Aplicação de Clustering para Segmentação de Clientes na Base de Dados da Justa

Clustering Application for Customer Segmentation in the JUSTA Database

Allana Rocha[1]
orcid.org/0000-0002-6787-2677

Ester de Macêdo[1]
orcid.org/0000-0003-4049-9937

Letícia Portela[1]
orcid.org/0000-0001-7214-1644

Vinícius Silva[1]
orcid.org/0000-0001-9889-2331

[1]Escola Escola Politécnica de Pernambuco, Universidade de Pernambuco, Recife, Brasil. E-mail: alsr@ecomp.poli.br

DOI: 10.25286/repa.v7i3.2458

Como citar este artigo pela NBR 6023/2018: Allana Rocha; Ester de Macêdo; Letícia Portela; Vinícius Silva. Clustering Application for Customer Segmentation in the JUSTA Database. Revista de Engenharia e Pesquisa Aplicada, Recife, v. 7, n. 3, p. 39-53.

RESUMO

Empresas de tecnologia financeira, mais conhecidas como fintechs, são companhias de inovação tecnológica com potencial transformador para o setor financial. Nelas, o tratamento personalizado requer a análise de quantidades expressivas de dados. Dessa forma, utilizar técnicas de mineração de dados pode oferecer maior facilidade em classificar e visualizar os consumidores. A empresa analisada nesse artigo, a Justa, é uma fintech que promove produtos e serviços através de uma conta digital, e que procurava aprimorar a classificação dos seus clientes. A partir das bases de dados anonimizadas, fornecida pela Justa, cada cliente foi representado por features consideradas importantes para a empresa. Para chegar na base final, foi feita a integração, redução, limpeza, e transformação dos dados. Os algoritmos testados para agrupar os clientes foram K-Means, fuzzy C-Means e K-Medoids, onde o K-medoids, aplicado com a distância de Gower, apresentou melhor resultado na delineação dos perfis. Os resultados indicaram que há perfis diferentes de clientes, mas que estes são pouco acentuados e estão concentrados em apenas algumas das características comportamentais.

PALAVRAS-CHAVE: Segmentação de mercado; Agrupamento; Mercado Financeiro; Python;

ABSTACT

Financial technology companies, also known as fintechs, are innovative technology companies with the potential to transform the financial sector. For them to apply a personalised treatment of clients, extensive data analysis is required. Therefore, employing data mining techniques can offer advantages in classifying and visualising costumers. Justa, the company explored in this work, is a fintech that provides products and services through digital bank accounts, and it sought to improve its understanding of its client base. Using anonymised datasets provided by Justa, each client was represented by features they considered relevant. To arrive at the final dataset, the integration, reduction, cleansing, and transformation of the original data was performed. The algorithms tested for grouping customers were K-Means, fuzzy C-Means and K-Medoids, where K-medoids presented better results in the delineation of the profiles. The results indicated that there are different profiles of clients, but that these are barely perceptible and are concentrated in a few behavioral characteristics.

KEY-WORDS: Market segmentation; Clustering; Financial market;

1 INTRODUÇÃO

Empresas de tecnologia financeira, mais conhecidas como fintechs, são companhias de inovação tecnológica com potencial transformador para o setor financial. Elas têm tido um papel ativo no incremento das experiências financeiras digitais, projetando soluções que impulsionam o mercado.

As fintechs vêm investindo em soluções para oferecer tratamento personalizado aos clientes. Estas organizações, no entanto, ao se expandirem no mercado, precisam lidar com quantidades expressivas de clientes e, consequentemente, de dados. Para tornar possível analisar e tomar decisões mais complexas, as técnicas de mineração de dados podem oferecer, entre outros, uma facilidade em classificar e visualizar quantidades numerosas de dados.

Nesse contexto, a Justa, fintech que promove produtos e serviços através de uma conta digital, precisava de um modelo inteligente que pudesse categorizar automaticamente seus clientes. Este problema surgiu da decisão de aplicar estratégias de promoção e fidelização personalizadas aos clientes, considerando seus históricos de transações realizadas através das soluções oferecidas pela empresa. Por ser inviável analisar cada cliente, individualmente e manualmente, optou-se por classificá-los em perfis e oferecer soluções de acordo com estes. Essa classificação, porém, exige avaliar uma extensa base de dados com mais de 3.000 clientes.

Atualmente, a Justa já faz uso de estratégias de benefícios para clientes específicos. Lojistas que utilizam unicamente a classe de produto denominada "POS" são classificados como “Super Heróis”. A partir de estímulos direcionados a essa classe, foi possível alavancar em 7% o volume transacional destes clientes.

A partir da base de dados anonimizada, fornecida pela Justa, cada cliente foi representado por features consideradas importantes para a empresa, geradas a partir da base de dados transacional. Estes dados foram avaliados para agrupar os clientes em 4, 5 e 6 classes diferentes (valores definidos pelos stakeholders).

O clustering, ou agrupamento, é um conjunto de técnicas de mineração de dados que identifica, de maneira não supervisionada, possíveis grupos em uma base de dados, a partir de semelhanças entre os indivíduos. Neste artigo, para essa finalidade, foram utilizados os algoritmos K-Means, com auxílio da metaheurística PSOC (Particle Swarm Optimization for Clustering), o Fuzzy C-Means e o K-Medoids. O K-Means é uma técnica consolidada e simples, mas limitada à aplicação da distância euclidiana. Como alternativa, utilizou-se o K-Medoids para aplicação da distância de Gower, mais adequada à natureza do problema.

Os resultados consistem em modelos possíveis para o agrupamento dos clientes, cabendo à empresa decidir se estes se enquadram em seu padrão de negócios. O desenvolvimento de uma aplicação para implantação no sistema da empresa e a criação de um modelo de aprendizado contínuo não foram contemplados no escopo deste projeto.

A seção 2 do artigo contém a fundamentação teórica e os trabalhos relacionados ao problema abordado. A seção 3 apresenta os dados utilizados e quais análises e tratamentos foram realizados, assim como a metodologia utilizada nos experimentos. Os resultados estão na seção 4 e a seção 5 contém conclusões e possíveis trabalhos futuros.

2 FUNDAMENTAÇÃO TEÓRICA

2.1 SEGMENTAÇÃO DE MERCADO

A segmentação de mercado pode ser descrita como o processo de particionamento de um grande mercado em grupos ou clusters menores **[1]**. Esse particionamento tem o objetivo de focalizar estratégias de marketing, ofertar e desenvolver novos produtos ou serviços baseado nas características específicas de cada grupo.

Em **[2]** foi proposto o modelo analítico RFM, que foi criado para ajudar os profissionais de marketing a segmentar grupos específicos de clientes, ou seja, para caracterizar clientes importantes em grandes conjuntos de dados por três variáveis: recência, frequência e valor monetário. O modelo RFM é baseado em três fatores quantitativos:

- Recência de um cliente (R): O intervalo de tempo entre a data mais recente nos dados e a data da transação mais recente do cliente.
- Frequência de um cliente (F): O número de transações do cliente em um período específico.
- Valor monetário de um cliente (M): A soma do valor gasto em cada transação do cliente.

