Aplicações pragmáticas de algoritmos de clusterização em instituições financeiras e bancos de varejo

Osvaldo Luiz dos Santos Pereira[1]*; Thiago Gentil Ramires[2]

[1] Mestre Msc em Física Nuclear (Unicamp). Doutorando em Física Teórica (UFRJ). Pesquisador convidado Instituto de Física UFRJ. Sua Música (suamusica.com.br). Head of Data Science.
[2] Universidade Tecnológica Federal do Paraná. Doutor de ciência. Rua Marcílio Dias, Jardim Paraíso; 86812-460. Apucarana, Paraná, Brasil.
*autor correspondente: osvald23@gmail.com

Aplicações pragmáticas de algoritmos de clusterização em instituições financeiras e bancos de varejo

Resumo

Neste trabalho foi uma apresentada metodologia sobre como reduzir a dimensão de análises no nicho de bancos de varejo através da utilização de algoritmos de clusterização. Foi utilizado o algoritmo K-Means para reduzir a quantidade de milhares de agências de um determinado banco de varejo para apenas seis grupos comparáveis (clusters) para os quais foram aplicados metas de performance semelhantes. Foi demonstrado neste trabalho que o K-means teve uma performance adequada em agrupar as agências em grupos semelhantes de acordo com as variáveis de comportamento e de características. O resultado final apresentou baixa dispersão das variáveis dentro dos clusters e diferenciando estas mesmas variáveis dentre os seis clusters encontrados utilizando o algoritmo de clusterização. Além disto foram demonstradas metodologias de análise dos resultados da clusterização, utilizando análise bivariada entre variáveis e os clusters. Os dados utilizados foram gerados sinteticamente em cumprimento da Lei Geral de Proteção aos Dados (LGPD) a partir de métodos estatísticos e baseados em dados reais de um determinado banco de varejo.

Palavras-chave: K-means; Performance; Estratégia.

Pragmatic applications of clustering algorithms to finance and retail banking

Abstract

In this survey, a methodology was presented on how to reduce the size of analyzes in the retail banking niche through the use of clustering algorithms. The K-Means algorithm was used to reduce the number of thousands of branches of a given retail bank to just six comparable groups (clusters) to which similar performance targets were applied. It was demonstrated in this work that the K-means had an adequate performance in grouping the agencies in similar groups according to the variables of behavior and characteristics, the final result showed low dispersion of the variables within the clusters and differentiating these same variables among the six clusters found using the clustering algorithm. In addition, methodologies for analyzing the results of clustering were demonstrated, using bivariate analysis between variables and clusters. The data used were synthetically generated in compliance with the General Data Protection Law (LGPD) based on statistical methods and based on real data from a given retail bank.

Keywords: K-means; Performance; Strategy.

Introdução

Algoritmos de aglomeração são algoritmos de aprendizado de máquina (*machine learning*) não supervisionados que tem como objetivo agrupar observações de um conjunto de dados utilizando variáveis métricas e calculando distâncias entre eles para definir indicadores de semelhanças entre estas observações*, por exemplo* o *K-Means* desenvolvido por (MacQuenn, 1967). Estes algoritmos foram desenvolvidos desde a década de 1960 com o objetivo de serem utilizados em diversos tipos de aplicações, como por exemplo processamento de sinais (primeira aplicação do algoritmo K-Means), através de métodos de quantização vetorial. Com o crescente desenvolvimento de tecnologias, as aplicações destes tipos de algoritmos se tornaram cada vez mais pragmáticas, tendo inclusive aplicações em

diversos nichos da indústria e do ramo empresarial, como apresentado por Tang et al. (2022), que utiliza diversas técnicas avançadas para criar modelos aplicados em risco de crédito, risco de mercado, otimização de portfólios, mercado financeiro e estratégias de trading. Marques et al. (2019) utilizaram algoritmos de clusterização com o objetivo de desenhar estratégias para identificar modelos de negócio bancários utilizando como variáveis de clusterização características de negócio de diferentes fontes e tipos de atividade envolvendo as agências bancárias analisadas no trabalho. Herrera-Restrepo et al. (2016) também utilizaram métodos de clusterização aliados à análises multivariadas com o objetivo de avaliar e identificar performance operacional de agências bancárias. (Sharahi e Aligholi, 2015) utilizaram algoritmos de clusterização para segmentar clientes de agências bancárias com o intuito de criar estratégias de marketing para cada um dos clusters encontrados no estudo.

Neste trabalho o objeto de estudo da segmentação de agências bancárias do setor de bancos de varejo em grupos com características comparáveis com o objetivo de atribuir metas equivalentes à estas agências de forma que a apuração destas metas possua comparação justa, ou seja, supondo duas agências em Estados diferentes, mas em contextos similares, por exemplo, níveis de transações, quantidade de clientes, renda média dos clientes, tamanho, contexto geográfico e social, possam se encontrar em clusters semelhantes. Este trabalho tem como objetivo principal apresentar aplicações pragmáticas de algoritmos de clusterização em instituições financeiras e bancos de varejo, apresentando o case de redução de quantidade de metas a serem atribuídas para rede de agências, ao invés de serem avaliadas individualmente as agências são atribuídas a grupos homogêneos e comparáveis que recebem conjuntos de metas semelhantes, com isto há redução na dimensionalidade do problema, ou seja, ao invés de avaliar milhares de agências, são utilizados poucos grupos homogêneos.

Foi apresentada uma metodologia de avaliação da qualidade do resultado final da clusterização. Para tal foi utilizada análise bivariada considerando o resultado obtido na utilização do algoritmo K-Means. Esta metodologia consiste em utilizar os valores médios e desvios padrões das variáveis de clusterização, calculando uma matriz de dispersões. Ainda não há na literatura aplicação direta deste método, e um dos objetivos finais é apresentar resultados para criação de benchmarks nos valores de referência para dispersão dos valores das variáveis de clusterização.

Foi demonstrado que métodos de *clusterização* podem ser utilizados como meio de redução dimensional da quantidade de dados para tomada de decisão em bancos de varejo, segmentando as agências da rede bancária em grupos comparáveis em termos de metas, desempenho e características operacionais. Além disso, foi proposta metodologia de análise dos resultados da clusterização através da avaliação das variáveis de clusterização

agrupadas em grupos de variáveis de performance com metodologia de score normalizado para efeito de comparação, e avaliar os grupos de agências encontrados, e comparando seus perfis operacionais, de resultado, características físicas e tamanho, e tipo de negócios. Estas análises são compostas de análises estatísticas bivariadas considerando as variáveis de clusterização justapostas contra os clusters encontrados, e também considerando os agrupamentos de variáveis (métricas de comparação) justapostos contra os clusters. Foram utilizados os valores médios, desvios padrões, e dispersão (razão entre o desvio padrão médio, e a média), e os valores normalizados pelo z-score dentre os clusters.

O case utilizado foi a segmentação de agências bancárias de um determinado banco, em grupos homogêneos e com variáveis de clusterização comparáveis dentro destes grupos, com o objetivo de serem atribuídas metas de performance para cada um dos clusters encontrados, reduzindo assim a quantidade de metas necessárias a serem designadas. Em geral estas metas são definidas por Diretorias de planejamento com foco na rede de agências. Exemplos de possíveis metas seriam: conquista de clientes PF e PJ, venda de produtos bancários, aumento de valores de aporte em produtos de investimento, e tomada de crédito.

Ao lidar com um número muito grande de agências, torna-se uma tarefa muito extensa a atribuição destas cinco mil metas individuais de performance, e mais extenso ainda a apuração mensal destas metas considerando todas as agências da rede bancária, porém com a criação dos clusters estes grupos de agências apresentarão comportamentos e características de negócio muito semelhantes entre si, e os grupos comparáveis poderão ser utilizados ao invés de atribuir cinco mil metas uma para cada agência. As milhares agências serão reduzidas a apenas algumas unidades de grupos comparáveis de agências, ou seja os clusters resultantes do algoritmo de clusterização K-Means. Este trabalho é dividido da seguinte maneira, na segunda seção são apresentados os materiais e métodos utilizados no trabalho, no qual é descrito como foram coletados os dados (sintéticos), as tecnologias (linguagem) utilizadas e as estratégias teóricas de análise de dados utilizadas. Na terceira seção são apresentados os resultados obtidos juntamente com uma discussão dos principais achados e insights encontrados. Na quarta seção são feitas as considerações finais e conclusões sobre o trabalho desenvolvido e os resultados encontrados.

Material e Métodos

A Lei Geral de Proteção de Dados (LGPD) é uma legislação brasileira que tem como objetivo principal a proteção da privacidade e dos dados pessoais dos cidadãos. Promulgada em agosto de 2018 e entrando em vigor em setembro de 2020, a LGPD estabelece diretrizes claras sobre como as organizações devem coletar, armazenar, processar e compartilhar

informações pessoais, visando garantir maior controle e transparência aos indivíduos em relação ao uso de seus dados. A lei também estabelece penalidades para o descumprimento das normas, visando assegurar uma cultura de responsabilidade na gestão de informações sensíveis. Com a LGPD, o Brasil se alinha a tendências internacionais de proteção de dados e busca fortalecer a segurança e a confiança na era digital.

