UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL

ESCOLA DE ENGENHARIA

DEPARTAMENTO DE ENGENHARIA QUÍMICA

TRABALHO DE CONCLUSÃO DE CURSO EM ENGENHARIA QUÍMICA

Avaliação de métodos de *clustering* para a segmentação geográfica de clientes

Autor: Guilherme Kauer De Nadal

Orientador: Marcelo Farenzena

Porto Alegre, abril de 2023

Autor: Guilherme Kauer De Nadal

Avaliação de métodos de *clustering* para a segmentação geográfica de clientes

Trabalho de Conclusão de Curso apresentado à COMGRAD/ENQ da Universidade Federal do Rio Grande do Sul como parte dos requisitos para a obtenção do título de Bacharel em Engenharia Química

Orientador: Marcelo Farenzena

Banca Examinadora:

Prof. Dr., Pedro Rafael Bolognese Fernandes, UFRGS

Prof. Dr., Jônathan William Vergani Dambros, UFRGS

Porto Alegre

2023

RESUMO

Em 2014 foram levantados a existência de mais de 1,2 mihões de pontos de vendas de bebidas no Brasil. Entre as empresas mais influentes do ramo, pode ser destacado a Ambev, que detém em seu portfólio marcas de bebidas como Budweiser, Brahma e Stella Artois. Com mais de 1 milhão de clientes em 2022 e 111 centros de distribuições, existe complexo problema para segmentar o atendimento destes pontos de vendas. Para atende-los de forma singular, são designados vendedores a cada ponto de venda, os quais auxiliam nas vendas de produtos como também ajudam preventivamente em possíveis problemas decorrentes da entrega. Levando em conta a quantidade de pontos de vendas que os vendedores da empresa precisam atender diariamente, uma metodologia para o planejamento de visitas se faz necessário. Desta maneira, o presente trabalho buscou propor um método para segmentar geograficamente as regiões de atendimento através de algoritmos de *clustering.* Para isso, foram estudados os fundamentos básicos do método de clusterização bem como testados diferentes métodos: *K-means, K-means ++ e Suppor Vector Machine*. Ao analisar os quatro métodos, foram avalidados parâmetros de velocidade de processamento do algoritmo, de distância média entre os pontos de venda de cada cluster gerado e da distribuição da quantidade de pontos de venda por região. O método que melhor obteve resultados nestes parâmetros foi o K-means, sendo assim utilizado em nas seguintes cidades: Sapucaia do Sul, Porto Alegre e Gravataí. Em um total de 210 clusters, a aplicação do *K-means* gerou uma redução de 69,1% da distância total de quilômetros dos pontos de venda frente aos pontos médios de seus respectivos clusters. Além disso, foi reduzido 67,7% do desvio padrão fem relação a quantidade de clientes alocados por clusters. Mesmo por vezes alocando clientes muitos próximos geograficamente em clusters diferentes, o *K-means* foi capaz de segmentar regiões e melhorar a disitrbuição de clientes por vendedores. Assim, pode ser concluido que métodos de clustering são eficazes para segmentar dados geográficos.

Palavras-chave: *K-means*, *K-means ++*, *SVM, clusterização*

ABSTRACT

In 2014, the existence of over 1.2 million beverage sales points in Brazil were identified. Among the most influential companies in the sector is Ambev, which holds brands such as Budweiser, Brahma, and Stella Artois in its portfolio. With over 1 million customers in 2022 and 111 distribution centers, it becomes complex to find an effective method to segment the service regions. To serve the customer in a singular way, salespeople are assigned to each point of sale, assisting in the sales of products and ensuring that there are no problems during delivery. Considering the number of sales points that the company's salespeople need to serve daily, a methodology for visit planning is necessary. Thus, this study proposed a method to geographically segment the service regions using clustering algorithms. For this purpose, the basic fundamentals of the clustering method were studied, and different methods were tested: K-means, K-means++ and Support Vector Machine. When analyzing the four methods, algorithm processing speed parameters, average distance between sales points in each generated cluster, and distribution of the number of sales points per region were evaluated. The method that performed best in these parameters was K-means, and it was therefore used on Ambev customer bases in Sapucaia do Sul, Porto Alegre, and Gravataí. In a total of 210 clusters, the application of K-means generated a 69.1% reduction in the total distance in kilometers of sales points compared to the midpoint of their respective clusters. In addition, there was a 67.7% reduction in the standard deviation compared to the number of customers allocated per cluster. Even when allocating customers who are geographically very close in different clusters at times, K-means was able to segment regions and improve the distribution of customers among salespeople. Thus, it can be concluded that clustering methods are effective for segmenting geographical data.

Keywords: *K-means*, *K-means ++*, *SVM, clustering*

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

PDV: Ponto de venda

KM: *K-means*

KMC++: *K-means constrained ++*

SVD*: Singular Value Decomposition*

SVM: Support Vector Machine

SI: Sillhoute Index

CH: Calinski-Harabasz Index

SUMÁRIO

1 Introdução

O avanço tecnológico e computacional proporcionado nos últimos anos foi de extrema importância para resolução de complexos problemas matemáticos. Somente em 2022, estima-se que 345 bilhões de reais tenham sido investidos na área de tecnologia no Brasil ("Investimento em TI chegará a R$345 bilhões até o fim de 2022 – IPNews – Comunicação Interativa", 2022). Dentro dos setores de tecnologia, pode ser destacado os avanços tecnológicos na área logística.

Muitos empresas são conhecidas por seus complexos e engenhosos sistemas logísticos, entre elas a Amazon, o Mercado Livre, a Via Varejo e a Magazine Luiza. Essas empresas estão em destaque pelos seus crescimentos em relação à variedade de produtos oferecidos como também pela capacidade de atender clientes com velocidade. Quanto mais rápido um pedido chega ao cliente, maior será a satisfação dele.

Para conseguir atender à clientela com maior agilidade, sistemas como Waze e Google Maps são utilizados para calcular a distância entre endereço de saída do produto e seu destino final, bem como estimar o tempo total do trajeto. Quando se fala de empresas atendendo milhares de clientes, a complexidade do problema aumenta proporcionalmente.

Sistemas logísticos empregam algoritmos para planejar entregas e visitas aos clientes. Nestes sistemas, são programados os melhores trajetos e a ordenação de clientes mais eficiente em termos financeiros. Estes sistemas exigem que seus métodos tenham alta performance na hora de gerar programações de entrega e atendimento, pois, dependendo da quantidade de clientes, pode se tornar muito complexo o planejamento.

A indústria de bebidas é um exemplo de setor industrial com alta complexidade no planejamento logístico. Em 2021, o faturamento deste setor representou 10,6% do PIB brasiliero (Associação Brasileira da Indústria de Alimentos, 2021). Além disso, em 2014 foi levantada a existência de 1,2 milhões de pontos de vendas de bebida (PDV's) no Brasil (Cervieri Júnior et al., [s.d.]). A grande quantidade de PDV's exige uma programação de visitas de vendas e de entregas de produtos. Logo, torna-se essencial um método para dividir geograficamente os clientes a fim de concentrar melhor as regiões de atendimento.

Neste contexto, o presente trabalho de conclusão visa otimizar a segmentação geográfica de clientes de uma empresa de bebidas através de métodos de *clustering.* Os clientes a serem segmentados pertencem à empresa Ambev, os quais estão situados nas cidades de Porto Alegre, Sapucaia do Sul e Gravataí.

O trabalho é divido como segue:

a) Definição e avaliação dos métodos de *clustering*;

b) Seleção de algoritmos para a segmentação dos clientes;

c) Realização de uma triagem dos métodos selecionados;

d) Aplicação do algoritmo escolhido para segmentação de clientes de Porto Alegre, Sapucaia do Sul e Gravataí;

e) Análise dos resultados obtidos;

f) Discussão das análises feitas e dos resultados obtidos.

2 Revisão Bibliográfica

Neste capítulo serão explicados os fundamentos básicos da metodologia de clusterização. Além disso, também serão revisados três algoritmos de *clustering*: *K-means*, *K-means ++* e *Support Vector Machine*.

2.1 Algoritmos de Clusterização

Cluster é um agrupamento de objetos que possuem algum tipo de semelhança entre si. A clusterização, que é a maneira como são criados esses grupos de objetos semelhantes, é muito utilizada na área de estatística exploratória. Entre as áreas de aplicação dessas técnicas, podem ser citados o reconhecimento digital de imagens, a compressão de dados, em *Machine Learning* e a área de bioinformática (Oyewole e Thopil, 2022).

Cada conjunto de dados tem suas particularidades, logo deve-se conhecer os principais tipos de algoritmos de clusterização e avaliar quais são os mais vantajosos para o objetivo definido. Algo comum entre todos métodos é a sua premissa básica: aprendizado não supervisionado, o qual rotula o conjunto de dados sem a necessidade de uma segmentação prévia. Assim, *clustering* é um processo exploratório de segmentação de dados usado e que agrupa os pontos de acordo com a similiaridade entre eles (Surya Priy, 2023).