2.2 MINERAÇÃO DE DADOS

A mineração de dados é utilizada para analisar grandes quantidades de dados e obter informações relevantes sobre eles, de acordo com a finalidade pretendida. Para alcançar esse objetivo, foram desenvolvidas diversas técnicas para cada tipo de problema de domínio conhecido, como agrupamentos, associações, etc. Outras técnicas se destinam a problemas cujo caminho para a solução não é conhecido, como na aprendizagem de máquina para classificações mais complexas.

Clustering é o processo de agrupar um conjunto de observações em grupos de observações semelhantes. Um cluster é uma coleção de observações que são semelhantes entre si dentro do mesmo cluster e são diferentes das observações em outros clusters **[2]**.

K-Means é um método de clusterização difícil e de centróides **[3]**. Para utilizá-lo, define-se um número K de grupos e escolhe-se um centróide aleatório para cada um deles. Os dados de treinamento são então distribuídos entre os grupos, sendo colocados sempre no mais próximo. Finalmente, o centróide é reposicionado para a média dos pontos, e esses passos se repetem durante o resto do treinamento.

No entanto, os centróides finais no K-Means não são interpretáveis, pois eles não representam pontos reais, apenas a média dos pontos no cluster. O K-Medoids **[4]** busca utilizar pontos da base como centróides. O método utilizado para determinar qual ponto será escolhido é o Particionamento Entorno de Medoids (PAM). Nele, a atualização dos centróides funciona testando todos os pontos dentro do cluster e selecionando o que tem a menor *loss*.

Uma tarefa importante ao agrupar os dados é decidir qual métrica será usada para calcular a distância entre cada ponto. A distância de Gower (1971) pode ser utilizada quando os dados são mistos (quantitativos, ordinais e nominais) **[5]**. É sempre um número entre 0 (idêntico) e 1 (máximo diferente).

2.3 TRABALHOS RELACIONADOS

Gerenciamento de relacionamento com clientes é amplamente estudado como uma forma de fidelização de clientes em empresas **[6] [7]**. Há diversas estratégias para categorização de consumidores, utilizando estatística, mineração de dados e/ou aprendizado de máquina **[8]**.

Shen **[2]** fornece um exemplo de segmentação de clientes, usando aprendizado de máquina não supervisionado para desenvolver estratégias diferenciadas para diferentes clientes de plataformas online. Para distinguir os clientes, as características comportamentais são obtidas a partir do modelo RFM (Recency, Frequency, Monetary Value). Além disso, os produtos adquiridos são classificados em diferentes categorias por Term Frequency - Inverse Document Frequency (TF-IDF) e métodos de clustering para refletir as preferências de produtos dos clientes. Para a parte de modelagem, os clientes foram segmentados em vários grupos usando o algoritmo de clustering K-Means, e os clusters são visualizados após a redução da dimensão dos dados pelos métodos de PCA (Principal Component Analysis) e T-Distributed Stochastic Neighbor Embedding (T-sne). Pensando nisso, foram exploradas e identificadas as principais características dos consumidores em cada segmento, e fornecidas algumas recomendações baseadas nos grupos de clientes.

Em [**9**], é desenvolvida uma segmentação dos clientes de um banco baseada em seus comportamentos. Também é utilizado o algoritmo K-Means através de junções (mergings), com a escolha dos parâmetros de entrada (entre os 35 disponíveis), auxiliada pela técnica da análise de fatores e o critério de Kaiser. Os 6 parâmetros de entrada selecionados incluíam valor das transações, quantidade de transações e produtos utilizados. Foram encontrados 4 clusters, dentre os quais: um cluster de pequenas companhias com poucos ativos, valores de transações baixos mas alta quantidade de transações, e outro, contendo clientes com valores baixos de transações, poucas transações e poucos produtos, geralmente identificando clientes novos. Algumas recomendações são sugeridas ao final do experimento, contextualizadas com a área de negócio de um banco.

3 MATERIAIS E MÉTODOS

3.1 STAKEHOLDERS ENVOLVIDOS

O Apoiaram o desenvolvimento deste trabalho os stakeholders Felippe Nagy (Analisa de Dados) e Genrique Feliciano (Engenheiro de Dados) ambos da empresa Justa. Como nível de importância e influência, os dois são considerados de alta influência na empresa por estarem à frente dos estudos dos dados.

DOI: 10.25286/repa.v7i3.2458

3.2 DESCRIÇÃO DA BASE DE DADOS

Os dados disponibilizados pela empresa Justa se distribuem em duas tabelas. Uma delas com 19 colunas e 1.651.385 instâncias, possuindo dados referentes às transações realizadas pelos lojistas. Contém informações como modo de pagamento da venda, data da transação, valor da transação, entre outros. Já em outra tabela, com dimensão de 3402 x 12, há informações cadastrais referentes a cada cliente, como data de cadastro, data da primeira e da última movimentação do cliente, e assim por diante. As duas tabelas podem se comunicar a partir do número de identificação do lojista. As informações são obtidas pelo processamento da transação junto às bandeiras de cartão, sendo formatadas e armazenadas na base de dados da empresa, posteriormente.

3.3 DICIONÁRIO DE DADOS

A Tabela 1 contém os atributos (features), por cliente, que serão utilizados como métricas para a classificação. Alguns não constam na base de dados disponibilizada para criação do modelo, mas foram derivados das informações fornecidas.

Tabela 1- Dicionário da Tabela de Atributos por Cliente

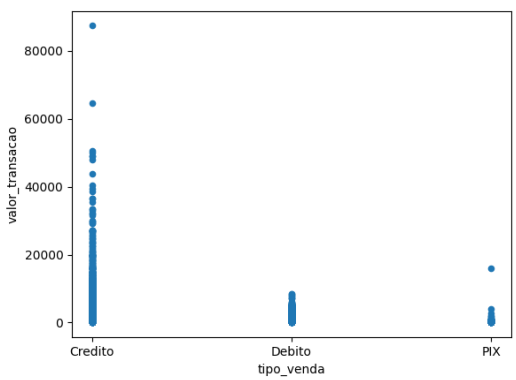
DADOS	DESCRIÇÃO	NÍVEL
qtd_total_transacoes	Qtd. de transações	4
media_por_crédito	Média ponderada da qtd. de transações por crédito. Transações mais recentes têm maior peso.	7
media_por_debito	Media ponderada da qtd. de transações por débito. Semelhante ao anterior.	7
prox_inicio_ano	Índice indicativo do quanto as transações de um cliente ocorreram próximas ao início do ano.	6
prox_meio_ano	Semelhante ao anterior. Considerando proximidade com a metade do ano.	6
prox_fim_ano	Semelhante ao anterior. Considerando proximidade com o fim do ano.	6
prox_manha	Índice indicativo do quanto as transações de um cliente ocorreram próximas ao horário da manhã.	8
prox_tarde	Semelhante ao anterior. Considerando proximidade com o período da tarde.	8
prox_noite	Semelhante ao anterior. Considerando proximidade com o período da noite.	8
receita total gerada	Somatório da receita total das transações do lojista.	1
potencial_mensal	O somatório de receita total, considerando o período de um mês, no mês mais próspero do cliente).	6
tempo_de_cadastro	Há quanto tempo o cliente está cadastrado.	3
recencia	Valor indicativo da assiduidade do lojista.	2
ntk_mes_n	Razão entre somatório da receita total sobre o somatório do valor da transação. Obtido para um período de um mês (indicado pelo valor de n = 1, 2 ou 3). Calculado para os últimos 3 meses de transações.	1

Fonte: Os Autores.

