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

Osvaldo Luiz dos Santos Pereira¹\*;Thiago Gentil Ramires2

1 Nome da Empresa ou Instituição (opcional). Titulação ou função ou departamento. Endereço completo (pessoal ou profissional) – Bairro; 00000-000 Cidade, Estado, País

2 Universidade Tecnológica Federal do Paraná. Doutor de ciência. Rua Marcílio Dias, Jardim Paraíso; 86812-460. Apucarana, Paraná, Brasil.

\*autor correspondente: nome@email.com

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

**Resumo**

Neste trabalho foi 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 entre os clusters. Os dados utilizados foram gerados sinteticamente em cumprimento da Lei Geral de Proteção aos Dados (LGPD) à 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**

Deixar para o final

**Keywords:** K-means; Performance; Strategy.

**Introdução**

Algoritmos de *clusterização* como o *K-Means* (MacQuenn, 1967) foram desenvolvidos desde a década de 1960 para serem utilizados em diversos tipos de aplicações, como por exemplo processamento de sinais, através de métodos de quantização vetorial. Desde então com o crescente desenvolvimento de tecnologias em termos de processamento computacional as aplicações destes tipos de algoritmos se tornaram cada vez mais pragmáticas, e mais amplas, tendo inclusive aplicações em diversos nichos da indústria e do ramo empresarial como mostrado no artigo de (Tang et al., 2022) no qual utiliza diversas técnicas avançadas de Analytics 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. Além disto foi apresentada uma metodologia de avaliação da qualidade do resultado final da clusterização ao utilizar análise bivariada considerando o resultado dos clusters, e 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 de dados para tomada de decisão em bancos. O case utilizado foi a segmentação de agências bancárias em grupos homogêneos e com variáveis comparáveis dentro destes grupos com o objetivo de serem atribuídas metas de performance para cada um dos clusters encontrados. Ao lidar com um número muito grande de agências, e.g. cinco mil ou mais agências espalhadas em diversas regiões do Brasil, é impraticável atribuir cinco mil metas customizadas de performance, e apurar estas metas mensalmente dentre todas as agências, com a criação dos clusters, estes grupos de agências apresentarão comportamentos e características de negócio muito semelhantes, então ao invés de atribuir cinco mil metas, este grupo de agências serão reduzidos a apenas alguns clusters. 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**

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 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. Cada uma das observações do *dataset* 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.

O *dataset* de modelagem contém 5200 observações 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 do qual cada agência pertence. Abaixo são apresentadas os nomes e descrições de cada uma das variáveis da tabela.

* 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 (estimada ou delcarada) de clientes pessoa física do tier A
* AvgMonthlyIncomePersonalClientTierB: valor médio de renda mensal (estimada ou delcarada) de clientes pessoa física do tier B
* AvgMonthlyIncomePersonalClientTierC: valor médio de renda mensal (estimada ou delcarada) de clientes pessoa física do tier C
* AvgMonthlyIncomePersonalClientTierD: valor médio de renda mensal (estimada ou delcarada) de clientes pessoa física do tier D
* AvgMonthlyINSSBenefitsAmount: valor médio do benefício recebido por clientes INSS
* AvgMonthlySalaryValue: valor médio da renda mensal de clientes pessoa física
* AvgMonthlyEBITDABusinessClientTierA: valor médio do faturamento mensal de clientes pessoa jurídica do tier A
* AvgMonthlyEBITDABusinessClientTierB: valor médio do faturamento mensal de clientes pessoa jurídica do tier B
* AvgMonthlyEBITDABusinessClientTierC: valor médio do faturamento mensal de clientes pessoa jurídica do tier C
* AvgMonthlyEBITDABusinessClientTierD: valor médio do faturamento mensal de clientes 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)

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.
* Finantial 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.

**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 desta tabela 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. Os resultados desta análise se encontram na Tabela 1. A Tabela 1 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. 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.