Para este trabalho foram utilizados dados sintéticos criados através de métodos estatísticos. Estes dados foram baseados em dados reais, seguindo as distribuições observadas em projetos reais. O intuito de utilizar dados sintéticos é de seguir as Leis Gerais de Proteção de Dados (LGPD) e cláusulas legais de não divulgação (NDA).

Outro ponto importante é que a demanda por novas metodologias de análise e criação de algoritmos e modelos de aprendizado de máquina está crescendo numa velocidade muito grande, e o acesso à dados cresce na mesma proporção o que causa o aumento do custo ao acesso à estes dados, portanto é comum que artigos metodológicos e até mesmo criação de novas tecnologias utilizem dados sintéticos nos treinamentos de modelos e desenvolvimento de novas metodologias (Raghunathan, 2021; Mahmoud, 2021; Hradec et al., 2022).

Os dados sintéticos foram baseados em dados reais de agências bancárias de um banco de varejo brasileiro, tendo suas distribuição calculadas através de métodos de inferência estatística. Cada uma das observações da base de dados representa uma das agências, e as colunas são variáveis que representam características, e comportamentos de cada uma destas agências, como por exemplo tamanho físico da agência, quantidade de clientes de diversos segmentos, quantidade de transações em caixa humano, quantidade de transações em caixa eletrônico, tipos de negócios realizados tanto em volume quanto em valores.

A base de dados utilizada neste trabalho contém 5.200 observações (linhas) que consistem em agências bancárias, e seus respectivos Ids, e contém cinquenta e três colunas que correspondem às variáveis de *clusterização* utilizadas neste trabalho. Ao final da aplicação do algoritmo K-means é acrescentada a coluna com o Id do cluster (grupo homogêneo) do qual cada agência pertence.

Abaixo são apresentados os nomes e descrições de cada uma das variáveis utilizadas no algoritmo de clusterização.

- Id agência
- NumberTellerCapacity: número de caixas humanos da agência
- NumberManagerPersonalCapacity: número de gerentes que atendem pessoa física
- NumberManagerBusinessCapacity: número de gerentes que atendem pessoa jurídica
- NumberATM: número de caixas eletrônicos

- NumberPersonalClientsTierA: número total de clientes pessoa física do tier A
- NumberPersonalClientsTierB: número total de clientes pessoa física do tier B
- NumberPersonalClientsTierC: número total de clientes pessoa física do tier C
- NumberPersonalClientsTierD: número total de clientes pessoa física do tier D
- NumberINSSClients: número de clientes INSS
- NumberSalaryAccounts: número total de clientes com conta salário
- NBusinessClientsTierA: número de contas pessoa jurídica do tier A
- NBusinessClientsTierB: número de contas pessoa jurídica do tier B
- NBusinessClientsTierC: número de contas pessoa jurídica do tier C
- NBusinessClientsTierD: número de contas pessoa jurídica do tier D
- AvgMonthlyIncomePersonalClientTierA: valor médio de renda mensal de clientes do tier A
- AvgMonthlyIncomePersonalClientTierB: valor médio de renda mensal de clientes tier B
- AvgMonthlyIncomePersonalClientTierC: valor médio de renda mensal de clientes tier C
- AvgMonthlyIncomePersonalClientTierD: valor médio de renda mensal de clientes tier D
- AvgMonthlyINSSBenefitsAmount: valor médio do benefício recebido por clientes INSS
- AvgMonthlySalaryValue: valor médio da renda mensal de clientes pessoa física
- AvgMonthlyEBITDABusinessClientTierA: faturamento mensal pessoa jurídica do tier A
- AvgMonthlyEBITDABusinessClientTierB: faturamento mensal pessoa jurídica do tier B
- AvgMonthlyEBITDABusinessClientTierC: faturamento mensal pessoa jurídica do tier C
- AvgMonthlyEBITDABusinessClientTierD: faturamento mensal pessoa jurídica do tier D
- BranchSizeSquareMeters: tamanho da agência em metros quadrados
- AvgMonthlyATMPaymentsTransactions: quantidade média mensal de transações do tipo pagamento realizados por caixa eletrônico na agência
- AvgMonthlyATMWithdrawTransactions: quantidade média mensal de transações do tipo retirada/recebimentos realizados por caixa eletrônico na agência
- AvgMonthlyATMTransferTransactions: quantidade média mensal de transações do tipo transferências realizados por caixa eletrônico na agência
- AvgMonthlyATMDepositTransactions: quantidade média mensal de transações do tipo depósito bancário por caixa eletrônico realizados na agência
- AvgMonthlyATMTransactions: quantidade média mensal de transações realizadas por caixa eletrônico da agência
- AvgMonthlyTellerPaymentsTransactions: quantidade média mensal de transações do tipo pagamento realizados por caixa humano da agência
- AvgMonthlyTellerWithdrawTransactions: quantidade média mensal de transações do tipo retirada realizados por caixa humano da agência

- AvgMonthlyTellerTransferTransactions: quantidade média mensal de transações do tipo transferência realizadas por caixa humano da agência
- AvgMonthlyTellerDepositTransactions: quantidade média mensal de transações do tipo depósito bancário realizados por caixa humano da agência
- AvgMonthlyTellerTransactions: quantidade média mensal de transações totais realizadas por caixa humano da agência
- AvgMonthlyManagerPersonalLoanTransactions: quantidade média mensal de transações do tipo empréstimo realizadas por gerentes de pessoa física da agência
- AvgMonthlyManagerBusinessLoanTransactions: quantidade média mensal de transações do tipo empréstimos realizadas por gerentes de pessoa jurídica da agência
- AvgMonthlyManagerPersonalInvestmentTransactions: quantidade média mensal de transações do tipo investimento realizadas por gerente de pessoa física da agência
- AvgMonthlyManagerBusinessInvestmentTransactions: quantidade média mensal de transações do tipo investimento realizadas por gerente pessoa jurídica da agência
- AvgMonthlyManagerTransactions: quantidade média mensal do total de transações realizadas por gerente (pessoa física ou jurídica) da agência
- AvgMonthlyRevenueThousands: quantidade média mensal do total do faturamento bruto mensal da agência (x1000)
- AvgMonthlyOperationalCostThousands: quantidade média mensal do total de custo operacional mensal da agência (x1000)
- AvgMonthlyOperationalLossThousands: quantidade média mensal do total de custo por perda operacional mensal da agência (x1000)
- AvgMontlhyEBITDA: EBITDA médio mensal da agência
- AvgMonthlySavingsAccountDeposit: valor médio mensal de valores depositados por conta poupança pessoa física da agência
- AvgMonthlyPersonalLoanAmount: valor médio mensal de valores de empréstimo mensal tomados por conta corrente pessoa física da agência
- AvgMonthlyBusinessLoanAmount: valor médio mensal de valores de empréstimos mensal tomados por conta corrente pessoa jurídica da agência
- AvgMonthlyPersonalCreditCardPaymentAmount: valor médio mensal do valor de pagamento de faturas mensais de cartão de crédito por contas corrente pessoa física
- AvgMonthlyBusinessCreditCardPaymentAmount: valor médio mensal do valor de pagamento de faturas mensais de cartão de crédito por contas corrente pessoa jurídica
- AvgMonthlyPersonalInvestmentsAmount: valor médio mensal de volume aplicado em investimentos por conta corrente pessoa física da agência

- AvgMonthlyBusinessInvestmentsAmount: valor médio mensal de volume aplicado em investimentos por conta corrente pessoa jurídica da agência
- DummyInsideShopping: dummy se a agência está presente em um shopping (1 se estiver dentro de um shopping, 0 se não estiver dentro de um shopping)
- DummyPrimeZone: dummy se a agência se encontra em um bairro classe A (1 se estiver localizada em um bairro classe A, 0 se não estiver localizada em um bairro classe A)

Das variáveis listadas acima, as principais são as variáveis de performance como volume de transações, quantidade de transações, média mensal de aporte em investimentos. As variáveis descritas acima são agregadas em tipos mais gerais de características:

- Size: tamanho das agências, avaliados em tamanho físico (metros quadrados), quantidade de clientes pessoa física e pessoa jurídica (por tier), INSS, contas salário, gerentes pessoa física, gerentes pessoa jurídica, caixa eletrônico, caixa humano. Estas variáveis avaliam a inércia da agência em poder gerar negócios devido ao seu porte.
- Business Potential: é o potencial de negócios que a agência possui, engloba as variáveis do tipo renda declarada/estimada dos clientes pessoa física, faturamento declarado/estimado dos clientes pessoa jurídica, benefício médio mensal de clientes INSS, depósito médio mensal em contas salário. Estas variáveis avaliam o quanto de volume de dinheiro se encontra disponível para captação dentro das agências.
- Financial Health: é a saúde financeira da agência. As variáveis que se encontram neste agrupamento são as médias do faturamento bruto mensal, custos operacionais mensais, custos variáveis mensais, e perdas operacionais da agência. Esta agregação geral avalia a lucratividade da agência, se é consistentemente positiva, ou negativa financeiramente.
- Transaction Volume: avalia o volume de fluxo de dinheiro movimentado na agência. Engloba variáveis como a quantidade média mensal de depósitos em conta poupança, valor médio mensal de empréstimos de pessoas física e jurídica, valores médios mensais de investimentos de conta correntes de pessoas física e jurídica.
- Credit: variáveis que envolvem o quantitativo do risco de crédito da agência. Engloba as variáveis de total de transações de empréstimo realizados por gerentes de contas corrente de pessoa física ou pessoa jurídica, e os valores médios mensais do pagamentos de faturas de cartões de crédito realizados por pessoa física ou pessoa jurídica. Esta agregação avalia a exposição à possíveis perdas por calote em produtos de crédito pessoal ou rotativos.
- Investiment: engloba as variáveis da média do total de transações do tipo investimento realizadas por gerentes de conta corrente de pessoas física ou pessoa jurídica. Esta agregação avalia o potencial em ganho financeiro gerado nas agências através do volume de transações por produtos de investimento.