3.4 ANÁLISE DESCRITIVA DOS DADOS

A Figura 1 ilustra a distribuição de valores de transação para vendas em "Crédito", "Débito" e "Pix". Observando a dispersão, vemos que "Crédito" apresenta uma maior variedade de valores. Já as transações em pix são as menos comuns, e apresentam uma distribuição similar ao tipo "Débito".

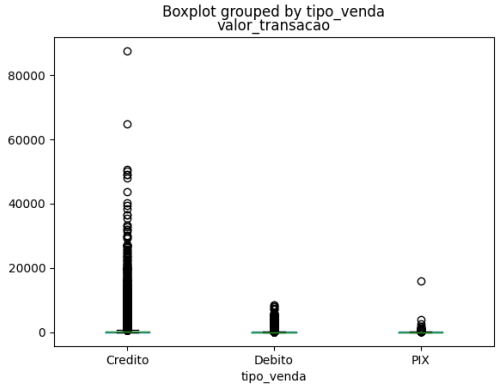
Figura 1 – Dispersão dos valores de transação em "Crédito", "Débito" e "Pix".



Fonte: Os Autores.

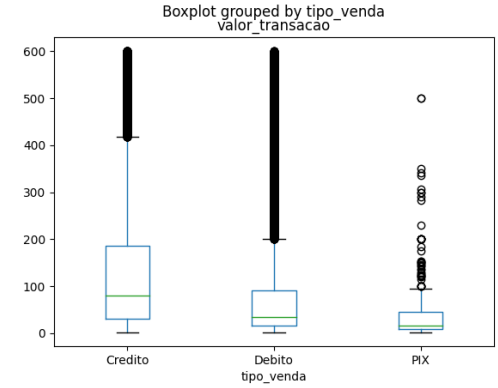
A partir dos boxplots presentes na Figura 2, observa-se que há um grande número de outliers nas transações em "Crédito", como esperado pelo visto no gráfico de dispersão. Para melhor visualização, foram gerados, também, os boxplots da Figura 3, em que os valores são limitados a R$ 600,00. Neles, vemos novamente que as transações em cartões de crédito têm valores mais altos, e as em pix têm valores mais baixos.

Figura 2 – Boxplots dos valores de transação em "Crédito", "Débito" e "Pix".



Fonte: Os Autores.

Figura 3 – Boxplots dos valores de transação em "Crédito", "Débito" e "Pix", limitados até 600.



Fonte: Os Autores.

3.5 PRÉ-PROCESSAMENTO DOS DADOS

Os fluxogramas das figuras 4, 5 e 6 demonstram o pré-processamento de dados realizado. Essa etapa consistiu, principalmente, em significar os valores ausentes, remover informações redundantes ou irrelevantes, corrigir inconsistências e realizar agrupamentos relacionais. A justificativa para estes passos vem do entendimento da área de negócios da Justa e foi realizada com auxílio dos stakeholders.

3.6 METODOLOGIA EXPERIMENTAL

Como citado anteriormente, foram realizados agrupamentos com 3 algoritmos: K-Means, fuzzy C-Means e K-Medoids. Para os agrupamentos com K-Means e C-Means, por serem algoritmos altamente influenciados pela inicialização, optou-se pela metaheurística PSOC. O PSOC é um algoritmo que busca a solução (modelo) que minimize uma função objetivo estabelecida. Para compor a função objetivo, utilizou-se a técnica estatística do teste U de Mann-Witney.

Todas as execuções foram realizadas em Python. Para o K-Means, o fuzzy C-Means, o K-Medoids e o teste U de Mann-Whitney utilizou-se, respectivamente, os modelos oferecidos pelas bibliotecas SciKit-Learn, SciKit-Fuzzy, SciKit-Learn-Extra e SciPy.

Para realizar o treino e o teste, realizou-se a técnica de Cross-Validation com K-Fold. Em todas as execuções, foram realizados 5 splits na base de dados (4 para treino e 1 para teste).

- Escolher lojistas com receita_total_gerada maiores que -10000 e menores 10000.
- Pegar lojistas com qtd_transacoes menores que e iguais a 5000.
- E logo após remover lojistas com potencial_mensal mínimo e máximo.

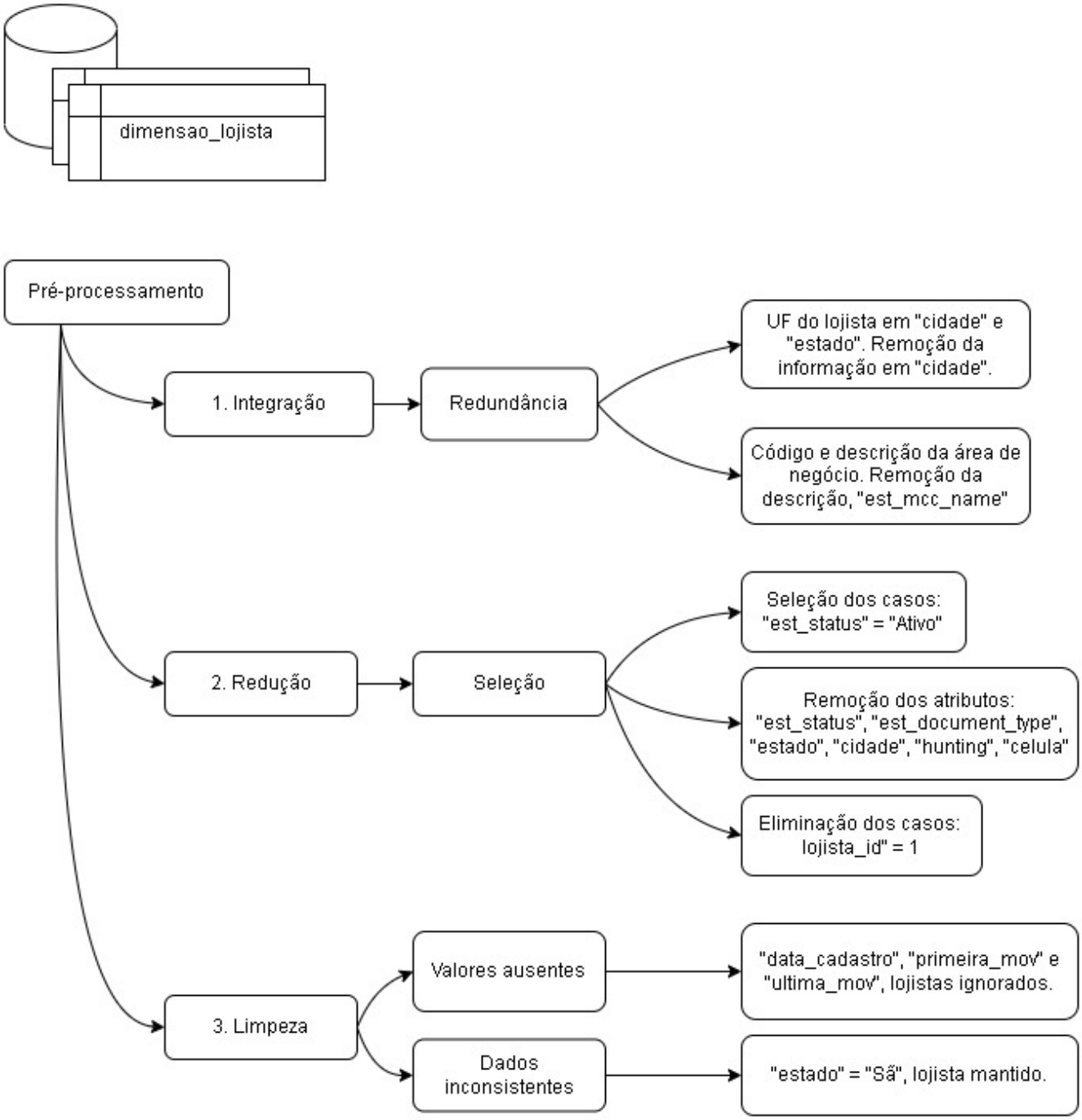
Após as filtragens, restaram 2.743 clientes dos 2.796 obtidos anteriormente pelo pré-processamento. Essa remoção representou cerca de 2% do total.