Existem diversas metodologias para a segmentação de dados disponíveis na literatura, as quais se diferenciam pela forma em que são geradas as regiões segmentadas. Pode-se enfatizar que cada modelo possui vantagens e desvantagens, sendo sugerido o uso deles com de acordo com o que mais atende o objetivo pré-definido. Os modelos mais conhecidos serão abordados nos subtópicos 2.1.1 até 2.1.4.

2.1.1 Clusterizaçao Hierárquica

Baseado no princípio de que cada objeto tem uma conexão com o seu vizinho de acordo com a proximidade entre eles, clusters são representados dentro de um diagrama de árvores, como disposto na Figura 1. Neste diagrama, são expostas as relações entre cada cluster dividido em níveis. Todos os clusters possuem uma ligação entre todos, sendo a proximidade entre eles o principal fator levado em conta. Entre os algoritmos mais conhecidos, podem ser citados *DIANA, AGNES e Birch* (Scikit-learn, 2023).

Vantagens: Fácil implementação, não necessitar ter número de regiões pré-definido e clusters menores são criados em comparação aos outros modelos (Scikit-learn, 2023).

Desvantagens: Dificilmente chega ao resultado ótimo, não é iterativo na alocação de objetos, a aplicação da metodologia inúmeras vezes geram resultados diferentes no mesmo grupo de informações e baixa performance para conjunto de dados extensos (Scikit-learn, 2023).

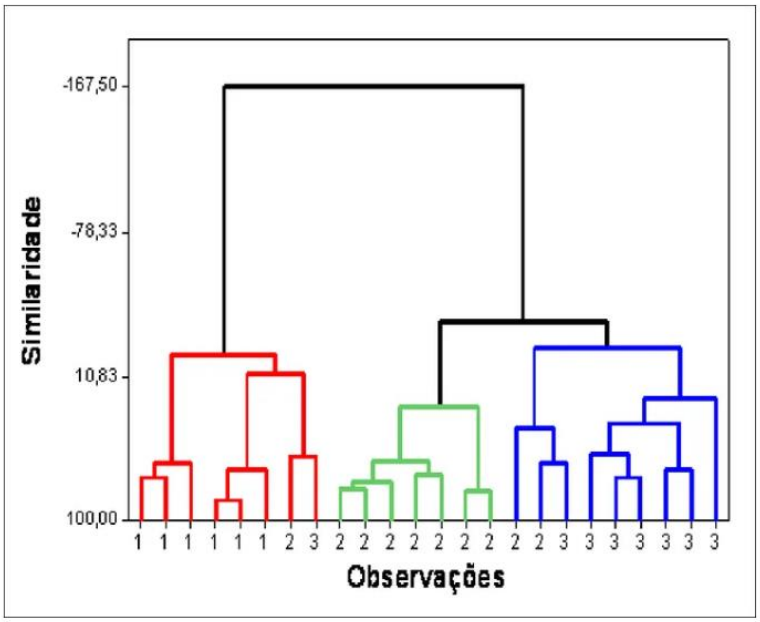


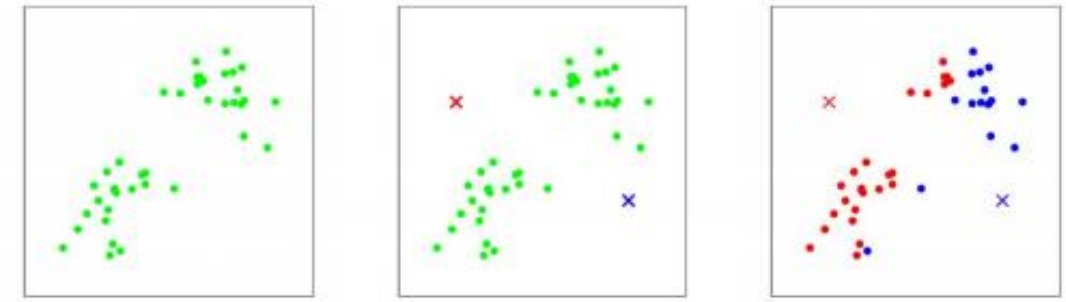
Figura 1: Exemplo do funcionamento de uma árvore hierárquica
Fonte: Lucena, 2019

2.1.2 Clusterização de Centróides

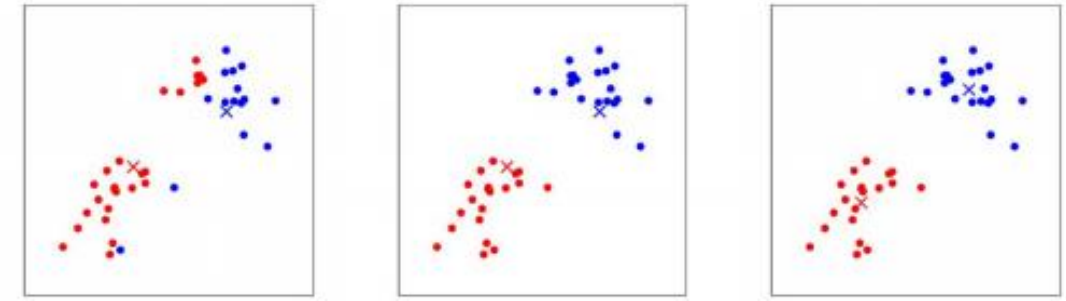
A metodologia de Clusterização de Centróides consiste em segmentar dados dispostos em um plano euclidiano através de círculos em volta dos dados, de maneira cada círculo represente uma classificação. Este método é um método iterativo, o qual busca atingir a reduzir a distância dos centros de cada círculo frente aos pontos dentro dele. Cada ponto tem um vetor com direção e sentido ao centro de cada cluster. O menor valor de vetor calculado corresponderá ao cluster em que deverá ser alocado. Diferentemente da clusterização hierárquica, um número *K* de clusters é definido antes do algoritmo inicializar. Além disso, como mostrado na Figura 2, os centróides dos clusters são escolhidos de maneira randômica inicialmente, o que gera resultados diferentes a cada vez que for aplicado o método. Os algoritmos mais conhecidos são o *K-means* e *MiniBatch K-Means* (Scikit-learn, 2023).

Vantagens: É simples de modelar e programar, ampla aplicabilidade em diversos tipos de dados, existência de iteratividade, a qual corrige alocações errôneas de pontos em clusters, alto desempenho de processamento com grandes quantidades de dados e garantia de convergência da segmentação (Scikit-learn, 2023).

Desvantagens: Necessário conhecer o número de *K* clusters inicialmente e é sensível a valores muito discrepantes (Scikit-learn, 2023).



(a) Conjunto de dados (b) Escolha dos centróides (c) Primeira iteração



(d) Segunda iteração (e) Terceira iteração (f) Última iteração

Figura 2: Etapas do funcionamento do *K-means*
Fonte: Piech, 2013

2.1.3 Clusterização de Densidade

Ao contrário das duas metodologias já mencionadas, este método prioriza regiões de alta densidade em detrimento da distância entre pontos. Os clusters possuem formas e quantidades de pontos bem variados. Devido à forma geométrica adquirida, consegue-se identificar facilmente quais os pontos que são atípicos no conjunto de dados. O algoritmo mais famoso é o *DBScan* (Joshi, 2022).

Como é visto na Figura 3, pontos de alta densidade se concentram bem ao centro dos clusters, enquanto pontos com baixa densidades estão mais afastados dos centros dos clusters. Por último, pontos com coloração preta são considerados *outliers*, sendo assim não classificados dentro de qualquer cluster, permitindo a sua identificação no conjunto de dados.

Vantagens: Não requerem um número definido de *K* clusters, possui bom desempenho em diferentes distribuições geométricas de dados e conseguem detectar pontos discrepantes dos demais com facilidade (Scikit-learn, 2023).

Desvantagens: Necessitam bom conhecimento da distribuição dos dados no conjunto de dados, não generalizam muito bem as regiões criadas com densidades diferentes entre si e exigem recursos computacionais intensos (Scikit-learn, 2023).