Para a geração dos dados sintéticos (tabela final de modelagem) foram utilizadas funções estatísticas de geração de números aleatórios com determinada distribuição (normal) da biblioteca numpy (https://numpy.org/) da linguagem Python (https://www.python.org/). As variáveis descritas acima foram geradas em arrays através do sorteio aleatório seguindo uma distribuição normal com média e desvio padrão pré-determinados por cluster e por variável. O dataset final de modelagem foi composto com os arrays sintéticos concatenados por cluster gerado. Cada linha desta tabela representa uma agência específica, com respectivo Id.

Para a geração da variável Cluster, contendo o Id do cluster de cada uma das agências foi utilizado o algoritmo de clusterização K-means. Foi utilizado o método KMeans da biblioteca scikitlearn (https://scikit-learn.org/stable/) da linguagem Python.

As tabelas e dados utilizados para a descrição da metodologia neste trabalho estão organizadas e explicadas no apêndice ao final deste documento, contendo os links de acesso aos arquivos que se encontram em um repositório de GitHub pessoal do autor deste trabalho.

Resultados e Discussão

Como resultado da aplicação do algoritmo de clusterização K-Means na tabela de modelagem foram encontrados no total seis clusters com características distintas. Como proposta de metodologia para averiguar a qualidade da clusterização, e das variáveis utilizadas, foi proposta a construção de uma tabela de contingência contendo os valores de dispersão por cluster de cada uma das variáveis utilizadas no modelo.

A construção da tabela de avaliação se predispõe da seguinte forma, primeiramente são calculadas as médias por variável de cada um dos cluster, e depois uma mesma tabela com os valores de desvio padrão de cada variável por cluster, por fim o cálculo da dispersão se dá pela divisão do desvio padrão pela média de cada uma das variáveis por cluster.

Foi utilizado também a normalização das variáveis pelo z-score, considerando os valores de médias e desvios padrões por cluster para efeito de comparação dentro de uma escala única, facilitando a comparação dos resultados e ordenação das características dentre os clusters. Também foram utilizados grupos macro de variáveis que visam agrupar as variáveis de clusterização em grupos semelhantes de métricas para avaliação,a partir desta agregação é possível calcular os valores médios de dispersão não apenas para as variáveis métricas de clusterização, mas também em agregações de variáveis. O intuito de criar agregações de variáveis é além de uma estratégia técnica para avaliação, mas também uma estratégia de negócios de forma a ter uma visão holística e global sobre um conjunto maior de variáveis de performance e de características das agências.

Estes grupos macro são os seguintes: Business Potential (potencial de realização de negócios), Credit (apetite por risco de crédito e exposição), Financial Health (saúde financeira, resultados e custos), Investiment (perfil de investidor dos clientes do tipo pessoa física e pessoa jurídica das agências), Size (tamanho físico, metragem quadrada média das agências, quantidade de clientes pessoa física e pessoa jurídica, clientes pensionistas, clientes com conta trabalho, quantidade de caixas humanos, caixas eletrônicos, gerentes pessoa física, gerentes pessoa jurídica), Transaction Value (valores médios transacionados pelas pessoas física e jurídica em transações do tipo pagamentos, depósitos, retiradas, transferências, investimentos, e empréstimos), e Transaction Volume (quantidade em número de transações, depósitos, retiradas, pagamentos, investimentos, e empréstimos). Estes grupos agregam seleções de variáveis de clusterização com o objetivo de facilitar a classificação dos clusters em características de performance.

Os resultados desta análise se encontram na Tabela 1, a qual mostra que o cluster 1 (C1) apresenta dispersão média de 11%, o cluster 2 (C2) apresenta dispersão média de 12%, cluster 3 (C3) de 19%, cluster 4 (C4) com 15%, cluster 5 (C5) com 25%, e o cluster 6 (C6) com 19%, ou seja, que em média as observações (agências) possuem uma dispersão média de 17% após a clusterização.

Tabela 1. Valores calculados de dispersão média para cada variável de clusterização.

Features	C1	C2	C3	C4	C5	C6
NumberTellerCapacity	25%	26%	39%	39%	60%	27%
NumberManagerPersonalCapacity	0%	0%	0%	0%	0%	0%
NumberManagerBusinessCapacity	0%	0%	0%	0%	0%	0%
NumberATM	19%	20%	33%	27%	64%	20%
NumberPersonalClientsTierA	20%	26%	32%	33%	71%	32%
NumberPersonalClientsTierB	12%	13%	20%	20%	43%	16%
NumberPersonalClientsTierC	12%	11%	19%	15%	23%	13%
NumberPersonalClientsTierD	10%	10%	16%	13%	20%	10%
NumberINSSClients	10%	10%	16%	13%	19%	10%
NumberSalaryAccounts	12%	11%	20%	19%	41%	12%

Tabela 1. Valores calculados de dispersão média para cada variável de clusterização (Continuação).

Features	C1	C2	C3	C4	C5	C6
NBusinessClientsTierA	6%	10%	11%	11%	23%	11%
NBusinessClientsTierB	9%	9%	14%	15%	17%	9%
NBusinessClientsTierC	9%	9%	14%	12%	17%	9%
NBusinessClientsTierD	11%	11%	19%	16%	24%	12%

AvgMonthlyIncomePersonalClientTierA	7%	7%	11%	9%	14%	14%
AvgMonthlyIncomePersonalClientTierB	10%	9%	16%	13%	19%	20%
AvgMonthlyIncomePersonalClientTierC	13%	14%	20%	18%	27%	27%
AvgMonthlyIncomePersonalClientTierD	21%	20%	33%	28%	40%	39%
AvgMonthlyINSSBenefitsAmount	18%	18%	29%	24%	36%	35%
AvgMonthlySalaryValue	11%	11%	17%	14%	21%	21%
AvgMonthlyEBITDABusinessClientTierA	10%	10%	16%	14%	19%	20%
AvgMonthlyEBITDABusinessClientTierB	20%	20%	33%	26%	42%	41%
AvgMonthlyEBITDABusinessClientTierC	31%	29%	47%	41%	62%	61%
AvgMonthlyEBITDABusinessClientTierD	6%	6%	9%	7%	12%	12%
BranchSizeSquareMeters	10%	13%	16%	13%	21%	10%
AvgMonthlyATMPaymentsTransactions	10%	9%	16%	13%	20%	36%
AvgMonthlyATMWithdrawTransactions	9%	10%	16%	13%	19%	35%
AvgMonthlyATMTransferTransactions	9%	10%	16%	12%	20%	37%
AvgMonthlyATMDepositTransactions	10%	10%	16%	14%	20%	35%
AvgMonthlyATMTransactions	10%	10%	16%	13%	21%	20%
AvgMonthlyTellerPaymentsTransactions	10%	10%	16%	13%	21%	20%
AvgMonthlyTellerWithdrawTransactions	10%	11%	16%	13%	20%	20%
AvgMonthlyTellerTransferTransactions	10%	10%	16%	13%	21%	19%
AvgMonthlyTellerDepositTransactions	10%	10%	16%	13%	21%	20%
AvgMonthlyTellerTransactions	10%	10%	16%	13%	20%	19%
AvgMonthlyManagerPersonalLoanTransactions	10%	10%	16%	13%	20%	13%
AvgMonthlyManagerBusinessLoanTransactions	9%	9%	17%	13%	19%	14%
AvgMonthlyManagerPersonalInvestmentTransactions	11%	20%	16%	13%	21%	14%
AvgMonthlyManagerBusinessInvestmentTransactions	9%	19%	16%	13%	21%	13%
AvgMonthlyManagerTransactions	10%	11%	16%	13%	20%	13%
AvgMonthlyRevenueThousands	10%	10%	13%	13%	20%	14%
AvgMonthlyOperationalCostThousands	9%	11%	35%	14%	19%	10%
AvgMonthlyOperationalLossThousands	10%	10%	36%	13%	20%	10%
AvgMontlhyEBITDA	10%	20%	14%	14%	20%	10%
AvgMonthlySavingsAccountDeposit	10%	10%	16%	14%	20%	12%
AvgMonthlyPersonalLoanAmount	10%	10%	16%	14%	20%	14%
AvgMonthlyBusinessLoanAmount	9%	11%	16%	12%	19%	13%
AvgMonthlyPersonalCreditCardPaymentAmount	10%	9%	16%	13%	21%	16%
AvgMonthlyBusinessCreditCardPaymentAmount	11%	10%	16%	14%	20%	16%
AvgMonthlyPersonalInvestmentsAmount	10%	20%	16%	14%	21%	19%
AvgMonthlyBusinessInvestmentsAmount	10%	20%	16%	13%	19%	15%

Fonte: Dados originais da pesquisa.