3.6.1 Aplicação do Teste U de Mann-Whitney

O teste de Mann-Whitney tem o objetivo de comparar tendências centrais de duas amostras independentes de tamanhos iguais. Esse teste é indicado para comparação de dois grupos não pareados, para verificar se pertencem ou não à mesma população. Os valores de U calculados pelo teste avaliam o grau de entrelaçamento dos dados de dois grupos após sua ordenação. A maior separação dos dados em conjunto indica que as amostras são distintas, rejeitando-se a hipótese de igualdade das medianas **[10]**.

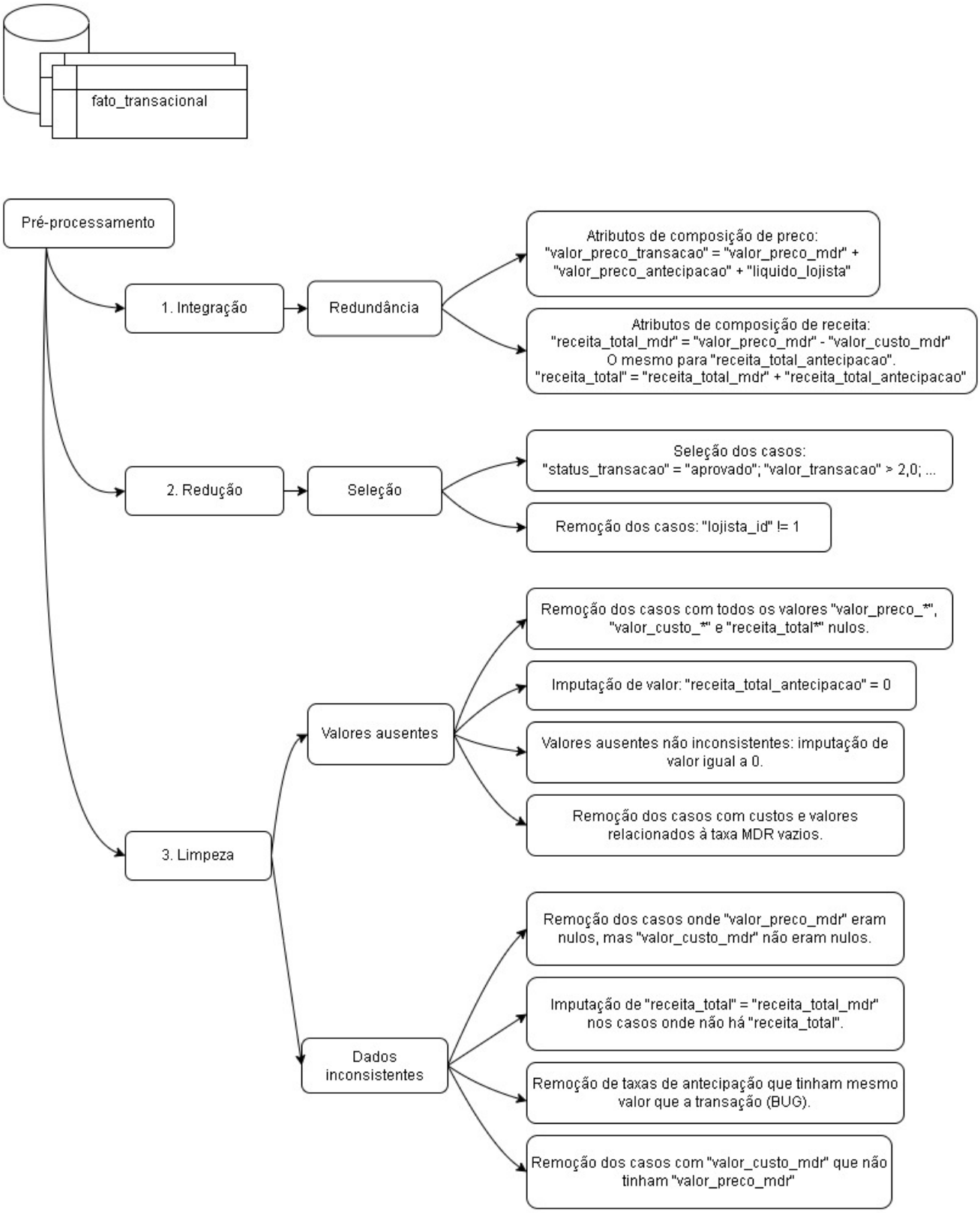
DOI: 10.25286/repa.v7i3.2458

Figura 4 – Fluxograma do pré-processamento da tabela de informações por cliente lojista.



Fonte: Os Autores.