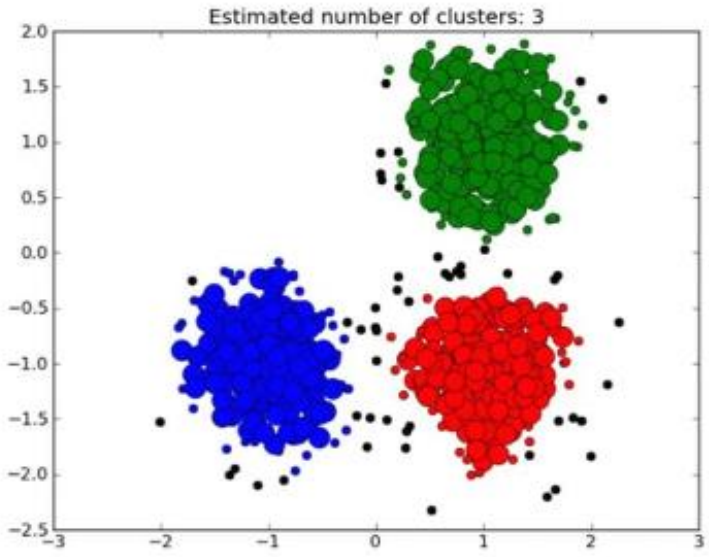


Figura 3: Exemplo de funcionamento do *DBScan*
Fonte: Joshi, 2022

2.1.4 Clusterização de Distribuição

Ao contrário dos outros modelos, essa metodologia busca estimar a probabilidade de certo ponto pertencer a cada cluster existente. Para a definição de clusters, são assumidas distribuições para cada um deles. Essa é uma metodologia muito ineficaz na falta de conhecimento prévio da distribuição de dados no plano. Um dos métodos mais conhecidos é a Mistura Gaussiana, que interpreta os dados dentro de modelos gaussianos (Joshi, 2022).

Na Figura 4 podem ser vistas regiões de alta probabilidade dentro dos próprios clusters que são definidas pelos centros geométricos dos clusters. Conforme há uma redução da distribuição, menores são as quantidades de pontos encontrados dentro das áreas centrais. Nota-se que, mesmo um ponto estando mais próximo do centro de certo cluster, não necessariamente fará parte daquele cluster. Portanto, este método é sensível a *outliers* no conjunto de dados.

Vantagens: Estima a probabilidade dos dados pertencerem a cada cluster, não são assumidos clusters com geometrias esféricas e tem usabilidade em diferentes formatos de dados (Scikit-learn, 2023).

Desvantagens: Difícil categorização de informações, baixo precisão em dados mal distribuídos e assume forma elíptica para os clusters (Scikit-learn, 2023).

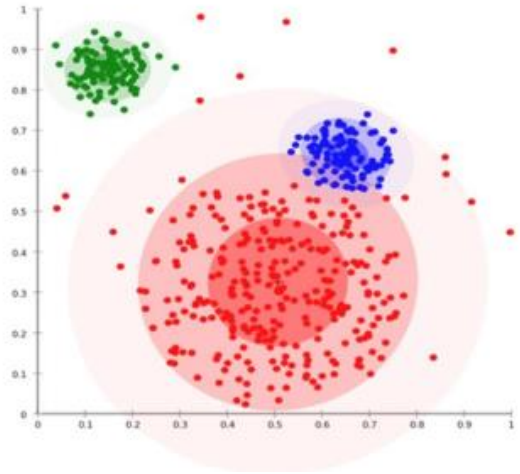


Figura 4: Exemplo de funcionamento da Mistura Gaussiana
Fonte: Joshi, 2022

2.2 K-Means (KMC)

Este algoritmo é um modelo de aproximação de centróides. *K* grupos são criados, os quais escolhem pontos randômicos como pontos centrais de cada cluster. A cada iteração, dados são alocados dentro de clusters próximos. Enquanto os clusters forem diminuindo de área, o algoritmo seguirá ativo. Vale ressaltar que o *K-means* usa a forma geométrica de um círculo para delimitar a sua área, o que pode vir a ser um problema em regiões de dados mal distribuídos (Joshi, 2022).

A distância entre um centróide e um ponto dentro do cluster é calculado através da distância Euclidiana. Desta forma, o algoritmo é finalizado quando *k* clusters têm a menor distância euclidana possível no conjuntos de dados ("Understanding K-Means, K-Means++ and, K-Medoids Clustering Algorithms | by Satyam Kumar | Towards Data Science", [s.d.]). A função matemática para o algoritmo pode ser encontrada na equação 1:

$$J = \sum_{j-1}^{k} \sum_{i-1}^{n} \left\| X_i^j - C_j \right\|^2 \quad (1)$$

Sendo,

k: número de Clusters;

n: número de objetos (neste caso PDVs);

Xi: ponto de informação i (PDV);

C_J: centro do cluster j;

J: diferença total de distância entre todos os pontos e seus respectivos clusters.

Uma das aplicações do *KMC* é a segmentação de clientes em empresas *de E-commerce* de acordo com os seus históricos de compras de produtos, com as suas localizações geográficas e com os seus comportamentos de pesquisa, com o intuito de otimizar os produtos oferecidos pelos vendedores das empresas (Tabianan et al., 2022). Outra

utilização do algoritmo é para a identificação e classificação de padrões de falhas em sistemas industriais, como em processos de manufatura para criação e reparo de materiais metálicos (Gaja e Liou, 2017).

2.3 K-Means Constrained ++ (KMC++)

Este algoritmo é uma variante do *K-Means*, o qual detém o mesmo método de alocar pontos em cada cluster de maneira iterativa utilizando a distância Euclidiana mínima. No entanto, diferentemente do algoritmo original, o KMC++ ajusta os centróides transformando pontos do próprio conjunto de dados como os centros dos clusters, enquanto o KM mantém o centro original definido de forma aleatória. Vale ressaltar que inicialmente ambos os algoritmos escolhem os centróides na primeira iteração de forma aleatória ("Understanding K-Means, K-Means++ and, K-Medoids Clustering Algorithms | by Satyam Kumar | Towards Data Science", [s.d.])

O principal objetivo do KMC++ é diminuir a sensibilidade inicial de alocação dos centróides criados pelo KM, pois as localizações dos centróides são criados de maneira aleatória, o que afeta na forma final dos clusters. Assim, para previnir esta sensibilidade, KMC++ defini pontos alocados dentro dos clusters como os novos centróides a cada iteração ("Understanding K-Means, K-Means++ and, K-Medoids Clustering Algorithms | by Satyam Kumar | Towards Data Science", [s.d.]). Essa definição de um ponto como um novo centróide é baseado na distância dos pontos até o centróide, sendo assim sempre escolhido o ponto mais distante do cluster alocado como o novo centro dele.

2.4 Support Vector Machine (SVM)

O SVM é um método de aprendizado supervisionado o qual prevê a classificação de um conjunto de dados. Ele utiliza dados já categorizados e aprende com eles, de forma que, quando surgem novas informações, é possível prever quais as categorizações a serem utilizadas nestes dados. Entre as principais vantagens do SVM, estão a alta velocidade de processamento e a capacidade de categorizar dados utilizando-se de uma pequena quantidade de informações fornecidas (Monkey Learn, [s.d.]).

A partir de um conjunto de dados já classificado em categorias, um separador é traçado entre os pontos destas categorias. Após traçado o separador, os dados são transformados para um espaço dimensional no qual o separador torna-se um hiperplano, o qual divide as categorizações. As transformações são feitas por funções kernel, as quais definem a maneira em que será criado esse espaço dimensional. As funções kernel podem ser lineares, polinomiais ou de bases radiais (RBF)(IBM, 2021). O processo de transformação e classificação pode ser visto na Figura 5.

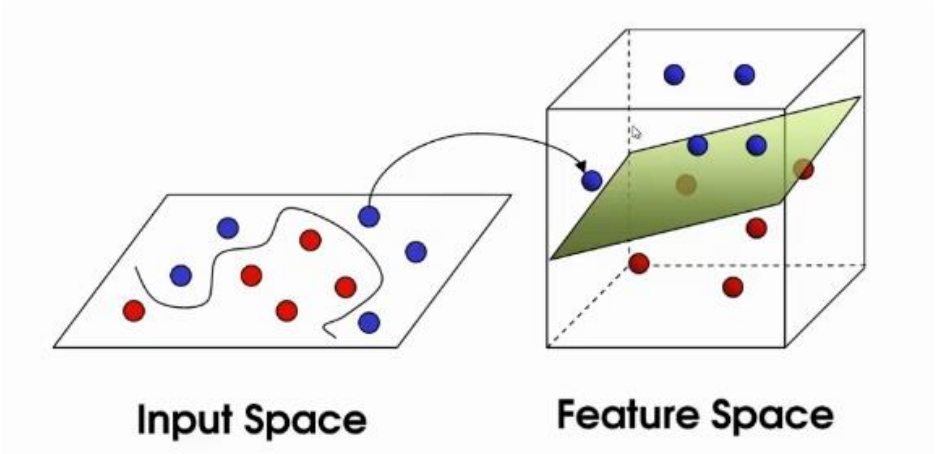
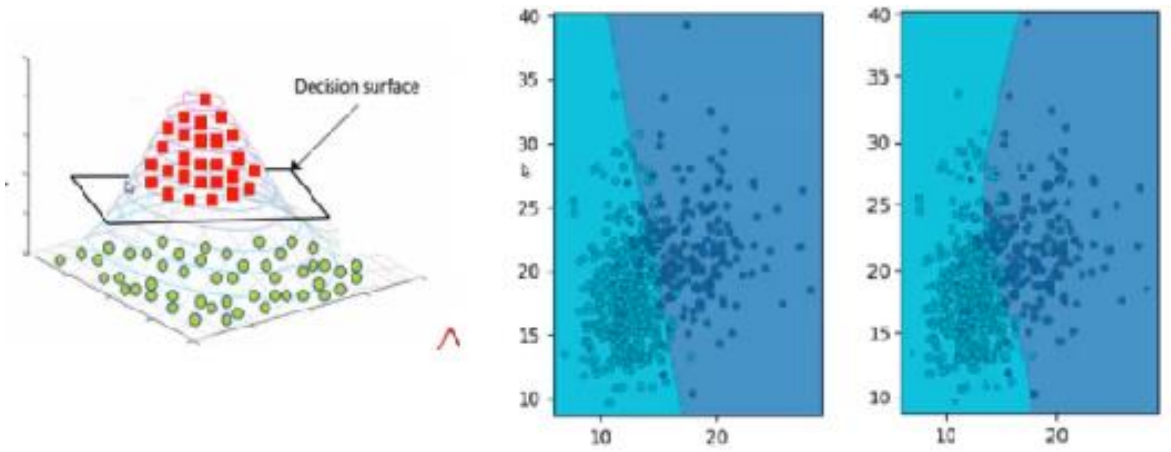


