacional de segmentação levando em consideração as etapas (b) até (e), utilizando a linguagem Python;
g) análise dos clusters: avaliar a qualidade dos clusters gerados a partir dos diferentes algoritmos de clusterização e seus comportamentos em múltiplas bases de dados, aplicando índices de validação e apresentando-os na forma de gráficos.

As etapas serão realizadas nos períodos relacionados no Quadro 3.

Quadro 3 – Cronograma de atividades a serem realizadas

	2022									
	fev.		mar.		abr.		maio		jun.	
etapas / quinzenas	1	2	1	2	1	2	1	2	1	2
levantamento bibliográfico										
seleção de bases de dados										
definição das características do modelo RFM										
definição de métricas do modelo RFM										
definição dos algoritmos de clusterização										
implementação										
análise dos clusters										

Fonte: elaborado pelo autor.

4 REVISÃO BIBLIOGRÁFICA

Neste capítulo serão brevemente aprofundados os assuntos que servirão de base para a realização deste trabalho. Serão tratados os temas de clustering, modelo RFM e índices de validação de clusters.

Para Cherkassky e Mulier (2007), clustering se trata do problema de separar um conjunto de dados em grupos chamados de "clusters" baseado em alguma medida de similaridade. O objetivo é encontrar um conjunto de clusters dos quais as amostras dentro dos mesmos são mais similares entre si do que quando comparadas com amostras de outros clusters. Existem vários algoritmos para clusterizar dados, especializando-se em situações específicas, como o *Density-based spatial clustering of applications with noise* (DBSCAN), que é um algoritmo que agrupa dados baseado em sua densidade, utilizando conceitos de pontos núcleo, pontos vizinhos e ruído. Outro método utilizado é o *Hierarchical Clustering*, que constrói uma hierarquia de clusters, podendo aplicar abordagens de cima para baixo, onde as observações parte de um cluster que é dividido em outros menores, ou a abordagem de baixo para cima, onde cada dado inicia como um cluster, e é então agrupado conforme o algoritmo é executado. Por fim, um dos algoritmos mais utilizados é o K-means, cuja orientação é focada em centroides, ou seja, seus clusters são representados por um ponto central (não necessariamente fazendo parte do conjunto de dados) que possui a menor distância possível entre si mesmo e o resto dos dados do cluster.

Segundo Hughes (2011), o modelo RFM é "Um meio antigo e altamente preditivo de determinar quem irá responder e comprar. Um método de codificar clientes existentes. Usado para prever resposta, tamanho médio de pedido, e outros fatores". Este modelo categoriza geralmente clientes através das características de recência (R), frequência (F) e monetária (M). As métricas utilizadas para medir tais características podem variar, porém geralmente classificam recência como a quantidade de dias desde a última compra, frequência como a quantidade de compras dentro de um determinado período, e monetária como o total acumulado de todas as vendas realizadas para um cliente. Este modelo é muito utilizado em marketing direto, onde os meios de comunicação são diretamente entre a empresa e o consumidor, realizado através de mídias sociais, e-mail, mensagens SMS ou até pelo correio. No estudo realizado por Verhoef *et al*. (2003), cerca de 90% das empresas questionadas sobre a aplicação métodos de segmentação (como o RFM) afirmam que possuíam como objetivo a seleção de alvos, ou seja, encontrar o segmento de clientes que mais se identificam com a empresa, e 64,4% citaram como objetivo o tratamento diferencial de clientes, com promoções, preços e ofertas especiais.

O índice de Silhouette é utilizado no conceito de clusterização, para analisar a qualidade de um dado localizado em determinado cluster, levando em conta a distância média entre clusters. Com o cálculo deste índice, valores próximos de 1 para um determinado dado em um cluster são considerados bons, valores perto de 0 indicam que o dado está entre clusters e caso o valor seja próximo de -1 significa que provavelmente o dado está no cluster errado. O índice de Calinski-Harabasz é definido como a razão de duas somas calculadas entre todos os clusters: a soma da dispersão intra-clusters e soma da dispersão inter-clusters. Quanto maior seu valor, melhor a performance da clusterização observada. Por fim, o índice de Davies-Bouldin indica a similaridade média entre clusters, levando em conta sua distância e tamanho. Um valor baixo para este índice indica uma melhor separação entre os clusters.

REFERÊNCIAS

CHERKASSKY, Vladimir S.; MULIER, Filip. Methods for data reduction and dimensionality reduction. In: CHERKASSKY, Vladimir S.; MULIER, Filip. **Learning from data**: concepts, theory, and methods. 2. ed. Hoboken: Ieee Press, 2007. Cap. 6, p. 191

GUSTRIANSYAH, Rendra; SUHANDI, Nazori; ANTONY, Fery. Clustering optimization in RFM analysis Based on k-Means. **Indonesian Journal Of Electrical Engineering And Computer Science**, [S. l], v. 18, n. 1, p. 470-477, abr. 2020. Mensal. Disponível em: http://ijeecs.iaescore.com/index.php/IJEECS/article/view/20264. Acesso em: 02 set. 2021

HUGHES, Arthur Middleton. **Strategic Database Marketing 4e**: the masterplan for starting and managing a profitable, customer-based marketing program. 4. ed. [S. l.]: McGraw-Hill, 2011. 608 p.

KUMAR, V. **Managing Customers for Profit**: strategies to increase profits and build loyalty. Upper Saddle River: Pearson Prentice Hall, 2008. 296 p.