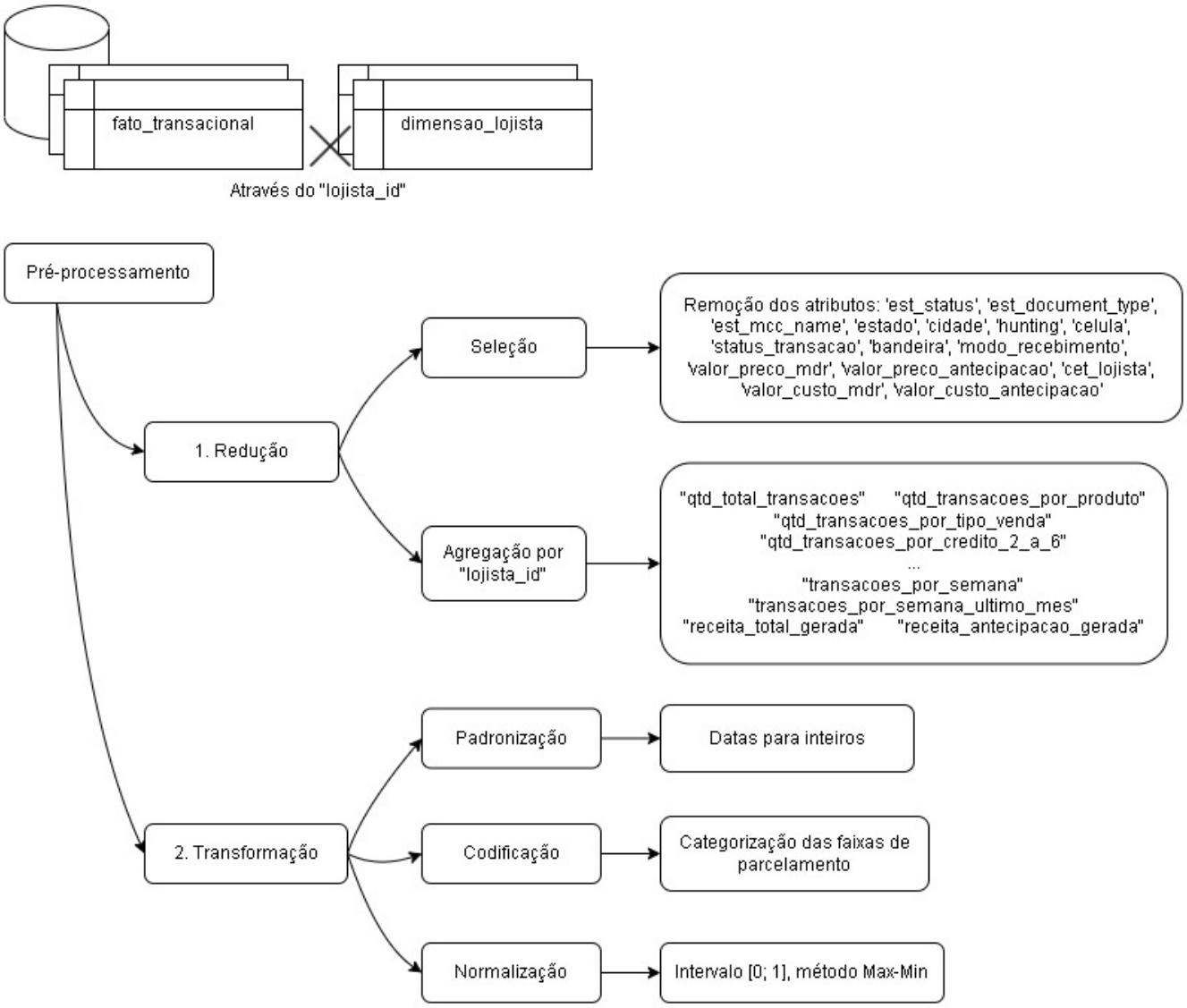
Figura 5 – Fluxograma do pré-processamento da tabela de informações por transação financeira.



Fonte: Os Autores.

DOI: 10.25286/repa.v7i3.2458

Figura 6 – Fluxograma do pré-processamento da tabela de informações por cliente lojista.



Fonte: Os Autores.

No experimento, por exemplo, para a feature "receita_total_gerada", foram comparadas as distribuições de seus valores entre os clientes de dois clusters diferentes. O teste fornece, como resultado, um valor p [0, 1], que caso seja menor que um valor crítico, indica que as distribuições de valores entre as duas populações são diferentes (ou seja, as populações são diferentes). Como função objetivo a ser minimizada pelo PSOC, utilizou-se a soma dos somatórios ponderados dos valores p de cada feature, em relação a cada par de clusters. O peso (nível de importância) das features foi definido pelos stakeholders conforme na Tabela 1.

3.6.1 Aplicação do Teste U de Mann-Whitney

O Particle Swarm Optimization for Clustering (PSOC) é uma adaptação do PSO, utilizando, como partículas, soluções possíveis à clusterização [11]. As partículas são iniciadas aleatoriamente e, a cada iteração, se movem no espaço de busca, através da alteração de seus centros de clusters. Os melhores resultados são avaliados de acordo com uma função objetivo pré-definida.

A forma que o PSOC foi utilizado é como citado em [**11**], alterando sua função objetivo para o cálculo discutido na seção anterior. Sua implementação foi realizada para buscar modelos ótimos tanto para o K-Means quanto para o fuzzy C-Means. Para o K-Medoids, o PSOC foi dispensado porque o cálculo deste algoritmo é

computacionalmente mais custoso que os dois anteriores.

3.6.1 Hiperparâmetros

As execuções realizadas e seus hiperparâmetros estão listadas a seguir. Houve outras execuções com configurações semelhantes, mas não foram exibidas por apresentarem resultados semelhantes ou inferiores aos gerados pelas expostas aqui.

Na primeira execução do K-Means com PSOC, os hiperparâmetros foram:

- Num. clusters: 4
- Inicialização(K-Means): K-Means++
- Máx. iterações (K-Means): 300
- Algoritmo (K-Means): Lloyd
- Num. iterações (PSOC): 100
- Num. partículas: 100
- Velocidade inicial das partículas: 0

Na segunda execução do K-Means com PSOC, com 4 features, os hiperparâmetros foram:

- Num. clusters: 6
- Inicialização(K-Means): K-Means++
- Máx. iterações (K-Means): 300
- Algoritmo (K-Means): Lloyd
- Num. iterações (PSOC): 100
- Num. partículas: 100
- Velocidade inicial das partículas: 0
- Features selecionadas: recencia, receita_total_gerada, tempo_de_cadastro, qtd_total_transacoes.

Na terceira execução foi usado Fuzzy C-Means com PSOC, e os hiperparâmetros foram:

- Num. clusters: 6
- Inicialização (C-Means): Aleatória
- Máx. iterações (C-Means): Não fornecido
- Num. iterações (PSOC): 200
- Num. partículas: 100
- Velocidade inicial das partículas: 0

Na terceira execução foi testado K-Medoids, e os hiperparâmetros foram:

- Num. clusters: 6
- Inicialização(K-medoids): K-medoids++
- Máx. iterações (C-Means): 1.000
- Algoritmo: Partition Around Medoids (PAM)
- Num. iterações (inicializações): 200

4 ANÁLISE E DISCUSSÃO DOS RESULTADOS

Um bom agrupamento é aquele para o qual a variação dentro do agrupamento é a menor possível. Utilizou-se o "Método de Elbow" para avaliar a quantidade de clusters ideal.

Para a primeira e a segunda execução, o Método de Elbow indicou como melhor resultado a separação dos clientes em 11 grupos. Esta quantidade, porém, é considerada inviável para os stakeholders.

Outro método utilizado foi o Gap-statistics. Sua execução não resultou, de forma conclusiva, em um número de clusters ideal. O número mais próximo do ideal foi de 15 clusters, valor este considerado impraticável.

4.1 PRIMEIRA EXECUÇÃO

Após escolher o número de clusters como 4, iniciou-se a procura pelo modelo K-Means que fosse o mais próximo de um ótimo global através do PSOC, juntamente com um método de comparação estatístico de amostras, Mann-Witney, como medida de certeza de que os grupos se diferem [**11**].