Figura 5: Funcionamento da transformação de dados em outra dimensão e da criação do hiperplano

Fonte:(Shipra Saxena, 2021)

Entre as funções *kernel* usadas para criarem os hiperplanos, a linear é a mais simples, pois traça uma linha reta para categorizar os dados, enquanto a RBF separa os dados de acordo com uma distribuição Gaussiana . Por último, a polinomial, é utilizada em dados não uniformes. Para segmentar dados não lineares, em geral são usadas funções *kernel* polinomiais (Shipra Saxena, 2021). A Figura 6 demonstra como estes hiperplanos separaram os conjuntos de dados.



(a) RBF (b) Linear (c) Polinomial

Figura 6: Categorização de dados a partir de diferentes funções *kernel*

Fonte:(Shipra Saxena, 2021)

Uma das aplicações do SVM é na determinação de sintomas chaves para a identificação de doenças, pois consegue encontrar padrões em conjuntos de dados extensos, como fichas médicas de pacientes (Qin et al., 2023). Outra aplicação é aumentar a precisão de previsões

climáticas, as quais são usadas como parâmetros na estimativa de energia fotovoltaica gerada (Alrashidi e Rahman, 2023). Também pode ser citado a criação de um novo modelo de avaliar e classificar filmes, o qual usa comentários dos filmes feitos por usuários para classifica-los (Jassim et al., 2023).

3 Materiais e Métodos

Neste capítulo, o caso de estudo que contempla a jornada de visita a clientes realizada por vendedores de uma cervejaria nas cidades de Porto Alegre, Sapucaia do Sul e Gravataí será descrito. Para isso, serão utilizados dados de latitudes e longitudes geográficas de clientes fornecidos pela empresa.

3.1 Descrição do caso em estudo

Uma segmentação de clientes será implementada em três cidades, nas quais representam mais de 8.000 pontos de venda (PDV's) e contam com 42 vendedores para o seu atendimento. Esses possuem nichos diferentes de acordo com o tipo de cliente atendido, sendo as segmentações dos clientes por exemplo bares, restaurantes, mercados, mercearias e casas de festas.

Tendo em vista a segmentação do mercado dos clientes, alguns vendedores atendem clientes com segmento de mercado A, enquanto outros atendem clientes do segmento B. Vale ressaltar que não se pode misturar PDV's de vendedores A com PDV's de vendedores B, pois cada categoria de vendedor irá atender uma única categoria de clientes. Desta maneira, vendedores do tipo A atendem clientes do tipo A, enquanto vendedores do tipo B atendem clientes do tipo B e assim por diante. Conforme a Tabela 1, é possível verificar os perfis de vendedores das salas de vendas.

Tabela 1: Quantidade de vendedores por cidade

Gerente Vendas	Cidade Atendida	PDV	Vendedores Total	Vendedores A	Vendedores B	Vendedores C	Vendedores D	Vendedores E
X	Sapucaia	2043	10	8	1	1	0	0
Y	Porto Alegre	3814	22	17	1	1	1	2
Z	Gravataí	2261	10	10	0	0	0	0
Total	-	**8118**	**42**	**35**	**2**	**2**	**1**	**2**

Fonte: Autor, 2023

Outro ponto importante é o funcionamento do agendamento de visitas, as quais ocorrem de segunda-feira a sexta-feira. Além disso, existem três frequências de visitas: semanais, quinzenais e mensais. Como pode ser visto na Tabela 2, a quantidade de visitas semanais é maior do que a de quinzenais e, por sua vez, a quantidade de visitas quinzenais é maior que a quantidades de visitas mensais. É importante pontuar que o ideal é aumentar o total de visitas semanais, pois ao ampliar as frequências de visitas, mais ocasiões de contato com os clientes, resultando assim em maiores probabilidades de acontecerem vendas.

Tabela 2: Frequência de visita aos clientes

Gerente Vendas	Cidade Atendida	PDV	Semanais	Quinzenais	Mensais
X	Sapucaia	2043	901	817	325
Y	Porto Alegre	3814	2088	1602	124
Z	Gravataí	2261	1227	792	242
Total	-	**8118**	**4216**	**3211**	**691**

Fonte: Autor, 2023

3.2 Método utilizado

A fim de formular quais serão os cenários em que serão aplicados os algoritmos, será fundamental fazer uma análise de dados e a escolha da metodologia que melhor atende os requisitos do estudo. Para isso, as cidades foram divididas em nove cenários conforme a Tabela 3. Em cada cenário são definidas as cidades de atendimento bem como os tipos de vendedores que estarão envolvidos nele.

Neste estudo, foi utilizado um computador Dell com o processador Intel(R) Core(TM) i5 CPU M 480 @ 2.67GHz 2.67 GHz, 4,00 GB de memória RAM e um sistema operacional de 64 bits. Já para a execução dos algoritmos, utilizou-se do programa Pycharm versão 2022.2.2, o qual foi usado a linguagem Python para a sua programação.

Após estas definições, serão calculados os totais de clientes e clusters que compõem cada cenário. Vale ressaltar que cada vendedor possui 5 clusters disponíveis, os quais referentes aos dias de atendimento da semana entre segunda feira e sexta feira. Em seguida ,é obtida a média de PDV's que cada cluster possuirá.

Com intuito de gerar regiões bem distribuídas em relação à quantidade de clientes diários e regiões com distâncias mínimas entre clientes, será usada a definição de *K* máximo e *K* mínimo. A primeira corresponde à maior quantidade de PDV's que os clusters poderão ter no cenário, enquanto o segundo equivale à menor quantidade respectivamente. Ambas as métricas foram definidas arbitrariamente pelos gerentes das salas de vendas para cada cenário.

Tabela 3: Cenários de atuação dos algoritmos

Cenário	Cidade Atendida	Tipo de Vendedor	Nº Vendedores	Total de PDV's	PDV's / K	K clusters	K Máximo	K Mínimo
1	Sapucaia	A	8	1780	44,5	40	50	35
2	Sapucaia	B	1	121	24,2	5	28	20
3	Sapucaia	C	1	142	28,4	5	32	23
4	Porto Alegre	A	17	3182	37,4	85	39	31
5	Porto Alegre	B	1	105	21,0	5	22	17
6	Porto Alegre	C	1	182	36,4	5	39	33
7	Porto Alegre	D	1	155	31,0	5	33	29
8	Porto Alegre	E	2	190	19,0	10	22	16
9	Gravataí	A	10	2261	45,2	50	50	40

Fonte: Autor, 2023

A fim de avaliar a segmentação dos algoritmos dos algoritmos K-means, K-means ++ e SVM, uma triagem aplicando-os nos cenários 1,2 e 3 foi definida. A ordem de atividades a ser seguida é a seguinte:

a) Separar a base de clientes com informações de latitudes e longitudes;

b) Definir os cenários com as quantidade clusters, de K máximo e de K mínimo;

c) Aplicar os algoritmos nos cenários 1,2 e 3;

d) Escolher o método que melhor desempenhar e replicar para os 9 cenários;

e) Analisar visualmente e analíticamente seus resultados.

3.3 Análise dos dados

Como levantado anteriormente, existem cinco perfis de vendedores, portanto é relevante analisar a distribuição de PDV's por vendedor e por dia da semana. A Tabela 4 apresenta grande variância na distribuição de clientes nos vendedores da categoria A. Já clientes do tipo B possuem uma menor quantidade de PDV's ao longo da semana e comparativamente quantidades parecidas de PDV's atendidos por dia. Ainda na Tabela 4, pode ser analisada a inexistência de um padrão de distribuição entre dias da semana, seja por sala, seja por vendedor.