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%),
* 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. Na tabela 3 são apresentadas as variáveis de clusterização com valores médios de dispersão calculados ponderação pela quantidade de agências por clusters, e também pela média simples. *As cinco variáveis que apresentaram os maiores valores de dispersão média*

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

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

* AvgMonthlyTellerTransferTransactions,
* AvgMonthlyBusinessInvestmentsAmount,
* AvgMonthlyIncomePersonalClientTierB,
* AvgMonthlyTellerWithdrawTransactions,
* AvgMonthlyManagerPersonalInvestmentTransactions

A tabela 4 apresenta os valores calculados de dispersão médios calculados por cluster e por agrupamento de variáveis de clusterização. Os clusters apresentaram as seguintes médias de dispersão C1 (11%), C2 (13%), C3 (19%), C4 (15%), C5 (23%), C6 (17%). Como valores muito acima de referência tem-se a variável BusinessPotential, FinancialHealth, e a variável Size. Para valores muito abaixo da referência (benchmark) tem-se as variáveis Credit, FinancialHealth, Investiment, TransactionValue, e a variável TransactionVolume.

*A tabela 5 apresenta apresenta os valores médios de dispersão calculados por agrupamento de variáveis de forma geral sem conserar os valores por cluster. Os grupos de variáveis que apresentaram valores de dispersão menores que os valores de referência (benchmark) são Credit, Size, TransactionValue, TransactionVolume, enquanto que as variáveis que tiveram valores acima da referência (benchmark) são BusinessPotential, e FinancialHealth.*

A tabela 6 apresenta os valores médios calculados por cluster, e por variáveis de clusterização. Alguns padrões observados são de que o cluster C1 apresenta em média quatidades maiores de clientes do tipo Pessoa Física. Os clusters C1, e C6 apresentam valores de metragem quadrada acima da mediana (agências grandes), enquanto que 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 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 que 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).

A tabela 7 ...

A tabela 8 ...

Os clusters encontrados podem ser resumidos nas seguintes características

* Cluster 1:
* Cluster 2:
* Cluster 3:
* Cluster 4:
* Cluster 5:
* Cluster 6:

Tabela 1 . A tabela mostra os valores calculados de dispersão média para cada variável de clusterização calculada por cluster resultante. Os clusters obtiveram os seguintes valores médios de dispersão C1 (11%), C2 (12%), C3 (19%), C4 (15%), C5 (25%), C6 (19%). A variável NumberTellerCapacity apresentou valores consideravelmente maiores que a mediana, principalmente para os clisters C3, C4, e C5. Outras variáveis que apresentaram valores altos de dispersão foram NumberPersonalClientsTierA (C3: 32%, C4: 33%, C5: 71%, C6: 32%), NumberSalaryAccounts (C5: 41%), AvgMonthlyIncomePersonalClientTierD (C3: 33%, C5: 40%, C6: 39%), AvgMonthlyEBITDABusinessClientTierC (C1: 31%, C3: 47%, C4: 41%, C5: 62%, C6, 61%). Como exemplo de baixíssima dispersão tem-se as variáveis NumberManagerPersonalCapacity, e NumberManagerBusinessCapacity que apresentaram 0% de dispersão dentre os clusters criados.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **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% |
| 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% |

Tabela 1. A tabela mostra os valores calculados de dispersão média para cada variável de clusterização calculada por cluster resultante. Os clusters obtiveram os seguintes valores médios de dispersão C1 (11%), C2 (12%), C3 (19%), C4 (15%), C5 (25%), C6 (19%). A variável NumberTellerCapacity apresentou valores consideravelmente maiores que a mediana, principalmente para os clisters C3, C4, e C5. Outras variáveis que apresentaram valores altos de dispersão foram NumberPersonalClientsTierA (C3: 32%, C4: 33%, C5: 71%, C6: 32%), NumberSalaryAccounts (C5: 41%), AvgMonthlyIncomePersonalClientTierD (C3: 33%, C5: 40%, C6: 39%), AvgMonthlyEBITDABusinessClientTierC (C1: 31%, C3: 47%, C4: 41%, C5: 62%, C6, 61%). Como exemplo de baixíssima dispersão tem-se as variáveis NumberManagerPersonalCapacity, e NumberManagerBusinessCapacity que apresentaram 0% de dispersão dentre os clusters criados. (continuação)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Features** | **C1** | **C2** | **C3** | **C4** | **C5** | **C6** |
| 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% |