NGUYEN, Thuyuyen H.; SHERIF, Joseph S.; NEWBY, Michael. Strategies for successful CRM implementation. Information Management & Computer Security, [S.l.], v. 15, n. 2, p. 102-115, maio 2007. Disponível em: https://www.emerald.com/insight/content/doi/10.1108/09685220710748001/full/html?journalCode=imcs. Acesso em: 26 set. 2021.

PEKER, Serhat; KOCYIGIT, Altan; EREN, P. Erhan. LRFMP model for customer segmentation in the grocery retail industry: a case study. **Marketing Intelligence & Planning**, [S.l.], v. 35, n. 4, p. 544-559, 6 maio 2017. Emerald. http://dx.doi.org/10.1108/mip-11-2016-0210. Disponível em: https://www.emerald.com/insight/content/doi/10.1108/MIP-11-2016-0210/full/html. Acesso em: 07 set. 2021.

PETRISON, Lisa A.; BLATTBERG, Robert C.; WANG, Paul. Database marketing: past, present, and future. **Journal Of Direct Marketing**, [S.l.], v. 11, n. 4, p. 109-125, mar. 1997. Wiley. http://dx.doi.org/10.1002/(sici)1522-7138(199723)11:43.0.co;2-g. Disponível em: https://onlinelibrary.wiley.com/doi/10.1002/(SICI)1522-7138(199723)11:4%3C109::AID-DIR12%3E3.0.CO;2-G. Acesso em: 19 set. 2021.

RASHID, Mohammad A.; HOSSAIN, Liaquat; PATRICK, Jon David. The Evolution of ERP Systems: a historical perspective. In: NAH, Fiona Fui-Hoon. **Enterprise Resource Planning**: solutions and management. Hershey: Irm Press, 2001. p. 35-50. Disponível em: https://books.google.com.br/books?id=qBcJwDWk4ioC&lpg=PR1&ots=9MrXoQhaRL&dq=Enterprise%20Resource%20Planning%3A%20Solutions%20and%20Management&lr&hl=pt-BR&pg=PR1#v=onepage&q=Enterprise%20Resource%20Planning:%20Solutions%20and%20Management&f=false. Acesso em: 19 set. 2021.

REINARTZ, Werner; THOMAS, Jacquelyn S.; KUMAR, V. Balancing Acquisition and Retention Resources to Maximize Customer Profitability. **Journal Of Marketing**, [S.l.], v. 69, n. 1, p. 63-79, jan. 2005.

TAVAKOLI, Mohammadreza *et al*. Customer Segmentation and Strategy Development Based on User Behavior Analysis, RFM Model and Data Mining Techniques: a case study. In: 2018 IEEE 15TH INTERNATIONAL CONFERENCE ON E-BUSINESS ENGINEERING (ICEBE), 15., 2018, Xiam. **Proceedings [...] .** [S.l.]: Ieee, 2018. p. 119-126. Disponível em: https://www.researchgate.net/publication/330027350_Customer_Segmentation_and_Strategy_Development_Based_on_User_Behavior_Analysis_RFM_Model_and_Data_Mining_Techniques_A_Case_Study. Acesso em: 11 set. 2021.

TSIPTSIS, Konstantinos K.; CHORIANOPOULOS, Antonios. Data Mining Techniques in CRM: inside customer segmentation. Chichester: John Wiley & Sons, 2009. 374 p.

VERHOEF, Peter C *et al*. The commercial use of segmentation and predictive modeling techniques for database marketing in the Netherlands. **Decision Support Systems**, [S.l.], v. 34, n. 4, p. 471-481, mar. 2003. Disponível em: https://liacs.leidenuniv.nl/~puttenpwhvander/library/Others/segmpredmodel-hoekstra.pdf. Acesso em: 19 set. 2021.

FORMULÁRIO DE AVALIAÇÃO – PROFESSOR TCC I

Avaliador(a): Dalton Solano dos Reis

	ASPECTOS AVALIADOS[1]	atende	atende parcialmente	não atende
ASPECTOS TÉCNICOS	1. INTRODUÇÃO O tema de pesquisa está devidamente contextualizado/delimitado?			
	O problema está claramente formulado?			
	2. OBJETIVOS O objetivo principal está claramente definido e é passível de ser alcançado?			
	Os objetivos específicos são coerentes com o objetivo principal?			
	3. JUSTIFICATIVA São apresentados argumentos científicos, técnicos ou metodológicos que justificam a proposta?			
	São apresentadas as contribuições teóricas, práticas ou sociais que justificam a proposta?			
	4. METODOLOGIA Foram relacionadas todas as etapas necessárias para o desenvolvimento do TCC?			
	Os métodos, recursos e o cronograma estão devidamente apresentados?			
	5. REVISÃO BIBLIOGRÁFICA (atenção para a diferença de conteúdo entre projeto e pré-projeto) Os assuntos apresentados são suficientes e têm relação com o tema do TCC?			
ASPECTOS METODOLÓGICOS	6. LINGUAGEM USADA (redação) O texto completo é coerente e redigido corretamente em língua portuguesa, usando linguagem formal/científica?			
	A exposição do assunto é ordenada (as ideias estão bem encadeadas e a linguagem utilizada é clara)?			
	7. ORGANIZAÇÃO E APRESENTAÇÃO GRÁFICA DO TEXTO A organização e apresentação dos capítulos, seções, subseções e parágrafos estão de acordo com o modelo estabelecido?			
	8. ILUSTRAÇÕES (figuras, quadros, tabelas) As ilustrações são legíveis e obedecem às normas da ABNT?			
	9. REFERÊNCIAS E CITAÇÕES As referências obedecem às normas da ABNT?			
	As citações obedecem às normas da ABNT?			
	Todos os documentos citados foram referenciados e vice-versa, isto é, as citações e referências são consistentes?			