Tabela 4: Distribuição de PDV's por vendedor em Sapucaia

Sala Sapucaia			Total PDVs				
Vendedor	Tipo	Total	SEG	TER	QUA	QUI	SEX
1	A	280	49	59	69	57	46
2	A	273	59	50	56	50	58
3	A	254	54	45	56	41	58
4	A	256	48	64	52	56	36
5	A	242	55	53	48	45	41
6	A	107	23	23	16	20	25
7	A	174	33	31	41	33	36
8	A	194	44	35	43	36	36
9	B	121	26	21	26	23	25
10	C	142	29	23	32	29	29
-	-	2043	420	404	439	390	390

Fonte: Autor, 2023

Na Tabela 5, pode ser visto a quantidade de PDV's por dias da semana e por vendedores na sala de Gravataí. Diferentemente da Tabela 4, só existem vendedores do tipo A na sala de Gravataí. Enquanto em Sapucaia a diferença máxima encontrada entre o total de clientes por vendedores do mesmo tipo é de 173 PDVs, a diferença máxima calculada entre vendedores do mesmo nicho em Gravataí é de 75 clientes.

Tabela 5: Distribuição de PDV's por vendedor em Gravataí

Sala Gravataí			Total PDVs				
Vendedor	Tipo	Total	SEG	TER	QUA	QUI	SEX
1	A	251	52	49	39	55	56
2	A	260	63	54	45	40	58
3	A	241	41	43	56	47	54
4	A	198	1	40	49	56	52
5	A	235	58	38	47	46	46
6	A	250	53	52	58	42	45
7	A	226	50	55	38	39	44
8	A	213	48	48	26	52	39
9	A	185	35	48	41	26	35
10	A	202	47	33	42	43	37
-	-	2261	448	460	441	446	466

Fonte: Autor, 2023

Por último, a Tabela 6 traz os PDV's distribuídos por vendedores em Porto Alegre. Enquanto na Tabela 4 e na Tabela 5 o máximo encontrado de PDVs em um dia de visita foi de 69 e 63 clientes respectivamente, em Porto Alegre o máximo encontrado foi de 86 clientes. Alem disso, o máximo de discrepância de PDVs antendidos entre vendedores do mesmo nicho foi maior em comparação às outras salas: 233 clientes.

Tabela 6: Distribuição de PDV's por vendedor em Porto Alegre

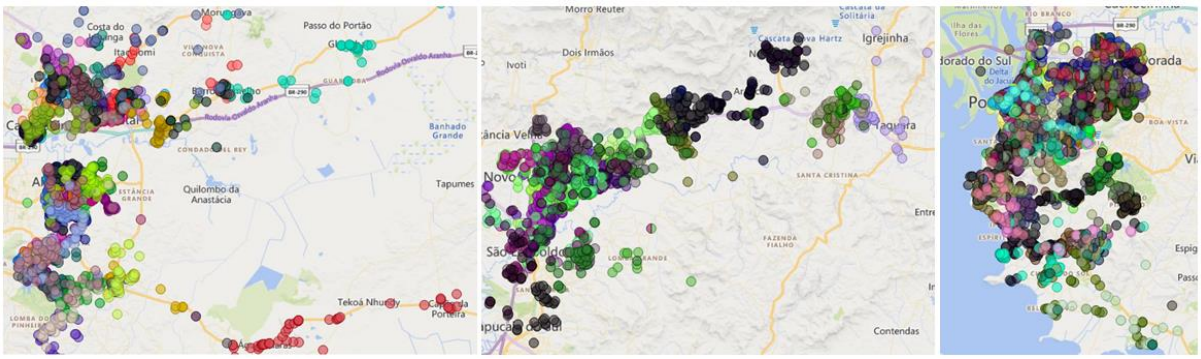
Sala Porto Alegre			Total PDVs				
Vendedor	Tipo	Total	SEG	TER	QUA	QUI	SEX
1	A	275	43	47	34	67	84
2	A	268	76	42	54	53	43
3	A	227	63	35	35	43	51
4	A	231	39	70	30	66	26
5	E	277	56	60	54	62	45
6	A	320	75	56	53	86	50
7	A	172	39	35	34	32	32
8	D	157	30	32	38	28	29
9	C	255	38	61	41	42	73
10	A	139	29	30	25	31	24
11	A	178	34	37	44	38	25
12	A	137	24	27	30	31	25
13	A	114	18	28	24	28	16
14	A	143	29	26	33	21	34
15	A	123	19	24	29	30	21
16	A	118	24	20	25	29	20
17	A	144	32	28	26	28	30
18	A	129	26	24	27	24	28
19	E	111	20	24	19	24	24
20	B	89	20	14	18	16	21
21	A	87	17	16	17	20	17
22	A	120	23	25	21	32	19
-	-	3814	774	761	711	831	737

Fonte: Autor, 2023

A fim de analisar a sobreposição de regiões de atendimento, foram coletadas as latitudes e longitudes de todos os PDV's e, utilizando-se do software Power BI de 2023, foram colocados em mapas a localização do estabelecimento de cada PDV conforme a

Figura 7. Cada círculo colorido representa um cliente único, já cada cor representa um dia de visita de certo vendedor.

Pode ser verificado que em (a) e (b) existem poucas regiões em que não existem sobreposições de vendedores, enquanto em (c) praticamente todas as regiões estão sobrepostas. Portanto, pode ser esperado que, após a segmentação geográfica, os vendedores tenham novas rotas diárias com clientes que antes não eram atendidos por eles.



(a) Mapa de Gravataí (b) Mapa de Sapucaia do Sul (c) Mapa de Porto Alegre

Figura 7: Mapas de Porto Alegre, Sapucaia e Gravataí com os seus respectivos PDV's
Fonte: Autor, 2023

4 Resultados

4.1 Triagem dos algoritmos

Conforme o capítulo anterior, foram escolhidos três algoritmos: *K-means*, *K-means++* e SVM. Será feita uma triagem com estes algoritmos, com o intuito de avaliar o desempenho deles na segmentação de dados geográficos. Os cenários escolhidos foram 1,2 e 3 para a triagem, os quaiscorrespondem à cidade de Sapucaia do sul e a regiões ao seu redor. Equanto o primeiro cenário possui uma alta quantidade de PDV's e de clusters, o segundo e o terceiro cenário possuem quantidades menores em ambos quesitos, sendo assim possível analisar a capacidade dos métodos diante de condições diferentes.

A primeira análise levou em conta o tempo de processamento em cada cenário. Na Tabela 7, encontra-se o tempo total em segundos que cada algoritmo levou para processar em cada cenário. O dado mais divergente aconteceu com *K-means* no cenário um, o qual levou em torno de 3 minutos de processamento. Considerando que o cenário 1 era o mais complexo, 3 minutos é um tempo aceitável de processamento.

Tabela 7: Tempo de processamento dos algoritmos de clusterização

Algoritmo	Cenário	Tempo (s)
KMC	1	186
KMC	2	16
KMC	3	14
KMC++	1	11
KMC++	2	10
KMC++	3	10
SVM	1	22
SVM	2	12
SVM	3	11

Fonte: Autor, 2023

Além do tempo de processamento, a qualidade das regiões geradas podem ser verificadas pela distância total dos PDV's em relação aos centróides de seus respectivos clusters. Também foi medido o desvio padrão da quantidade de clientes entre os clusters gerados, com o intuito de avaliar a dispersão entre eles.

Outra métrica utilizada para avaliar qualidade da classificação gerada foi o *Sillhoute Index* (SI), o qual pode ter um valor de -1 a +1, sendo que -1 indica uma má categorização dos clusters e +1 que os clusters gerados estão bem afastados e distinguíveis entre eles (Ashutosh Bhardwaj, 2020). Conforme a equação (2), é demonstrado como é calculado o SI:

$$SI = \frac{b-a}{\max(b,a)} \quad (2)$$

Sendo que:

SI: Sillhoute Index;

b: Distância média entre todos os centróides;

a: Distância média entre cada ponto de um cluster.

Outra forma de analisar a qualidade dos clusters é avaliar o quão bem separados eles estão entre eles. Desta forma, o *Calinski-Harabasz index* (CH) assume que uma boa solução na segmentação de dados devem ter pontos dentro dos clusters bem próximos como ao mesmo tempo ter clusters bem separados uns dos outros. Quanto maior o resultado do CH, melhor é a separação dos clusters e maior é a densidade deles. O CH pode ser calculado como a razão da soma das distâncias entre clusters pela soma das distâncias dentro dos clusters, e então multiplicado pela razão entre o número de observações e o número de clusters menos um (Haitian Wei, 2020). Na tabela 8, são expostos os dados de *Sillhoute Index, CH*, distância total dentro dos clusters e desvio padrão do tamanho deles.