Esta execução demonstrou uma divisão perceptível entre o comportamento de cada grupo. Porém, para a feature mais importante para a empresa (receita_total_gerada), os valores foram pouco conclusivos.

4.2 SEGUNDA EXECUÇÃO

Após a verificação dos resultados anteriores, utilizou-se outro método de análise de clustering, Sillhouette. Pode-se observar um aumento no valor, enquanto havia a diminuição dos números de features, levando em consideração seu rankeamento. A Figura 7 exibe os valores de Sillhouette nessa execução.

Tabela 2 – Valores de Sillhouette utilizando uma seleção de 4 features.

N_clusters	Silhouette_score
2	0.5751
3	0.4739
4	0.3884
5	0.4162
6	0.4185
7	0.3422
8	0.3414

Fonte: Os Autores.

DOI: 10.25286/repa.v7i3.2458

Quanto menor a quantidade de grupos, maior é o índice de Sillhouete. A utilização de 2 a 3 grupos é uma classificação que agrega pouca informação para a empresa. A técnica do Gap-Statistic foi executada para o cenário das 4 features, mostrando que um bom agrupamento para esse caso contém 6 grupos.

Houve uma considerável diferença levando-se em conta o tempo de cadastro e a recencia de cada grupo. Já para receita total há uma pequena variação no tamanho das faixas de valores. Na quantidade de transações, um grupo se destaca, abrangendo uma faixa mais alta e diferente dos demais. Esta foi a execução que delineou com maior resolução o limite dos grupos. Porém, por conter apenas 4 features, agrega pouca informação ao modelo.

4.3 TERCEIRA EXECUÇÃO

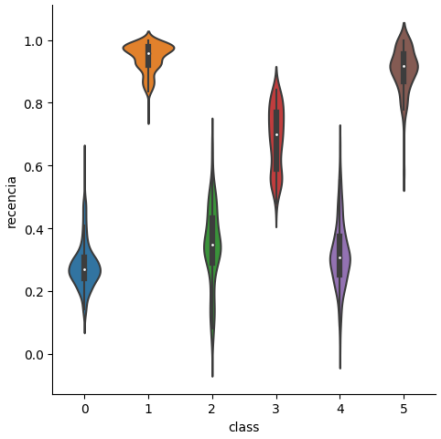
Da terceira execução em diante, todas as features por cliente foram utilizadas para o agrupamento. Para avaliar graficamente o resultado do Fuzzy C-means, uma filtragem foi realizada para manter, como membros de cada cluster, apenas aqueles com porcentagem de pertencimento maior que 50%.

Estes resultados, porém, mostraram pouco significado. Após a filtragem, permaneceram cerca de 5 clientes por cluster. Considerando como pertencente a cada cluster os clientes que possuem a porcentagem de pertencimento maior para aquele cluster, observou-se que há pouca ou nenhuma distinção entre os clusters, tornando o resultado inconclusivo. Todas as tentativas de representar graficamente as características de cada cluster resultaram em uma das situações citadas acima.

4.4 QUARTA EXECUÇÃO

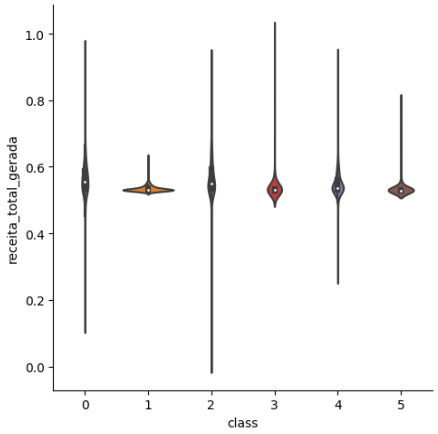
O K-Medoids foi utilizado com a distância Gower, que fornece um índice de similaridade entre os indivíduos. Em vez do PSOC, foram realizadas 200 inicializações e comparados seus valores de inércia, mantendo o menor encontrado. Os gráficos desta execução estão nas figuras de 7 a 10.

Figura 7 – Gráfico violino para "recencia", 4ª execução.



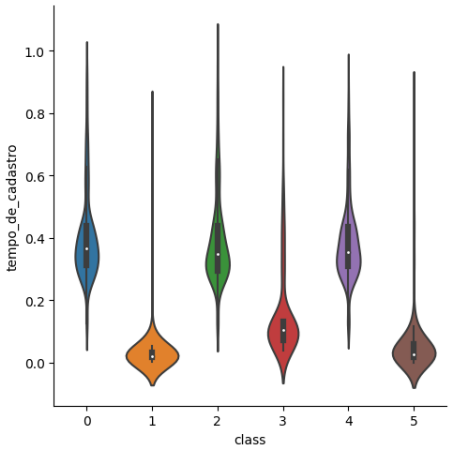
Fonte: Os Autores.

Figura 8 – Gráfico violino para "receita_total_gerada", 4ª execução.



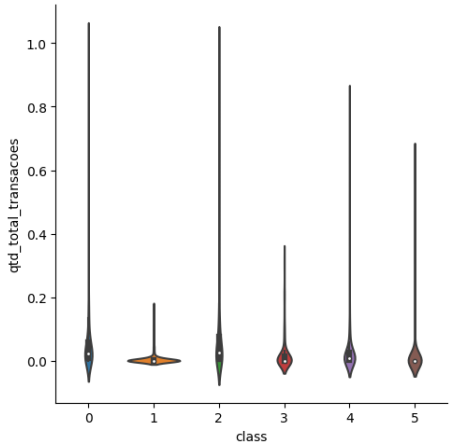
Fonte: Os Autores.

Figura 9 – Gráfico violino para "tempo_de_cadastro", 4ª execução.



Fonte: Os Autores.

Figura 10 – Gráfico violino para "qtd_total_transacoes", 4ª execução.



Fonte: Os Autores.

Os resultados dessa execução foram considerados como os melhores obtidos. Foi possível, ao menos, identificar intervalos de valores para as features mais importantes, mesmo que estas não demonstrem uma divisão clara entre os clusters. A distribuição de clientes por grupo, para essa execução, é: 29% (0), 12% (1), 23% (2), 7% (3), 23% (4) e 6% (5).

Para entender como as features se relacionam, foram gerados heat-maps com todas as features presentes (figuras de 11 a 16).

Logo após a análise desses mapas foram traçados perfis e características que se destacam para cada grupo, mostrados no Quadro 1.

Quadro 1 - Perfis analisados por grupo.

GRUPO	DESCRIÇÃO
0	Apresentaram relações proporcionais entre os valores de NTK.
1	Maior realização de transações por crédito está relacionada a um aumento na receita total do cliente.
2	Não gerou conclusões expressivas únicas para esse grupo
3	Utilizam mais crédito do que débito.
4	Possui comportamento mais distribuído entre os tipos de venda e apresenta valores menores de receita gerada, apesar do volume transacional.
5	Maiores valores de receita gerada não estão, necessariamente, relacionados ao maior uso de crédito ou débito.

Fonte: Os Autores.