Como objetivo deste trabalho é propor uma nova metodologia de avaliação de clusters, ainda não existem benchmarks para efeito de comparação, mas é possível inferir valores que sejam razoáveis para justificar a utilização de um resultado pós K-Means. Neste trabalho foi proposto o valor de até 20% como média geral, e até 25% como média da dispersão por cluster.

A variável Number Teller Capacity apresentou valores consideravelmente maiores que a mediana, principalmente para os clisters C3, C4, e C5. Outras variáveis que apresentaram valores altos de dispersão foram Number Personal Clients Tier A (C3: 32%, C4: 33%, C5: 71%, C6: 32%), Number Salary Accounts (C5: 41%), Avg Monthly Income Personal Client Tier D (C3: 33%, C5: 40%, C6: 39%), Avg Monthly EBITDA Business Client Tier C (C1: 31%, C3: 47%, C4: 41%, C5: 62%, C6, 61%). Como exemplo de baixíssima dispersão tem-se as variáveis Number Manager Personal Capacity, e Number Manager Business Capacity que apresentaram 0% de dispersão dentre os clusters criados.

As variáveis que apresentaram maiores dispersões, considerando a média dentre todos os clusters foram as seguintes: AvgMonthlySalaryValue (27%), NumberATM (31%), AvgMonthlyIncomePersonalClientTierD (30%), AvgMonthlyEBITDABusinessClientTierB (30%), NumberTellerCapacity (35%), NumberPersonalClientsTierA (36%) e AvgMonthlyEBITDABusinessClientTierC (45%).

Algumas variáveis apresentam valores de dispersão em média maiores do que as outras, fato este que evidencia a diferenciação dentre os clusters. Por exemplo, a variável AvgMonthlySalaryValue que representa o valor médio da renda mensal de clientes pessoa física na agência que apresenta dispersão de 27%. Outra variável que apresentou valores de dispersão significativamente maiores foi a NumberATM que representa a quantidade de caixas eletrônicos dentro das agências. Esta diferenciação é esperada dado a natureza física e geográfica da rede de agências. Agências maiores, localizadas em grandes centros como avenidas movimentadas de capitais precisam ter um número maior de caixas eletrônicos dimensionado para suprir a demanda local dos clientes que passam pela agência, enquanto que agências menores localizadas em municípios menores não necessitam de um número grande de caixas eletrônicos.

Na Tabela 2 é possível ver as variáveis de clusterização com valores médios de dispersão acima da mediana calculada para todas as variáveis. Estas variáveis apresentam variabilidade considerável dentro dos clusters, ajudando na diferenciação.

Tabela 2. Valores médios de dispersão calculados por variável de clusterização, sem considerar a segmentação por clusters.

Features	Avg Weighted Dispersion	Avg Dispersion
AvgMonthlyEBITDABusinessClientTierC	48%	45%
NumberTellerCapacity	37%	36%

NumberPersonalClientsTierA	34%	36%
AvgMonthlyEBITDABusinessClientTierB	33%	30%
AvgMonthlyIncomePersonalClientTierD	32%	30%
NumberATM	31%	31%
AvgMonthlyINSSBenefitsAmount	29%	27%
AvgMonthlyOperationalLossThousands	24%	16%
AvgMonthlyOperationalCostThousands	24%	16%
NumberPersonalClientsTierB	21%	21%
AvgMonthlyIncomePersonalClientTierC	21%	20%
NumberSalaryAccounts	19%	19%
AvgMonthlyATMTransferTransactions	18%	17%
AvgMonthlyATMDepositTransactions	18%	17%
AvgMonthlyATMPaymentsTransactions	18%	17%
AvgMonthlyATMWithdrawTransactions	18%	17%
NumberPersonalClientsTierC	17%	16%
AvgMonthlySalaryValue	17%	16%
NBusinessClientsTierD	17%	15%
AvgMonthlyPersonalInvestmentsAmount	16%	17%
AvgMonthlyTellerDepositTransactions	16%	15%
AvgMonthlyATMTransactions	16%	15%
AvgMonthlyTellerPaymentsTransactions	16%	15%
AvgMonthlyEBITDABusinessClientTierA	16%	15%
AvgMonthlyTellerTransactions	16%	15%

Fonte: Dados originais da pesquisa.

Na Tabela 3 são apresentadas as variáveis de clusterização com valores médios de dispersão calculados ponderação pela quantidade de agências por clusters, e pela média simples. As cinco variáveis que apresentaram os maiores valores de dispersão média foram:

- *AvgMonthlyEBITDABusinessClientTierC*,
- *NumberTellerCapacity,*
- *NumberPersonalClientsTierA,*
- *AvgMonthlyEBITDABusinessClientTierB,*
- *AvgMonthlyIncomePersonalClientTierD.*

As cinco variáveis que apresentaram os menores valores de dispersão foram:

- AvgMonthlyTellerTransferTransactions,
- AvgMonthlyBusinessInvestmentsAmount,
- AvgMonthlyIncomePersonalClientTierB,
- AvgMonthlyTellerWithdrawTransactions,
- AvgMonthlyManagerPersonalInvestmentTransactions

A partir dos dados mostrados na Tabela 3 é possível concluir que o resultado da clusterização se mostrou com valores médios de 13% de dispersão, um valor que pode ser utilizado como *benchmark* para futuros trabalhos envolvendo clusterização.

Tabela 3. A Valores médios de dispersão calculados por variável de clusterização, sem considerar a segmentação por clusters.

Features	Avg Weight. Dispersion	Avg Dispersion
AvgMonthlyTellerTransferTransactions	16%	15%
AvgMonthlyBusinessInvestmentsAmount	16%	16%
AvgMonthlyIncomePersonalClientTierB	16%	15%
AvgMonthlyTellerWithdrawTransactions	16%	15%
AvgMonthlyManagerPersonalInvestmentTransactions	16%	16%
AvgMonthlyManagerBusinessInvestmentTransactions	15%	15%
AvgMonthlyPersonalCreditCardPaymentAmount	15%	14%
AvgMonthlyBusinessCreditCardPaymentAmount	15%	14%
AvgMonthlyPersonalLoanAmount	15%	14%
AvgMonthlyManagerBusinessLoanTransactions	15%	14%
AvgMonthlyManagerTransactions	15%	14%
AvgMonthlyManagerPersonalLoanTransactions	15%	14%
AvgMonthlyBusinessLoanAmount	15%	14%
AvgMonthlySavingsAccountDeposit	15%	14%
NumberPersonalClientsTierD	14%	13%
BranchSizeSquareMeters	14%	14%
NumberINSSClients	14%	13%
AvgMontlhyEBITDA	14%	15%
AvgMonthlyRevenueThousands	14%	13%
NBusinessClientsTierB	13%	12%
NBusinessClientsTierC	13%	12%
NBusinessClientsTierA	12%	12%
AvgMonthlyIncomePersonalClientTierA	11%	10%
AvgMonthlyEBITDABusinessClientTierD	9%	9%
NumberManagerPersonalCapacity	0%	0%
NumberManagerBusinessCapacity	0%	0%

Fonte: Dados originais da pesquisa.

A Tabela 4 apresenta os valores calculados de dispersão médios por cluster e por agrupamento de variáveis. Os clusters apresentaram as seguintes médias de dispersão C1 (11%), C2 (13%), C3 (19%), C4 (15%), C5 (23%), C6 (17%). Como valores acima de referência tem-se a variáveis Business Potential, Financial Health, e a variável Size. Para valores abaixo da referência tem-se as variáveis Credit, Financial Health, Investiment, Transaction Value, e a variável Transaction Volume. Os grupos de variáveis que apresentaram valores de dispersão menores que os valores de referência são Credit, Size, Transaction Value, Transaction Volume, enquanto que as variáveis que tiveram valores acima da referência são Business Potential, e Financial Health. Os clusters apresentaram as seguintes médias de dispersão C1 (11%), C2 (13%), C3 (19%), C4 (15%), C5 (23%), C6 (17%). Para valores muito abaixo da referência tem-se a variável Credit nos clusters C1, e C2,

a variável Financial Health no cluster C1, a variável Investiment no cluster C1, a variável Transaction Value no cluster C1, e a variável Transaction Volume nos clusters C1, e C2.