*Tabela 2 – A tabela exibe os valores médios de dispersão calculados por variável de clusterização, sem considerar a segmentação por clusters. A primeira coluna Avg Weighted Dispersion é calculada considerando a quantidade de agências por clusters. A coluna Avg Dispersion é a média simples dos valores de dispersão para todos os clusters sem nenhuma ponderação. As cinco variáveis que apresentaram os maiores valores de dispersão média*

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

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

* AvgMonthlyTellerTransferTransactions,
* AvgMonthlyBusinessInvestmentsAmount,
* AvgMonthlyIncomePersonalClientTierB,
* AvgMonthlyTellerWithdrawTransactions,
* AvgMonthlyManagerPersonalInvestmentTransactions

|  |  |  |
| --- | --- | --- |
| **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% |

*Tabela 3 - A tabela exibe os valores médios de dispersão calculados por variável de clusterização, sem considerar a segmentação por clusters. A primeira coluna Avg Weighted Dispersion é calculada considerando a quantidade de agências por clusters. A coluna Avg Dispersion é a média simples dos valores de dispersão para todos os clusters sem nenhuma ponderação. As cinco variáveis que apresentaram os maiores valores de dispersão média*

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

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

* AvgMonthlyTellerTransferTransactions,
* AvgMonthlyBusinessInvestmentsAmount,
* AvgMonthlyIncomePersonalClientTierB,
* AvgMonthlyTellerWithdrawTransactions,
* AvgMonthlyManagerPersonalInvestmentTransactions (continuação)

|  |  |  |
| --- | --- | --- |
| **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% |

Tabela 4 – A tabela exibe os valores calculados de dispersão médios calculados por cluster e por agrupamento de variáveis de clusterização. Os grupos de variáveis que apresentaram valores de dispersão menores que os valores de referência (benchmark) são Credit, Size, TransactionValue, TransactionVolume, enquanto que as variáveis que tiveram valores acima da referência (benchmark) são BusinessPotential, e FinancialHealth. Os clusters apresentaram as seguintes médias de dispersão C1 (11%), C2 (13%), C3 (19%), C4 (15%), C5 (23%), C6 (17%). Como valores muito acima de referência tem-se a variável BusinessPotential nos clusters C5, e C6, a variável FinancialHealth no cluster C3, A variável Size no cluster C5. Para valores muito abaixo da referência (benchmark) tem-se a variável Credit nos clusters C1, e C2, a variável FinancialHealth no cluster C1, a variável Investiment no cluster C1, a variável TransactionValue no cluster C1, e a variável TransactionVolume nos clusters C1, e C2.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Feature Grouping** | **Cluster 1 (Dispersão)** | **Cluster 2 (Dispersão)** | **Cluster 3 (Dispersão)** | **Cluster 4 (Dispersão)** | **Cluster 5 (Dispersão)** | **Cluster 6 (Dispersão)** |
| 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% |

Tabela 5 – A tabela apresenta os valores médios de dispersão calculados por agrupamento de variáveis. A primeira coluna Avg Weighted Dispersion calcula a dispersão média considerando a ponderação pela quantidade de agências por cluster, enquanto que 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, TransactionValue, TransactionVolume, enquanto que as variáveis que tiveram valores acima da referência (benchmark) são BusinessPotential, e FinancialHealth.

|  |  |  |
| --- | --- | --- |
| **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% |

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). Alguns padrões observados são de que o cluster C1 apresenta em média quatidades maiores de clientes do tipo Pessoa Física. Os clusters C1, e C6 apresentam valores de metragem quadrada acima da mediana (agências grandes), enquanto que 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 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 que 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).