5 CONCLUSÕES E TRABALHOS FUTUROS

Observou-se que os clientes da empresa possuem, em sua maioria, comportamentos muito próximos em relação às métricas exploradas neste trabalho.

Uma possível ameaça à validade dos resultados é a quantidade limitada de iterações e inicializações nas execuções que permitiam essas configurações. Valores maiores explorariam mais regiões do espaço de soluções. Apesar disso, a base não apresenta um número significativamente maior de clientes em relação à quantidade de inicializações realizadas, o que pode invalidar essa ameaça.

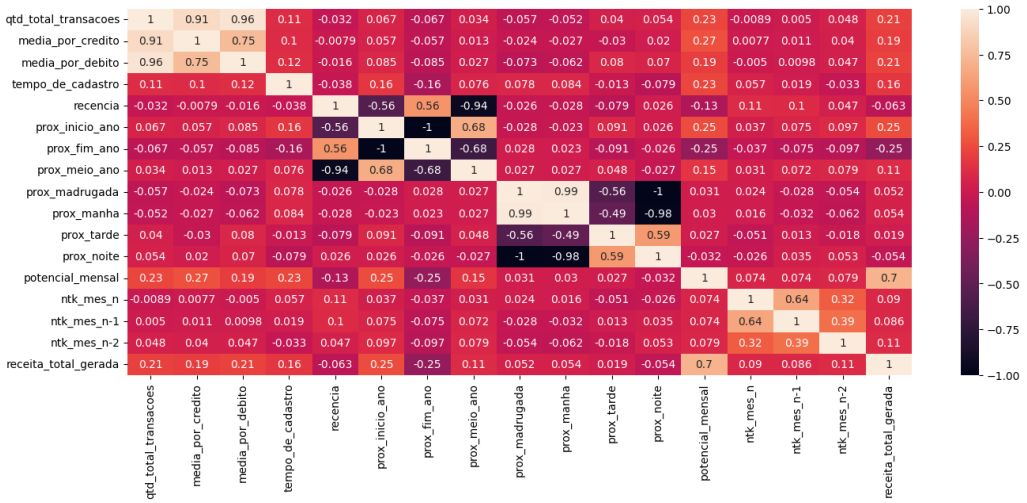
Outra ameaça consiste no período em que os dados fornecidos foram coletados. Não foi possível realizar conclusões sobre a sazonalidade das transações de um cliente porque tal informação exigiria os dados de, no mínimo, dois períodos (2 anos).

Para os stakeholders, os resultados obtidos não forneceram insights suficientes para aplicar os modelos encontrados na rotina de planejamento estratégico da empresa. Apesar disso, houve interesse em entender quais as possíveis relações observadas para posterior refinamento.

Uma continuação para o que foi desenvolvido poderia envolver a testagem de algoritmos de agrupamento diferentes. Também seria necessário obter uma base de dados mais ampla.

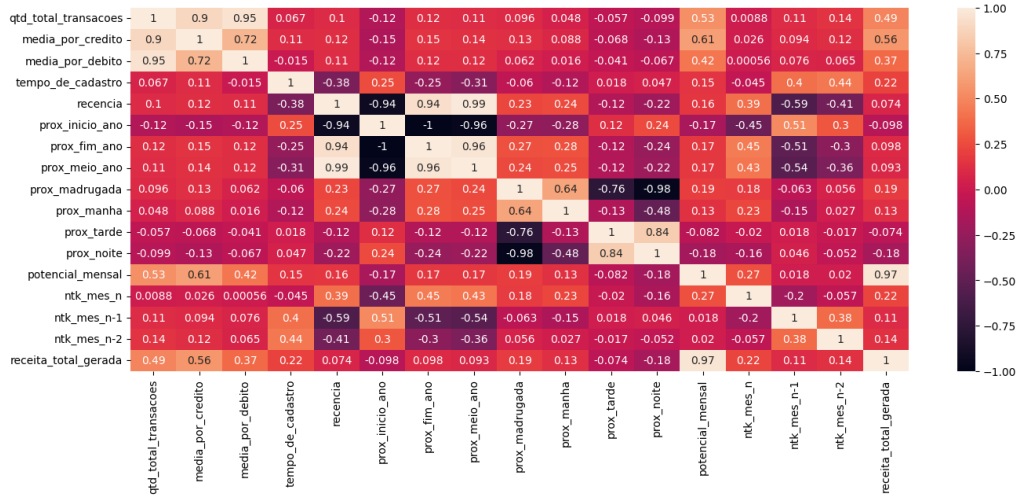
DOI: 10.25286/repa.v7i3.2458

Figura 11 – Heat-map do grupo 0, quarta execução.



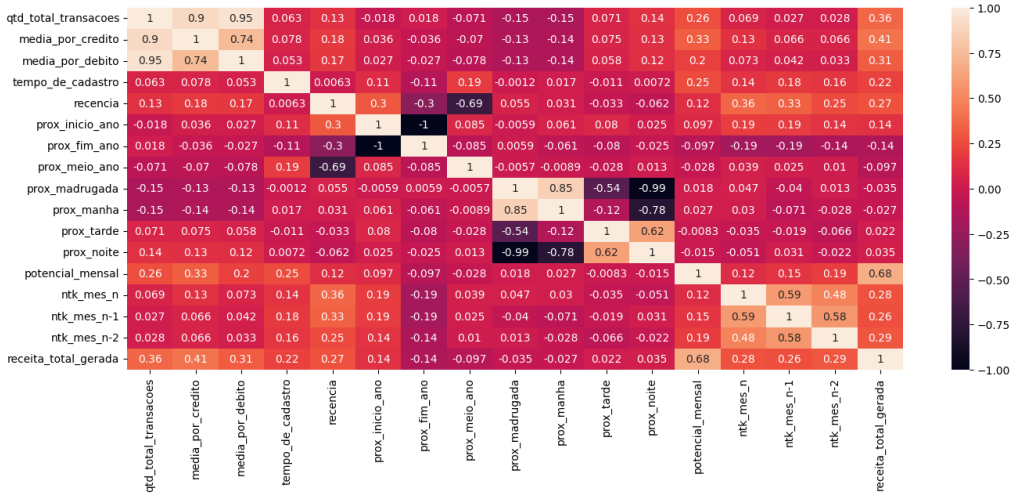
Fonte: Os Autores.

Figura 12 – Heat-map do grupo 1, quarta execução.



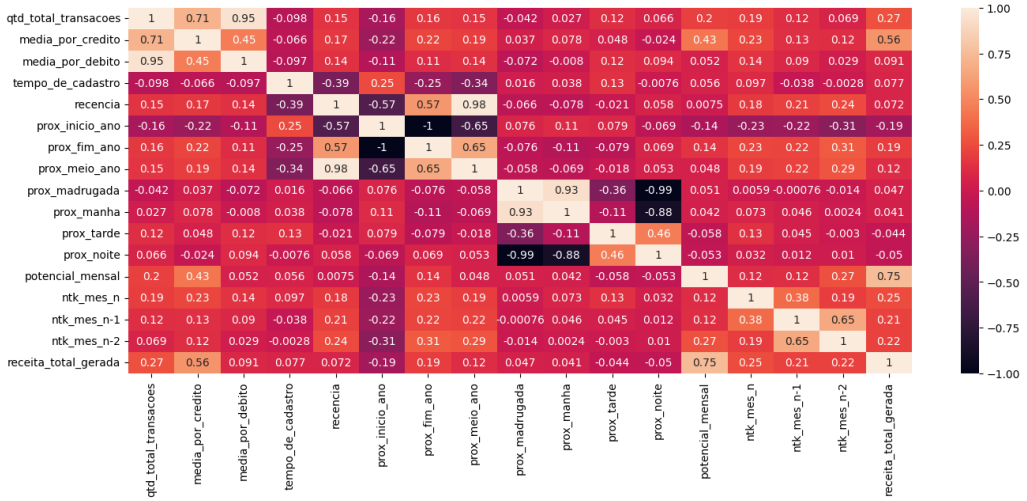
Fonte: Os Autores.