Tabela 4. Valores calculados de dispersão médios calculados por cluster e por agrupamento de variáveis de clusterização.

Feature Grouping	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
BusinessPotential	15%	14%	23%	19%	29%	29%
Credit	10%	10%	16%	13%	20%	15%
FinancialHealth	10%	13%	24%	13%	20%	11%
Investment	10%	20%	16%	13%	21%	13%
Size	11%	12%	18%	17%	30%	13%
TransactionValue	10%	14%	16%	13%	20%	15%
TransactionVolume	10%	10%	16%	13%	20%	25%

Fonte: Dados originais da pesquisa.

A Tabela 5 apresenta os valores médios de dispersão calculados por agrupamento de variáveis de forma geral sem considerar os valores por cluster. Os grupos de variáveis que apresentaram valores de dispersão menores que os valores de referência (benchmark) são Credit, Size, Transaction Value, Transaction Volume, enquanto que as variáveis que tiveram valores acima da referência (benchmark) são Business Potential, e Financial Health.

*Tabela 5. V*alores médios de dispersão calculados por agrupamento de variáveis.

Feature Grouping	Avg Weighted Dispersion	Avg Dispersion
BusinessPotential	23%	22%
Credit	15%	14%
FinancialHealth	19%	15%
Investment	15%	15%
Size	17%	17%
TransactionValue	15%	15%
TransactionVolume	17%	16%

Fonte: Dados originais da pesquisa.

A primeira coluna Avg Weighted Dispersion calcula a dispersão média considerando a ponderação pela quantidade de agências por cluster, enquanto a coluna Avg Dispersion é a média simples calculada sem considerar ponderação. Os grupos de variáveis que apresentaram valores de dispersão menores que os valores de referência (benchmark) são

Credit, Size, Transaction Value, Transaction Volume, enquanto que as variáveis que tiveram valores acima da referência (benchmark) são Business Potential, e Financial Health.

A Tabela 6 apresenta os valores médios calculados por cluster, das variáveis de clusterização. Alguns padrões observados são de que o cluster C1 apresenta em média quantidades maiores de clientes do tipo Pessoa Física, enquanto as agências do cluster C5 apresentam menores quantidades de clientes do tipo Pessoa Física, isto pode ser confirmado observando os valores das variáveis Number Personal Clients Tier A, Number Personal Clients Tier B, Number Personal Clients Tier C, e Number Personal Clients Tier D. O mesmo pode ser observado para a quantidade de clientes beneficiários de INSS, conforme pode ser observado através dos valores da variável Number INSS Clients.

Em relação ao tamanho físico das agências podemos observar os valores da variável Branch Size Square Meters. Os clusters C1, e C6 apresentam valores de metragem quadrada acima da mediana (agências grandes), enquanto o cluster C5 apresenta valores de metragem quadrada bem abaixo da mediana (agências pequenas), e os clusters C2, C3, C4 apresentam valores de metragem quadrada próximo do valor da mediana (agências médias).

O cluster C1 apresenta valores de EBITDA para contas pessoa física maiores que os valores da mediana calculada para todos os clusters, o que evidencia que as agências deste cluster apresentam maior potencial de realizar negócios com seus clientes, além de apresentar um maior equilíbrio financeiro de suas contas. O mesmo pode ser observado para as agências do cluster C2, que também apresentam valores consideravelmente altos de EBITDA. O cluster C5 apresenta valores de EBITDA para contas pessoa física abaixo do valor da mediana.

O cluster C6 apresenta níveis de transações no caixa eletrônico abaixo da mediana, enquanto os clusters C1, e C2 apresentam altos níveis de transação no caixa eletrônico. Considerando as transações de depósito, saque, pagamento, e transferência bancária.

A mesma configuração se apresenta nos níveis de transações no caixa humano, nos negócios realizados com os gerentes de conta, fato este que leva a crer que os clusters C1, e C2 são de agências que apresentam um grande fluxo de negócios, clientes e transações (alto fluxo).

Tabela 6. Valores médios calculados por cluster, e por variáveis de clusterização. Os clusters apresentam as seguintes quantidades de agências C1 (400), C2 (300), C3 (2500), C4 (700), C5 (500), C6 (800).

Features	Cluster 1 (Média)	Cluster 2 (Média)	Cluster 3 (Média)	Cluster 4 (Média)	Cluster 5 (Média)	Cluster 6 (Média)
NumberTellerCapacity	8	8	5	5	2	8
NumberManagerPersonalCapacity	3	3	2	2	1	3
NumberManagerBusinessCapacity	3	3	2	2	1	3
NumberATM	10	10	6	7	3	10
NumberPersonalClientsTierA	1.586	1.209	1.006	977	443	995
NumberPersonalClientsTierB	3.195	3.185	2.005	2.013	900	2.388
NumberPersonalClientsTierC	3.983	4.036	2.495	3.016	1.967	3.986
NumberPersonalClientsTierD	5.554	5.538	3.493	4.223	2.833	5.556
NumberINSSClients	1.616	1.594	995	1.201	800	1.599
NumberSalaryAccounts	2.026	2.002	1.251	1.259	557	2.007
NBusinessClientsTierA	120	75	75	75	34	75
NBusinessClientsTierB	351	351	220	221	176	351
NBusinessClientsTierC	717	719	452	540	361	721
NBusinessClientsTierD	1.365	1.377	848	1.025	677	1.362
AvgMonthlyIncomePersonalClientTierA	7.166	7.158	4.494	5.401	3.579	3.610
AvgMonthlyIncomePersonalClientTierB	4.816	4.801	3.007	3.633	2.390	2.403
AvgMonthlyIncomePersonalClientTierC	3.597	3.554	2.250	2.688	1.808	1.780
AvgMonthlyIncomePersonalClientTierD	2.751	2.822	1.737	2.089	1.430	1.408
AvgMonthlyINSSBenefitsAmount	3.125	3.134	1.953	2.353	1.541	1.581
AvgMonthlySalaryValue	4.518	4.476	2.799	3.373	2.239	2.250
AvgMonthlyEBITDABusinessClientTierA	798.971	795.564	499.852	601.541	395.625	400.210
AvgMonthlyEBITDABusinessClientTierB	511.614	521.253	319.601	383.358	247.501	257.656
AvgMonthlyEBITDABusinessClientTierC	234.016	235.871	150.661	180.402	118.902	118.179
AvgMonthlyEBITDABusinessClientTierD	140.693	140.278	87.845	105.593	70.371	70.210
BranchSizeSquareMeters	805	601	502	599	396	800
AvgMonthlyATMPaymentsTransactions	39.790	40.054	25.045	29.838	20.211	11.320
AvgMonthlyATMWithdrawTransactions	32.125	32.004	20.073	24.120	16.178	9.202
AvgMonthlyATMTransferTransactions	23.988	23.959	15.001	18.054	12.030	6.742
AvgMonthlyATMDepositTransactions	30.270	30.528	18.949	22.674	15.039	8.744
AvgMonthlyATMTransactions	127.229	126.952	79.428	95.167	62.548	63.846
AvgMonthlyTellerPaymentsTransactions	31.954	31.913	19.979	24.026	15.684	15.913
AvgMonthlyTellerWithdrawTransactions	28.780	28.879	18.097	21.618	14.400	14.483
AvgMonthlyTellerTransferTransactions	24.007	23.710	14.974	18.048	11.927	12.099
AvgMonthlyTellerDepositTransactions	39.683	39.545	24.840	30.008	19.976	20.011
AvgMonthlyTellerTransactions	126.287	123.783	77.713	94.124	62.721	62.457
AvgMonthlyManagerPersonalLoanTransactions	323	319	200	241	162	240
AvgMonthlyManagerBusinessLoanTransactions	480	482	301	359	240	363
AvgMonthlyManagerPersonalInvestmentTransactions	159	79	100	120	79	120

Tabela 6 – Valores médios calculados por cluster, e por variáveis de clusterização. Os clusters apresentam as seguintes quantidades de agências C1 (400), C2 (300), C3 (2500), C4 (700), C5 (500), C6 (800). (continuação)

Features	Cluster 1 (Média)	Cluster 2 (Média)	Cluster 3 (Média)	Cluster 4 (Média)	Cluster 5 (Média)	Cluster 6 (Média)
AvgMonthlyManagerBusinessInvestmentTransactions	238	120	150	178	120	181
AvgMonthlyManagerTransactions	1.203	1.187	754	899	596	904
AvgMonthlyRevenueThousands	879.283	876.584	662.031	657.211	442.454	662.513
AvgMonthlyOperationalCostThousands	239.104	241.601	67.424	179.663	120.175	240.248
AvgMonthlyOperationalLossThousands	80.660	81.093	22.529	60.066	39.405	80.205
AvgMontlhyEBITDA	556.841	280.023	420.083	418.740	282.672	558.893
AvgMonthlySavingsAccountDeposit	875	881	549	662	434	662
AvgMonthlyPersonalLoanAmount	1.785	1.824	1.124	1.354	892	1.336
AvgMonthlyBusinessLoanAmount	39.720	39.972	24.962	30.300	20.133	30.075
AvgMonthlyPersonalCreditCardPaymentAmount	1.281	1.270	798	962	646	796
AvgMonthlyBusinessCreditCardPaymentAmount	7.988	7.996	5.043	6.014	4.003	5.050
AvgMonthlyPersonalInvestmentsAmount	1.928	958	1.190	1.429	958	972
AvgMonthlyBusinessInvestmentsAmount	15.971	8.074	10.067	12.029	7.991	10.022

Fonte: Dados originais da pesquisa.