A função *kernel* a ser utilizada no SVM necessita ser definida. Para estre trabalho, foi escolhido a função a *kernel* polinomial. A fim de gerar os clusters necessários, foram feitos testes em cada cenário de maneira que o grau escolhido gerasse a quantidade mínima estipulada. Enquanto escolhendo a função polinomial de grau três gerou a quantidade de clusters necessárias nos cenários 2 e 3, somente usando função polinomial de grau cinco foi capaz de gerar os quarenta clusters necessários para o cenário 1. Desta forma, nos três cenários foram utilizadas funções *kernel* polinomiais de grau 5.

Tabela 8: Análise da qualidade dos clusters gerados

Cenário	Algoritmo	Distância Total (km	Sillhoute Index	CH	Desvio Padrão (tamanho dos clusters)
1	KM	1473,4	0,357	51700021	6,3
1	KMC++	1361,4	0,448	7831	26,1
1	SVM	1497,0	0,368	4485	6,6
2	KM	152,1	0,308	95291	3,5
2	KMC++	118,7	0,521	281	18,2
2	SVM	152,1	0,323	114	3,5
3	KM	263,8	0,439	179503	3,0
3	KMC++	256,1	0,506	171	9,1
3	SVM	280,9	0,429	122	3,4

Fonte: Autor, 2023

Para KM e KMC++, ambos algoritmos obtiveram resultados na mesma ordem de grandeza na distância dos PDVs em relação ao seus centróid es. No entanto, no cenário 1, o desvio padrão de KMC++ foi muito superior a gerada pelo KM. A variante do *K-means* obteve resultados similares em cenários em que a quantidade de dados e a quantidade de clusters eram pequenos, todavia no cenário com a maior quantidade de dados e clusters a dispersão

foi grande. Em relação ao *Sillhoute Index,* KM teve os piores resultados nos 3 cenários, enquanto KMC++ teve os melhores resultados neste item.

Os resultados obtidos de distância total dentro dos clusters pelo SVM inferiores nos três cenários frente ao KM. Além disso, no primeiro cenário, o SVM atingiu o dobro de desvio padrão em relação ao tamanho do clusters frente ao KM. Nos três cenários o SVM teve resultados superiores ao KM no *Sillhoute Index*.

Por último, vale ressaltar os resultados de CH. Os melhores resultados foram do KM, que condiz com clusters bem densos e bem espaçados entre si. Ao comparar os resultados de CH do KM em relção aos obtidos pelo SVM e KMC++, mostrou-se uma grande superioridade neste quesito.

Entre os três algoritmos, o KM foi o mais consistente em termos desvio padrão e da distância, mesmo tendo tempo de processamento maior no primeiro cenário que todos os demais. O mais próximo em termos de resultados foi o SVM, mesmo assim, alcançou resultados inferiores em quase todos cenários frente ao KM. Mesmo possuindo um *Sillhoute Index* inferior nos três cenários, o resultado do CH calculado foi muito superior aos demais algoritmos. Desta maneira, o *K-means* foi o algoritmo escolhido para ser utilizado nos demais cenários.

4.2 Resultados da distribuição de clientes

Esta seção expõe os resultados em relação à quantidade de clientes por cluster gerado. Nela foram utilizadas métricas estatísticas para avaliar a distribuição de PDVs alocados por dias de visita e por vendedores.

Na tabela 9, são apresentados dados de desvio padrão em relação ao conjunto de dados. Considerando a soma total dos desvios padrões dos 9 cenários, foi observado um grande decréscimo dos mesmos após a aplicação do algoritmo. Examinou-se também que, nos cenários em que haviam maiores quantidades de clusters, foram alcançadas maiores diferenças nas distribuições. Já em cenários com menores quantidades de clusters, foi observado que nos cenários 2 e 3 os desvios padrões aumentaram.

Ao analisar a qualidade dos clusters gerados, o cenário 7 foi o que demonstrou menores resultados de Sillhoute Index (SI) e CH quando comparado aos demais cenários. Enquanto o melhor resultado de SI foi alcançado pelo cenário 5, o melhorde CH foi do cenário 4, que também é o que possui maior quantidade de clusters. O cenário com os melhores resultados combinados foi o 6, na qual conseguiu resultados acima da média no SI e no CH. De forma geral, os números de SI como também os de CH demonstraram que os clusters gerados estão bem segmentados e bem distribuídos.

Tabela 9: Resultados de Sillhoute Index, CH e desvio padrão por clusters gerados

						Pré Algoritmo	Pós Algoritmo
Cenário	Cidade Atendida	Tipo de Vendedor	Clusters	Sillhoute Index	CH	Desvio Padrão	Desvio Padrão
1	Sapucaia	A	40	0,357	51700021	12,6	4,1
2	Sapucaia	B	5	0,308	95291	1,9	3,5
3	Sapucaia	C	5	0,439	179503	2,9	3,4
4	Porto Alegre	A	85	0,331	116444588	15,7	1,9
5	Porto Alegre	B	5	0,576	80685	2,6	2,0
6	Porto Alegre	C	5	0,391	434204	13,7	2,8
7	Porto Alegre	D	5	0,193	245651	3,6	1,8
8	Porto Alegre	E	10	0,362	310004	17,2	2,0
9	Gravataí	A	50	0,342	51920798	10,3	4,5

Fonte: Autor, 2023

Conforme a Tabela 10, destacou-se a distância total encontrada entre *K* máximos e *K* mínimos na situação inicial do estudo. Enquanto no cenário 4 pré algoritmo o *K* mínimo foi de 16 e o *K* máximo foi de 86, após a clusterização foi encontrado um *K* Mínimo de 32 e um *K* Máximo de 39. As condições de contorno para o *K-means* foram escolhidas pelos gerentes de vendas das respectivas salas, portanto há cenários em que a dispersão foi um pouco maior como pode ser visto no cenário 1. Já o cenário 9, que possuía vendedores do mesmo tipo e quantidade de clusters parecidas, atingiu uma dispersão menor entre *K* máximo e *K* mínimo.

Tabela 10: Variância e desvio padrão da quantidade de clientes por cluster

					Pré Algoritmo					Pós Algoritmo				
Sala	Vendedor	Tipo	Cenário	Total	SEG	TER	QUA	QUI	SEX	SEG	TER	QUA	QUI	SEX
Sapucaia	1	A	1	280	49	59	69	57	46	50	50	44	50	44
Sapucaia	2	A	1	273	59	50	56	50	58	43	43	43	43	43
Sapucaia	3	A	1	254	54	45	56	41	58	48	48	48	47	47
Sapucaia	4	A	1	256	48	64	52	56	36	48	48	40	48	41
Sapucaia	5	A	1	242	55	53	48	45	41	48	44	48	48	44
Sapucaia	6	A	1	107	23	23	16	20	25	35	36	35	43	43
Sapucaia	7	A	1	174	33	31	41	33	36	40	48	48	48	40
Sapucaia	8	A	1	194	44	35	43	36	36	40	48	48	40	40
Sapucaia	9	B	2	121	26	21	26	23	25	26	20	28	20	27
Sapucaia	10	C	3	142	29	23	32	29	29	32	23	32	28	27
Porto Alegre	1	A	4	275	43	47	34	67	84	38	37	38	36	37
Porto Alegre	2	A	4	268	76	42	54	53	43	32	39	38	38	39
Porto Alegre	3	A	4	227	63	35	35	43	51	38	38	36	39	39
Porto Alegre	4	A	4	231	39	70	30	66	26	36	38	39	37	39
Porto Alegre	5	E	8	277	56	60	54	62	45	17	22	19	16	18
Porto Alegre	6	A	4	320	75	56	53	86	50	33	38	37	36	39
Porto Alegre	7	A	4	172	39	35	34	32	32	39	39	38	39	39
Porto Alegre	8	D	7	157	30	32	38	28	29	29	33	29	33	31
Porto Alegre	9	C	6	255	38	61	41	42	73	33	33	38	39	39
Porto Alegre	10	A	4	139	29	30	25	31	24	39	36	39	38	37
Porto Alegre	11	A	4	178	34	37	44	38	25	37	38	39	39	38
Porto Alegre	12	A	4	137	24	27	30	31	25	36	37	38	38	39
Porto Alegre	13	A	4	114	18	28	24	28	16	38	35	35	31	35
Porto Alegre	14	A	4	143	29	26	33	21	34	37	39	39	39	39
Porto Alegre	15	A	4	123	19	24	29	30	21	36	37	36	37	38
Porto Alegre	16	A	4	118	24	20	25	29	20	34	35	39	38	39
Porto Alegre	17	A	4	144	32	28	26	28	30	39	39	39	33	39
Porto Alegre	18	A	4	129	26	24	27	24	28	39	39	39	39	39
Porto Alegre	19	E	8	111	20	24	19	24	24	21	18	21	17	21
Porto Alegre	20	B	5	89	20	14	18	16	21	22	22	22	17	22
Porto Alegre	21	A	4	87	17	16	17	20	17	37	34	36	37	38
Porto Alegre	22	A	4	120	23	25	21	32	19	39	37	32	38	38
Gravataí	1	A	9	251	52	49	39	55	56	50	42	50	40	50
Gravataí	2	A	9	260	63	54	45	40	58	40	50	40	40	50
Gravataí	3	A	9	241	41	43	56	47	54	43	50	50	50	40
Gravataí	4	A	9	198	1	40	49	56	52	40	40	40	42	45
Gravataí	5	A	9	235	58	38	47	46	46	50	40	40	50	46
Gravataí	6	A	9	250	53	52	58	42	45	45	40	40	50	50
Gravataí	7	A	9	226	50	55	38	39	44	50	43	50	50	50
Gravataí	8	A	9	213	48	48	26	52	39	50	42	50	50	46
Gravataí	9	A	9	185	35	48	41	26	35	42	40	40	40	45
Gravataí	10	A	9	202	47	33	42	43	37	50	40	40	50	50