Figura 13 – Heat-map do grupo 2, quarta execução.



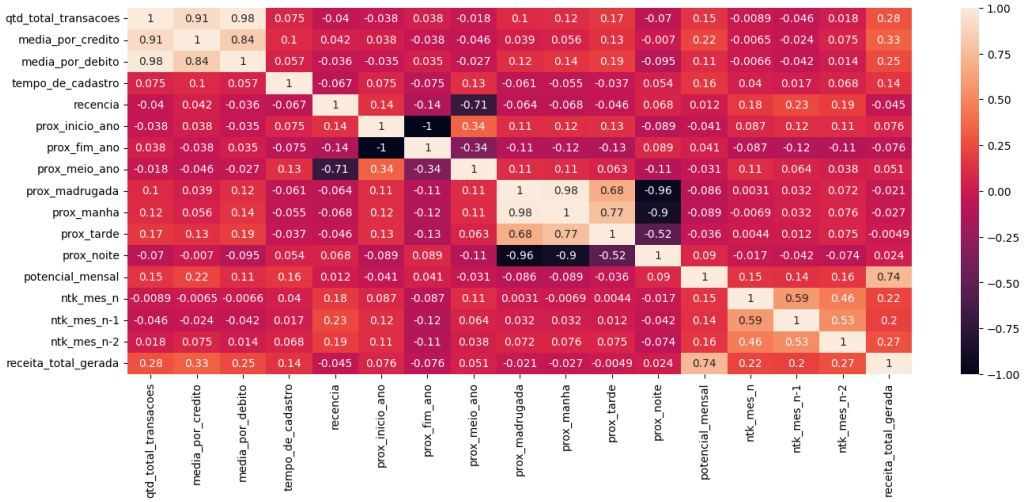
Fonte: Os Autores.

Figura 14 – Heat-map do grupo 3, quarta execução.



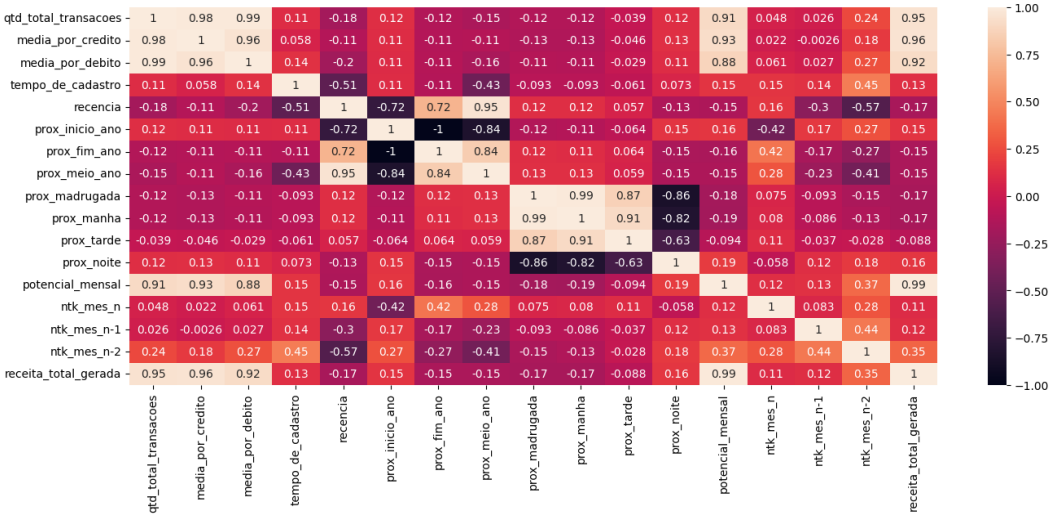
Fonte: Os Autores.

Figura 15 – Heat-map do grupo 4, quarta execução.



Fonte: Os Autores.

Figura 16 – Heat-map do grupo 5, quarta execução.



Fonte: Os Autores.

REFERÊNCIAS

[1] LIU, Hsiang-Hsi; ONG, Chorng-Shyong. Variable selection in clustering for marketing segmentation using genetic algorithms. **Expert systems with applications**, v. 34, n. 1, p. 502-510, 2008.

[2] SHEN, Boyu. E-commerce Customer Segmentation via Unsupervised Machine Learning. In: **The 2nd International Conference on Computing and Data Science**. 2021. p. 1-7.

[3] LIKAS, Aristidis; VLASSIS, Nikos; VERBEEK, Jakob J. The global k-means clustering algorithm. **Pattern recognition**, v. 36, n. 2, p. 451-461, 2003.

[4] JUDSON, Dean. **CLUSTER: Stata module to perform nonhierarchical k-means (or k-medoids) cluster analysis**. 1998.

[5] EVERITT, Brian S. et al. **Cluster analysis: Wiley series in probability and statistics**. Southern Gate, Chichester, West SussexUnited Kingdom: John Wiley & Sons, 2011.

[6] FARHAN, Marwa Salah; ABED, Amira Hassan; ABD ELLATIF, Mahmoud. A systematic review for the determination and classification of the CRM critical success factors supporting with their metrics. **Future Computing and Informatics Journal**, v. 3, n. 2, p. 398-416, 2018.

[7] KAMPANI, Nidhi; JHAMB, Deepika. Analyzing the role of e-crm in managing customer relations: A critical review of the literature. **Journal of Critical Review**, v. 7, n. 4, p. 221-226, 2020.

[8] TEMBHURNE, Durga Sadanand; ADHIKARI, Jayant; BABU, Rajesh. A Review study on Application of Data Mining Techniques in CRM of Pharmaceutical Industry. **International Journal of Scientific Research in Science and Technology**, v. 6, n. 2, p. 1-7, 2019.

[9] MARQUES, Pedro Afonso Bandeira Ferreira et al. **Business clients´ segmentation based on activity: a banking approach**. 2019. Tese de Doutorado.

[10] UMEH, Edith Uzoma et al. Comparison of two sample tests using both relative efficiency and power of test. **Open Journal of Statistics**, v. 6, n. 02, p. 331, 2016.

[11] SANTOS, Pedro et al. Application of PSO-based clustering algorithms on educational databases. In: **2017 IEEE Latin American Conference on Computational Intelligence (LA-CCI)**. IEEE, 2017. p. 1-6.

DOI: 10.25286/repa.v7i3.2458