A Tabela 7 mostra os valores médios normalizados através do z-score dos agrupamentos de tipos de variáveis de clusterização e por cluster. Os valores das variáveis foram normalizados calculando-se o z-score por variável em cada um dos clusters, utilizando os valores de média e desvio padrão por variável. Os valores médios para cada variável se encontram na tabela 9, enquanto os valores médios ponderados por grupo de variável, e o valor médio da dispersão por grupo de variável se encontram na tabela 8. Os valores normalizados podem ser utilizados como nota (escore) para diferenciar as características dentre cada um dos clusters. Por exemplo, a tabela evidencia que o Cluster 1 apresentou valores destacados em todos os agrupamentos, ou seja as agências que aparecem neste cluster possuem alto potencial de negócios (Business Potential = 1,43), são bastante voltadas à crédito (Credit = 1,51), possuem saúde financeira mais alta (Finantial Health = 1,24), possuem muitos clientes com altos investimentos (Investment = 2,12), possuem agências grandes em tamanho físico e em quantidade de clientes (Size = 1,29), são agências com bastante fluxo de negócios tendo altos níveis de transações em caixa eletrônico, caixa humano e atendimentos por gerentes comerciais (Transaction Volume = 1,43), e possuem altos valores transacionados (Transaction Value = 1,83).

O cluster 1 apresenta os maiores valores de agregação de variáveis, enquanto que o cluster 3 apresenta os menores valores. É possível concluir que o cluster 1 apresenta agências com perfis mais competitivos dentro da rede, são agências maiores, com maior volume de transações, valores mais altos de transações, maior potencial de negócios, possuem um perfil diversificado de clientes com valores altos de investimento, e com apetite acima da média para crédito, devido à essas características apresenta uma saúde financeira bem acima da média em relação às agências dos outros clusters.

Tabela 7. Valores médios de z-score dos agrupamentos de variáveis de clusterização.

Feature Grouping	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
BusinessPotential	1,43	1,44	-0,55	0,12	-1,23	-1,22
Credit	1,51	1,49	-0,85	-0,07	-1,63	-0,44
FinancialHealth	1,24	0,51	-1,03	-0,10	-1,42	0,80
Investment	2,12	-1,29	-0,42	0,43	-1,29	0,45
Size	1,29	0,86	-0,68	-0,37	-1,92	0,83
TransactionValue	1,83	0,49	-0,67	0,19	-1,51	-0,33
TransactionVolume	1,43	1,41	-0,50	0,15	-1,15	-1,34

Fonte: Dados originais da pesquisa.

A Tabela 8 apresenta os valores de média ponderada calculada para cada um dos agrupamentos de variáveis de clusterização, e valores de dispersão média para cada um dos grupos. Os valores médios ponderados têm como peso (ponderador) a quantidade de agências por cluster, e foram calculados depois da normalização dos valores pelo z-score. Os valores médios de dispersão foram calculados considerando a média ponderada e o desvio padrão sem considerar a ponderação pela quantidade de agências por cluster.

Tabela 8 – Valores de média ponderada calculada para cada um dos agrupamentos de variáveis de clusterização, e valores de dispersão média para cada um dos grupos.

Feature Grouping	Weighted Average	Average Dispersion
BusinessPotential	-0,36	0,28
Credit	-0,44	0,21
FinancialHealth	-0,40	0,26
Investment	-0,11	0,21
Size	-0,29	0,26
TransactionValue	-0,32	0,21
TransactionVolume	-0,35	0,27

Fonte: Dados originais da pesquisa.

A agregação de variáveis que apresentou menores valores de dispersão foi a Investiment, enquanto a agregação com maior valor de dispersão foi a Business Potential, e o valor médio de dispersão para todas as agregações é de 0,24. Um fato que chama atenção é de que todas as agregações tiveram valores médios ponderados negativos, o que pode significar que em média as agências da rede como um todo apresentam indicadores de performance ruins quando comparados com as agências benchmark (cluster 1).

A Tabela 9 apresenta os valores médios normalizados pela transformação de z-score para cada uma das variáveis de clusterização, e para cada um dos clusters resultantes do

algoritmo K-Means encontrados para as agências. A normalização das variáveis foi considerada nas linhas, e o z-score calculado considerando o valor do desvio padrão médio para os seis clusters, e a média ponderada foi calculada considerando como peso os valores da quantidade de agências em cada um dos clusters.

Tabela 9. Valores normalizados pela transformação de z-score para cada uma das variáveis de clusterização, para cada um dos clusters.

Feature Grouping	Features	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
BusinessPotential	AvgMonthlyIncomePersonalClientTierA	1,44	1,44	-0,55	0,12	-1,24	-1,21
BusinessPotential	AvgMonthlyIncomePersonalClientTierB	1,44	1,42	-0,55	0,14	-1,23	-1,22
BusinessPotential	AvgMonthlyIncomePersonalClientTierC	1,48	1,41	-0,54	0,11	-1,21	-1,25
BusinessPotential	AvgMonthlyIncomePersonalClientTierD	1,38	1,52	-0,59	0,10	-1,18	-1,23
BusinessPotential	AvgMonthlyINSSBenefitsAmount	1,43	1,45	-0,56	0,12	-1,26	-1,19
BusinessPotential	AvgMonthlySalaryValue	1,47	1,42	-0,56	0,12	-1,23	-1,21
BusinessPotential	AvgMonthlyEBITDABusinessClientTierA	1,45	1,42	-0,55	0,13	-1,24	-1,21
BusinessPotential	AvgMonthlyEBITDABusinessClientTierB	1,40	1,50	-0,55	0,10	-1,28	-1,18
BusinessPotential	AvgMonthlyEBITDABusinessClientTierC	1,39	1,44	-0,51	0,17	-1,24	-1,25
BusinessPotential	AvgMonthlyEBITDABusinessClientTierD	1,45	1,43	-0,56	0,12	-1,22	-1,23
Credit	AvgMonthlyManagerPersonalLoanTransactions	1,54	1,46	-0,96	-0,13	-1,75	-0,15
Credit	AvgMonthlyManagerBusinessLoanTransactions	1,48	1,52	-0,95	-0,16	-1,78	-0,11
Credit	AvgMonthlyPersonalCreditCardPaymentAmount	1,52	1,47	-0,76	0,02	-1,47	-0,77
Credit	AvgMonthlyBusinessCreditCardPaymentAmount	1,50	1,50	-0,74	0,00	-1,53	-0,73
FinancialHealth	AvgMonthlyRevenueThousands	1,51	1,49	-0,29	-0,33	-2,10	-0,28
FinancialHealth	AvgMonthlyOperationalCostThousands	0,98	1,02	-1,93	-0,03	-1,04	1,00
FinancialHealth	AvgMonthlyOperationalLossThousands	1,00	1,02	-1,91	-0,03	-1,06	0,98
FinancialHealth	AvgMontlhyEBITDA	1,49	-1,51	0,01	-0,01	-1,48	1,51
Investment	AvgMonthlyManagerPersonalInvestmentTransactions	2,11	-1,29	-0,42	0,46	-1,29	0,43
Investment	AvgMonthlyManagerBusinessInvestmentTransactions	2,14	-1,30	-0,42	0,40	-1,29	0,47
Size	NumberTellerCapacity	1,02	0,93	-0,58	-0,52	-1,90	1,05
Size	NumberManagerPersonalCapacity	1,00	1,00	-0,50	-0,50	-2,00	1,00
Size	NumberManagerBusinessCapacity	1,00	1,00	-0,50	-0,50	-2,00	1,00
Size	NumberATM	1,02	1,00	-0,68	-0,31	-2,01	0,98
Size	NumberPersonalClientsTierA	2,28	0,72	-0,12	-0,24	-2,46	-0,17
Size	NumberPersonalClientsTierB	1,42	1,41	-0,43	-0,42	-2,15	0,17

Tabela 9. Valores normalizados pela transformação de z-score para cada uma das variáveis de clusterização, para cada um dos clusters. (continuação)

Feature Grouping	Features	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
Size	NumberPersonalClientsTierC	0,98	1,05	-1,00	-0,31	-1,70	0,98