Fonte: Autor, 2023

4.3 Resultados da distribuição geográfica

Levando em consideração todas as coordenadas geográficas disponíveis dos clientes, foram calculados os pontos centrais de cada dia de visita do vendedor a partir da latitude média e longitude média dos clusters. Os pontos centrais permitiram que fossem calculadas

as distâncias entre todos os PDV's e os pontos médios, conseguindo assim gerar a diferença proporcionada pela clusterização. Foi utilizada esta métrica por ser fácil e rápida de calcular, além de conseguir expressar a distribuição geográfica dos clusters em quilômetros.

Todos os cenários obtiveram redução da distância total conforme a Tabela 11. Seis dos nove cenários alcançaram reduções percentuais superiores 60%, sendo que os três maiores em termos de quantidade de clusters diminuíram a distância acima dos 64%. Atenta-se ao cenário 3, o qual o impacto percentual da clusterização foi mais brando que os demais. Na Tabela 11 é possível ver os cenários consolidados.

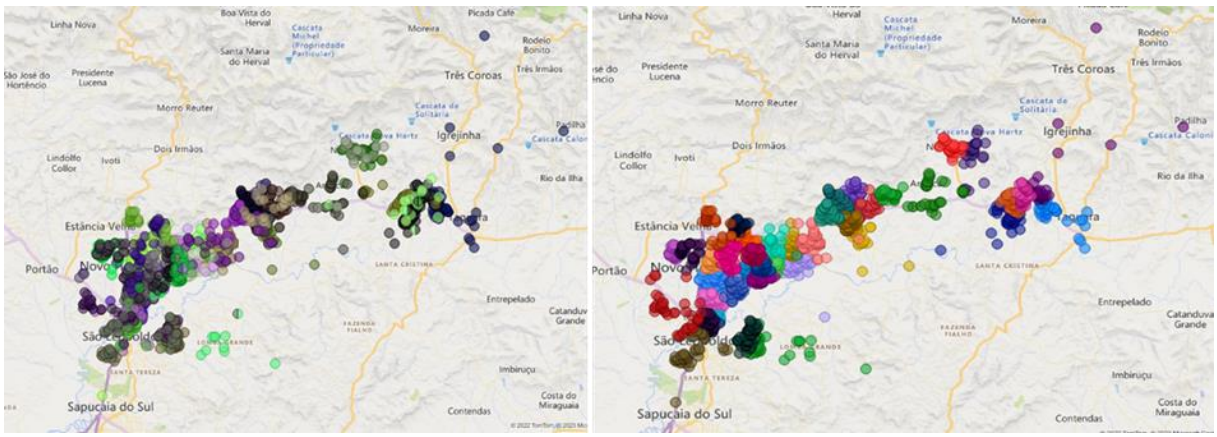
Tabela 11: Distância total quilômetros em todos os cenários dos PDV's frente aos pontos médios de visita do dia

Cenário	Cidade Atendida	Tipo de Vendedor	Clusters	Distância Pré (km)	Distância Pós (km)	Diferença percentual
1	Sapucaia	A	40	5810	1825	-69%
2	Sapucaia	B	5	264	152	-42%
3	Sapucaia	C	5	311	281	-10%
4	Porto Alegre	A	85	9934	2456	-75%
5	Porto Alegre	B	5	542	142	-74%
6	Porto Alegre	C	5	753	463	-38%
7	Porto Alegre	D	5	278	73	-74%
8	Porto Alegre	E	10	1930	416	-78%
9	Gravataí	A	50	6366	2272	-64%
Total	**-**	**-**	**210**	**26.188**	**8.079**	**-69,1%**

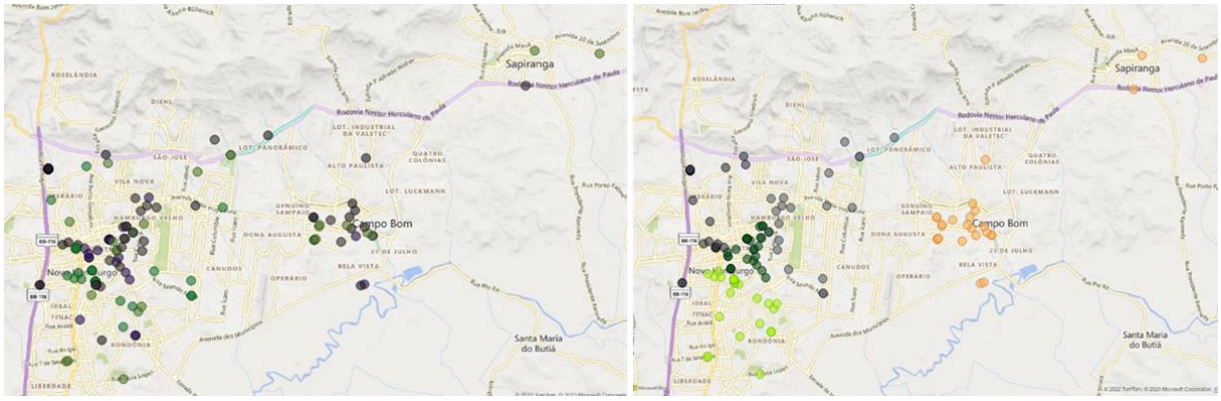
Fonte: Autor, 2023

4.3.1 Distribuição geográfica de Sapucaia

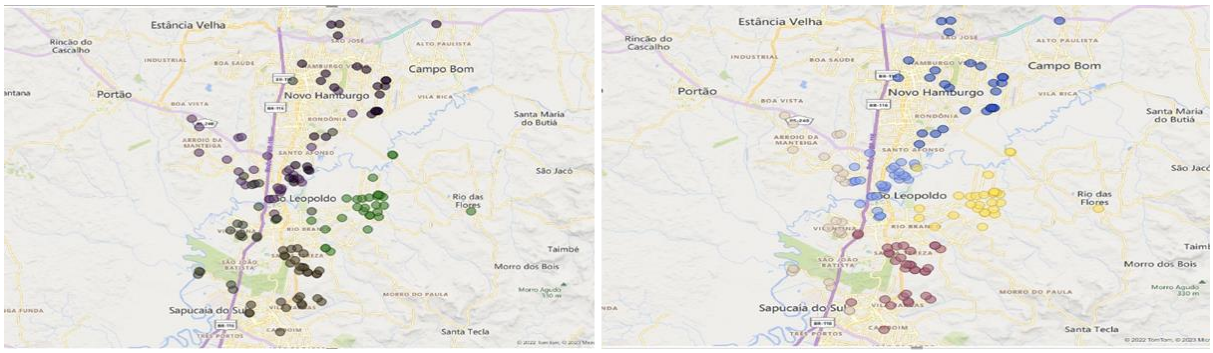
Na Figura 8 foram dispostos os PDV's dos cenários de Sapucaia, sendo que cada círculo representa um PDV e a coloração associada ao círculo equivale a um vendedor. Em Sapucaia, o cenário 1 atingiu a maior diferença percentual em distância. Em (b) é mostrado que não há sobreposição de clusters aparente. Foram mostrados na figura pontos bem distantes como os clientes do município de Três Coroas, os quais impactaram diretamente no tamanho do cluster.