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **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 |

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). Alguns padrões observados são de que o cluster C1 apresenta em média quatidades maiores de clientes do tipo Pessoa Física. Os clusters C1, e C6 apresentam valores de metragem quadrada acima da mediana (agências grandes), enquanto que 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 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 que 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). (continuação)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Features** | **Cluster 1 (Média)** | **Cluster 2 (Média)** | **Cluster 3 (Média)** | **Cluster 4 (Média)** | **Cluster 5 (Média)** | **Cluster 6 (Média)** |
| 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 |
| 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 |

Tabela 7 – Valores médios

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Feature Grouping** | **Cluster 1 (Média)** | **Cluster 2 (Média)** | **Cluster 3 (Média)** | **Cluster 4 (Média)** | **Cluster 5 (Média)** | **Cluster 6 (Média)** |
| BusinessPotential | 171.127 | 171.891 | 107.420 | 129.043 | 84.539 | 85.929 |
| Credit | 2.518 | 2.517 | 1.586 | 1.894 | 1.263 | 1.612 |
| FinancialHealth | 438.972 | 369.825 | 293.017 | 328.920 | 221.177 | 385.465 |
| Investment | 199 | 100 | 125 | 149 | 100 | 150 |
| Size | 1.423 | 1.381 | 891 | 1.011 | 610 | 1.324 |
| TransactionValue | 12.056 | 10.342 | 7.578 | 9.155 | 6.082 | 8.613 |
| TransactionVolume | 45.938 | 45.683 | 28.623 | 34.416 | 22.846 | 20.520 |

Tabela 8 -

|  |  |  |
| --- | --- | --- |
| **Feature Grouping** | **Weighted Average** | **Average** |
| BusinessPotential | 113.444 | 124.991 |
| Credit | 1.726 | 1.898 |
| FinancialHealth | 320.823 | 339.562 |
| Investment | 134 | 137 |
| Size | 1.016 | 1.107 |
| TransactionValue | 8.310 | 8.971 |
| TransactionVolume | 29.917 | 33.004 |

**Considerações Finais**

O título da seção Conclusão(ões) ou Considerações Finais deve ser alinhado à esquerda e grafado em negrito. Fica a critério do aluno e do orientador a escolha de qual termo melhor se adequa ao trabalho. Esta seção deve conter frases curtas, apresentando as conclusões e inferências elaboradas a partir da discussão dos resultados. É importante que estas frases não sejam meras reproduções dos resultados, respondendo aos objetivos propostos no trabalho. Os autores não devem, em hipótese alguma, mencionar, citar ou reproduzir resultados de outros estudos na(s) conclusão(ões) ou considerações finais do TCC. Por fim, salienta-se que essa seção não deve conter tabelas ou figuras, sendo redigida de forma sucinta.

**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](https://publications.jrc.ec.europa.eu/repository/handle/JRC128595), EUR 31116 EN, Publications Office of the European Union, Luxembourg, ISBN 978-92-76-53478-5.

**Apêndice ou Anexo** (opcional)

**deparas.xlsx**: arquivo do tipo Excel contendo os metadados para a criação dos dados sintéticos utilizados no projeto. Este arquivo serve como um tipo de MDM estatístico para simulação dos dados de cada variável de clusterização. O arquivo contém as seguintes abas

* **ProbsFeatures**: contém os valores de média e desvio padrão para cada uma das cinquenta e uma variáveis de clusterização para ocorrência e valores destas variáveis para cada um dos seis clusters criados. É a tabela central utilizada na metodologia para geração dos dados sintéticos e do dataframe de modelagem.
* **Dummy**: contém o metadados das probabilidades de ocorrência das variáveis dummy DummyInsideShopping e DummyPrimeZone.
* **State**: contém o metadados com a distribuição de probabilidade de ocorrência de agências por Estado brasileiro (UF).
* **ds\_cluster**: contém a agregação das variáveis de clusterização com a descrição das características de negócios de cada um dos clusters.
* **Aux**: Contém metadados auxiliares, não é utilizado diretamente na metodologia.
* **Dpintensity**: contém depara de intensidades utilizado para designar os valores para médias e desvios padrões de cada uma das variáveis de caracterização das agências.
* **Dpds**: contém o metadados geral de médias, desvios e intensidades para cada um dos clusters criados. É utilizado auxiliarmente para gerar a tabela ProbFeatures.

**ClusterAnalysis.ipynb**: arquivo notebook jupyter contendo o algoritmo de geração de clusters, e algumas análises de clusterização, investigações estatísticas, e gráficos.