Size	NumberPersonalClientsTierD	1,00	0,99	-1,02	-0,31	-1,67	1,01
Size	NumberINSSClients	1,04	0,97	-1,01	-0,33	-1,66	0,99
Size	NumberSalaryAccounts	1,03	0,98	-0,54	-0,52	-1,94	0,99
Size	NBusinessClientsTierA	3,00	-0,04	-0,06	-0,07	-2,81	-0,02
Size	NBusinessClientsTierB	1,00	1,00	-0,80	-0,80	-1,41	1,00
Size	NBusinessClientsTierC	0,99	1,00	-0,99	-0,34	-1,67	1,01
Size	NBusinessClientsTierD	0,99	1,04	-1,01	-0,32	-1,67	0,98
Size	BranchSizeSquareMeters	1,52	-0,13	-0,93	-0,14	-1,79	1,48
TransactionValue	AvgMonthlySavingsAccountDeposit	1,48	1,52	-0,96	-0,11	-1,81	-0,11
TransactionValue	AvgMonthlyPersonalLoanAmount	1,43	1,57	-0,94	-0,11	-1,77	-0,18
TransactionValue	AvgMonthlyBusinessLoanAmount	1,48	1,52	-0,98	-0,09	-1,79	-0,13
TransactionValue	AvgMonthlyPersonalInvestmentsAmount	2,35	-0,96	-0,17	0,65	-0,96	-0,91
TransactionValue	AvgMonthlyBusinessInvestmentsAmount	2,39	-1,19	-0,28	0,61	-1,23	-0,30
TransactionVolume	AvgMonthlyATMPaymentsTransactions	1,36	1,39	-0,30	0,24	-0,85	-1,85
TransactionVolume	AvgMonthlyATMWithdrawTransactions	1,38	1,36	-0,31	0,26	-0,86	-1,83
TransactionVolume	AvgMonthlyATMTransferTransactions	1,37	1,36	-0,30	0,27	-0,86	-1,84
TransactionVolume	AvgMonthlyATMDepositTransactions	1,36	1,40	-0,31	0,24	-0,88	-1,81
TransactionVolume	AvgMonthlyATMTransactions	1,45	1,44	-0,55	0,11	-1,25	-1,20
TransactionVolume	AvgMonthlyTellerPaymentsTransactions	1,44	1,43	-0,54	0,13	-1,25	-1,21
TransactionVolume	AvgMonthlyTellerWithdrawTransactions	1,44	1,46	-0,55	0,11	-1,23	-1,22
TransactionVolume	AvgMonthlyTellerTransferTransactions	1,47	1,40	-0,56	0,13	-1,24	-1,20
TransactionVolume	AvgMonthlyTellerDepositTransactions	1,44	1,42	-0,56	0,13	-1,22	-1,22
TransactionVolume	AvgMonthlyTellerTransactions	1,49	1,38	-0,57	0,12	-1,21	-1,22
TransactionVolume	AvgMonthlyManagerTransactions	1,54	1,46	-0,94	-0,14	-1,81	-0,11

Fonte: Dados originais da pesquisa.

As comparações nesta tabela devem ser feitas linha a linha, ou seja, cluster a cluster. O cluster 1 apresentou os maiores valores em média para cada uma das variáveis de clusterização, enquanto o cluster 5 apresentou os valores mais baixos para as variáveis de clusterização. O cluster 4 apresentou em média o maior valor de dispersão, e o cluster 1 foi o que apresentou o menor valor de dispersão média. O cluster 2 apresentou valores médios normalizados muito próximos do cluster 1, sendo, portanto, o segundo cluster com melhores indicadores de performance relativa.

Os clusters que apresentaram os piores indicadores de performance foram os cluster 5, e cluster 6, enquanto o cluster 4 ficou em terceiro lugar, e o cluster 3 obteve valores que o colocaram na mediana dentre os clusters.

A partir dos valores mostrados na tabela 9 podemos concluir que as agências do cluster C1 podem ser considerados benchmark, pois são agências de tamanho físico muito acima da média, com uma quantidade maior de clientes Pessoa Física, e Pessoa Jurídica, além de apresentar um maior potencial de realizar negócios, pois seus clientes possuem perfis de investidor, e tomadores de crédito, além destas agências apresentarem altos níveis de

fluxo de transações, em caixa humano, caixa eletrônico, e na atuação de seus gerentes de conta pessoa física, e que atendem empresas.

As agências do cluster C2 apresentaram perfis muito similares às agências do cluster C1, as agências com um perfil muito abaixo do esperado se encontram no cluster C5, as agências deste cluster apresentam um perfil muito inferior às agências do cluster C1, sendo pequenas em tamanho físico, com pouca quantidade de clientes pessoa física, e menor quantidade de clientes pessoa jurídica, apresentando também um baixo fluxo de realização de negócios, apresentando também resultado financeiro muito abaixo da média, tendo portanto estabilidade financeira baixa, e um baixo potencial para expansão de negócios.

O cluster 1 apresentou os maiores valores em média para cada uma das variáveis de clusterização, enquanto o cluster 5 apresentou os valores mais baixos para as variáveis de clusterização. O cluster 4 apresentou em média o maior valor de dispersão, e o cluster 1 foi o que apresentou o menor valor de dispersão média. O cluster 2 apresentou valores médios normalizados muito próximos do cluster 1, sendo, portanto, o segundo cluster com melhores indicadores de performance relativa. Os clusters que apresentaram os piores indicadores de performance foram os cluster 5, e cluster 6, enquanto o cluster 4 ficou em terceiro lugar, e o cluster 3 obteve valores que o colocaram na mediana dentre os clusters. Os clusters apresentaram os seguintes valores médios, desvio padrão, e de dispersão média calculados para cada uma das variáveis de clusterização

- Cluster1: média = 1,45 ; desvio padrão = 0,25 ; dispersão = 0,17
- Cluster2: média = 0,99 ; desvio padrão = 0,54 ; dispersão = 0,54
- Cluster3: média = -0,64 ; desvio padrão = 0,27 ; dispersão = -0,42
- Cluster4: média = -0,03 ; desvio padrão = 0,23 ; dispersão = - 7,17
- Cluster5: média = -1,49 ; desvio padrão = 0,35 ; dispersão = - 0,23
- Cluster6: média = -0,27 ; desvio padrão = 0,90 ; dispersão = - 3,35

Considerando todas as análises apresentadas, e utilizando os valores médios normalizados (z-score) para cada uma das variáveis de clusterização, e dos agrupamentos de variáveis é possível caracterizar os clusters de forma relativa considerando os agrupamentos como indicadores de performance.

Cluster 1 são agências (400 no total) benchmark, grandes em tamanho físico, e em quantidade de clientes pessoa física e pessoa jurídica (empresas), possuem alto potencial de negócios, alto apetite por crédito, boa saúde financeira, e possuem altos níveis transacionais tanto em volume de transações quanto em valores médios transacionados, e os clientes destas agências possuem também um perfil investidor.

Cluster 2 tem total de 300 agências, possuem perfis similares às agências do Cluster 1, com alto potencial de negócios, boa saúde financeira, alto apetite por crédito, são agências

grandes, e possuem altos níveis transacionais em volume e valores, porém os clientes das agências deste cluster não possuem perfil de investidores.

Cluster 3 possui 2500 agências, possuem um perfil de agências com baixa saúde financeira, pouco potencial de negócios, são agências de médio porte, possuem pouco apetite por crédito, os níveis transacionais são baixos tanto em volume quanto por valores, e os clientes destas agências não possuem perfil de investidores.

Cluster 4 com total de 700 agências, e possuem um perfil mediano quando comparado relativamente com os outros clusters, todas as suas características são medianas, ou seja são agências de médio porte, com níveis transacionais médios tanto em volume quanto em valores, possuem potencial de negócios, e saúde financeira medianas, os clientes destas agências possuem um perfil conservador em relação à investimentos.

Cluster 5 tem 500 agências, e possuem as agências com os piores perfis possíveis, tendo os piores indicadores, com baixa saúde financeira, baixo potencial de negócios, são agências muito pequenas, com pouco apetite por crédito, baixos valores e volumes transacionais, e seus clientes não possuem perfil de investidores.

Cluster 6 tem 800 agências de grande porte em tamanho físico, e em quantidade de clientes, tem saúde financeira relativamente boa em relação aos piores clusters neste sentido, mas possuem baixo potencial de negócios, pouco apetite por crédito, seus clientes não tem perfis de investidores, possuem níveis medianos de transações em volumes, porém possuem baixos valores transacionados em média.

A Tabela 10 mostra a densidade de agências por cluster, e por Estado (UF) no Brasil. Cada célula da tabela abaixo representa o percentual do total de agências do cluster (coluna) em questão em relação aos Estados brasileiros (linhas) Os cinco Estados com maior representatividade dentro dos clusters são SP, MG, RJ, BA, e PR, notando-se que este fato pode ser devido à estratégias de abertura de agências do banco de varejo em questão, não tendo nada relacionado com o resultado da clusterização em si. Os três Estados com menos agências são RR, AC, e AP.

Tabela 10. Densidade de agências por cluster, e por Estado (UF) no Brasil.

Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6

SP	22%	25%	22%	22%	24%	21%
MG	13%	10%	10%	9%	11%	10%
RJ	8%	10%	8%	9%	7%	7%
BA	8%	7%	7%	9%	9%	7%
PR	6%	4%	6%	7%	5%	6%
RS	5%	4%	6%	4%	5%	6%
PE	6%	6%	5%	4%	4%	5%
PA	4%	4%	5%	5%	5%	4%
CE	4%	4%	4%	4%	4%	3%
SC	3%	4%	4%	4%	4%	4%
GO	4%	3%	3%	4%	3%	4%
MA	4%	2%	3%	4%	3%	4%
PB	2%	1%	2%	2%	2%	2%
AM	1%	1%	2%	2%	1%	2%
MT	2%	1%	2%	2%	2%	2%
ES	1%	1%	2%	1%	2%	2%
PI	1%	2%	1%	1%	1%	3%
DF	2%	1%	2%	2%	2%	1%
AL	2%	2%	2%	1%	1%	2%
RN	1%	2%	1%	2%	2%	0%
MS	1%	1%	1%	1%	1%	2%
SE	2%	2%	1%	1%	1%	1%
TO	1%	1%	1%	1%	1%	1%
RO	1%	1%	1%	0%	0%	1%
AP	0%	0%	0%	1%	0%	1%
AC	0%	0%	0%	0%	0%	0%
RR	0%	0%	0%	0%	0%	0%

Fonte: Dados originais da pesquisa.

Relembrando que os seis clusters encontrados possuem as seguintes características de performance avaliando-se as métricas de clusterização.

Cluster 1: Agências benchmark de grande porte, com alta saúde financeira, apetite por crédito e volumes transacionais elevados. Clientes têm perfil investidor.

Cluster 2: Agências similares ao Cluster 1, porém seus clientes não têm perfil investidor.

Cluster 3: Agências de baixa saúde financeira, pouco potencial de negócios e transações baixas. Clientes não investidores.

Cluster 4: Agências medianas em porte e desempenho, com clientes conservadores em investimentos.

Cluster 5: Agências de pior perfil, baixa saúde financeira, pequenas e sem potencial de negócios. Clientes não investidores.

Cluster 6: Grandes agências com saúde financeira relativamente melhor, porém com pouco potencial de negócios e clientes não investidores.

A Figura 1 abaixo traz um resumo do resultado da clusterização de agências de acordo com os valores calculados da métrica de dispersão para os agrupamentos de variáveis. A tabela traz os nomes dos agrupamentos na primeira coluna, a descrição de negócios para cada um dos agrupamentos na segunda coluna, e nas colunas seguintes indicadores visuais sobre os valores encontrados para a dispersão de cada agrupamento para cada um dos cluster. O valor de referência (benchmark) utilizado para efeito de comparação foi a média geral por variável macro para todos os clusters. Os círculos vermelhos (resultados negativos) representam valores médios acima do benchmark, os círculos amarelos (resultados medianos) apresentam valores próximos da média geral, e círculos verdes (resultados positivos) representam valores abaixo do benchmark. Ou seja, a partir da Figura 1 é possível concluir que as agências com melhores performances são as agências do Cluster 1 e do Cluster 2, as agências medianas são aquelas que pertencem ao Cluster 4, e as agências dos clusters 3, 5 e 6 são agências com performances pobres em relação às métricas de resultado avaliadas. Os dois extremos dentre os seis clusters são os Cluster 1 (melhor) que apresenta valores acima da média de performance para todos os âmbitos avaliados, enquanto o Cluster 5 (pior) apresenta valores médios de performance muito abaixo dos valores medianos.

Figura 1. Descrição dos seis clusters encontrados como resultado do algoritmo K-Means.

Agrupamentos de Variáveis	Descrição do agrupamento	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
Business Potential	Capacidade da agência em gerar negócios através de produtos bancários comuns, linhas de crédito, e investimentos.	🟢	🟢	🔴	🟡	🔴	🔴
Credit	Apetite por crédito via perfil dos clientes pessoa física, e pessoa jurídica da agência.	🟢	🟢	🔴	🟡	🔴	🔴
Financial Health	Saúde financeira da agência, baseado na análise contábil (média mensal) entre custo, receita, e perdas.	🟢	🟢	🔴	🟡	🔴	🟢
Investment	Apetite por investimento dos clientes pessoa física e pessoa jurídica da agência.	🟢	🔴	🔴	🟡	🔴	🟡
Size	Tamanho físico da agência (metragem), quantidade de clientes PF e PJ, níveis de transações, e quantidade de colaboradores.	🟢	🟢	🟡	🟡	🔴	🟢
Transaction Value	Valores médios por transação realizados na agência. É a média dos valores realizados por clientes PF e por PJ.	🟢	🟢	🔴	🟡	🔴	🟡
Transaction Volume	Quantidade média das transações realizadas na agência por seus clientes PF e PJ.	🟢	🟢	🔴	🟡	🔴	🔴

Considerações Finais

Conclui-se que com o resultado da aplicação do algoritmo de clusterização K-Means nas cinco mil agências bancárias do banco de varejo em questão foram encontrados seis grupos homogêneos comparáveis, reduzindo a necessidade de utilizar o total de agências para a atribuição de metas, reduzindo assim tempo, e custo de operação nas análises de performance das agências, utilizando apenas seis atribuições mensais de metas para

acompanhamento de resultados. A metodologia de análise proposta mostrou-se eficaz para identificar os perfis de agências, a métrica dispersão foi utilizada qualitativamente para identificar se as variáveis de clusterização foram eficazes em sua aplicação no algoritmo. Como resultado original foi atribuído o valor de benchmark de 17% no valor de dispersão das variáveis como identificação de valor limite (*benchmark treshold*), pois não há ainda na literatura investigada nenhuma menção à sistematização de análise de resultados de clusterização, ou valores de corte para considerar que os algoritmos foram eficazes ou não.

Para comparação das agências foi proposta agregação das variáveis de clusterização em grupos de variáveis de performance: Business Potential, Credit, Finantial Health, Investiment, Size, Transaction Value, Transaction Volume. Foram propostas também metodologias de análise de comparação entre os clusters utilizando análise bivariada entre os clusters encontrados (colunas) e variáveis (linhas) tanto de clusterização, quanto o grupo de variáveis de performance. Os clusters Cluster1, e Cluster2 apresentaram características semelhantes, o Cluster1 apresentou os melhores indicadores de resultados, enquanto que o Cluster2 ficou com o segundo lugar em termos de performance dos indicadores, o pior cluster foi o Cluster5, e o segundo pior foi o Cluster3, o terceiro pior cluster em termos de resultados foi o Cluster6, enquanto que o Cluster4 ficou na mediana de resultados comparativos.

Este trabalho teve como objetivo concretizado dar início ao que pode ser utilizado na definição de novos benchmarks de dispersão dos clusters, melhorando os valores limites considerando novos grupos de variáveis agregadas, realizando testes de melhores grupos de variáveis de clusterização, ou utilizando novas estratégias de clusterização.

Agradecimento

Agradeço minha esposa por seu eterno suporte e carinho por mim, e meu orientador pelas imensas e significantes sugestões e discussões.

Referências

MacQueen, J.B. (1967). Some Methods for classification and Analysis of Multivariate Observations. Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability. Vol. 1. University of California Press. pp. 281–297.

Tang, G.; Tian, R.; Wu, B. (2022). An Overview of Clustering Methods in The Financial World. Proceedings of the 2022 7th International Conference on Financial Innovation and Economic Development Atlantis Press.

Herrera-Restrepo, O.; Triantis, K.; Seaver, W.L.; Paradi, J.C.; Zhu, H.; Bank branch operational performance: A robust multivariate and clustering approach, Expert Systems with Applications, Volume 50, 2016, Pages 107-119, ISSN 0957-4174.

Marques, B.P.; Alves,C.F. (2020) Using clustering ensemble to identify banking business models. Intell Sys Acc Fin Mgmt. 27: 66– 94.

Domeniconi, C.; Gunopulos, D.; Ma, S.; Papadopoulos, D.; Yan, B. (2007) Locally adaptive metrics for clustering high dimensional data. Data Mining and Knowledge Discovery, Volume 14, Issue 1, pp 63–97.

Sharahi, M.; Aligholi, M. (2015) Classify the Data of Bank Customers Using Data Mining and Clustering Techniques (Case Study: Sepah Bank Branches Tehran), J. Appl. Environ. Biol. Sci., 5(5) 458-464.

T. E. Raghunathan, Synthetic data, Annual Review of Statistics and Its Application, 8, 129-140, 2021.

Dankar, K.; Mahmoud, I. (2021) Fake it till you make it: guidelines for effective synthetic data generation, Applied Sciences 11.5: 2158.

Hradec, J.; Craglia, M.; Di Leo, M.; De Nigris, S.; Ostlaender, N.; Nicholson, N. (2022) Multipurpose synthetic population for policy applications, EUR 31116 EN, Publications Office of the European Union, Luxembourg, ISBN 978-92-76-53478-5.