(a) Antes da clusterização - Cenário 1 (b) Depois da clusterização – Cenário 1



(c) Antes da clusterização - Cenário 2 (d) Depois da clusterização – Cenário 2



(e) Antes da clusterização - Cenário 3 (f) Depois da clusterização – Cenário 3

Figura 8: Mapas rodoviários com os clusters gerados dos cenários de Sapucaia

Fonte: Autor, 2023

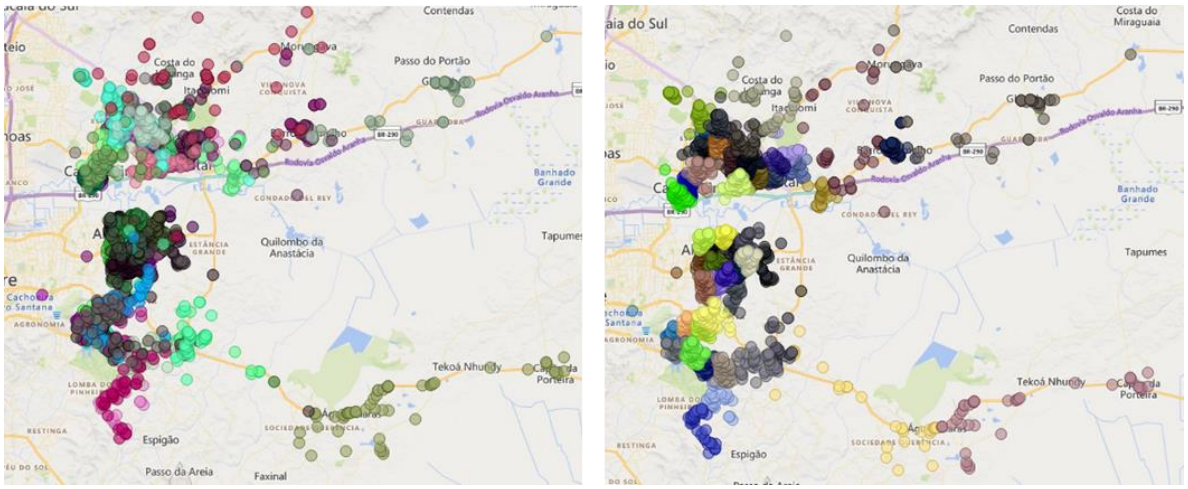
No cenário 2 e 3 ocorreram somente trocas de dia de atendimento dos clientes, considerando que havia somente 1 vendedor para cada um desses cenários. Enquanto em (c) aparenta não ter nenhuma divisão de regiões por dia de visita, (d) já estão bem delimitadas as fronteiras entre cada uma das regiões. Já no cenário 3, em (e) é visível uma organização de regiões de atendimento antes mesmo do algoritmo ser aplicado, impactando diretamente na diferença de distância total. São poucas alterações que ocorreram entre (e) e (f), também confirmando que organização anterior citada.

4.3.2 Distribuição geográfica de Gravataí

Em Gravataí, ocorreu a clusterização de clientes no cenário 9 somente, contudo entre todos os cenários esse era o segundo maior em quantidade de clusters. Conforme a Figura 9, este caso é muito semelhante ao cenário 1 de Sapucaia, pois tanto a quantidade de clusters como a homogeneidade da concentração de clientes eram próximas.

Diferentemente dos casos anteriores, percebeu-se a existência de regiões mais espaçadas. Algumas dessas regiões exibem PDV’s bem distantes e com maiores afinidades

a outros clusters do que o seu de origem. Já as regiões de centro das cidades apresentaram clusters uniformes e sem sobreposição deles.



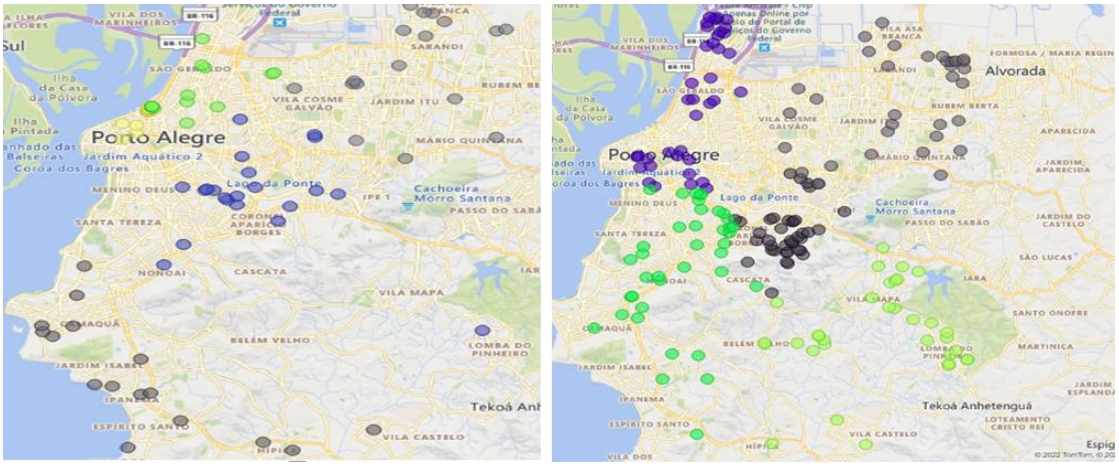
(a) Antes da clusterização - Cenário 9 (b) Depois da clusterização – Cenário 9

Figura 9: Mapas rodoviários com os clusters gerados no cenário de Gravataí

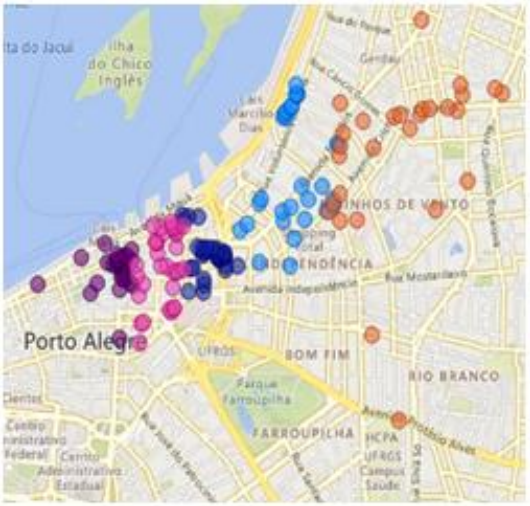
Fonte: Autor, 2023

4.3.3 Distribuição geográfica de Porto Alegre

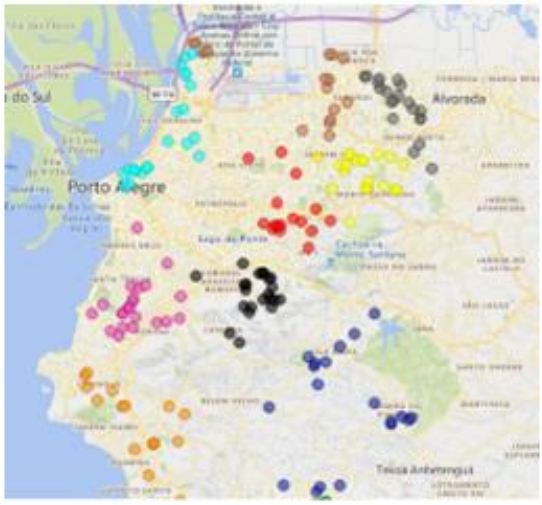
Semelhantemente aos cenários 1 e 9, o cenário 4 apresentou um contraste de regiões atendidas conforme a Figura 10 em (e). Por se tratar de clientes somente da mesma cidade, não foram retratadas regiões remotas nem impedimentos geográficos dentro dos clusters gerados.



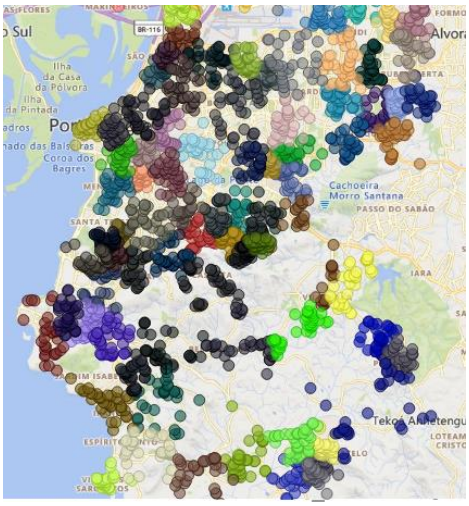
(a) Resultados do cenário 5 (b) Resultados do cenário 6



(c) Resultados do cenário 7



(d) Resultados do cenário 8



(e) Resultados do cenário 4

Figura 10: Mapas rodoviários com os clusters gerados nos cenários de Porto Alegre

Fonte: Autor, 2023

Os cenários com menores números de clusters apresentaram maiores inconsistências frente ao demais. Na Figura 10 (a), foi reparado um ponto azul muito distante dos demais. Já em (b) examinou-se a existência de interferências nas fronteiras dos clusters, visto que no bairro Cascata existem três clientes que estão muito próximos e são de clusters diferentes.

No item (c), a concentração de PDV's é muito próxima, o que ocasionou em clientes da mesma rua sendo alocados em clusters diferentes. Um ponto que chama atenção nesse item é um PDV do cluster azul numa região onde somente existiam clientes do cluster laranja. Por último, as regiões em (d) foram melhores distribuídas e obtiveram clusters mais espaçados. Contudo, poucos clientes da região próxima ao lago foram alocados na cor

marrom, sendo que o cluster dessa cor está muito longe do lago. Isto se deve aos clusters próximo destes pontos terem alcançado os valores máximos de *K* previamente definidos para cada cenário.

4.3.4 Frequência de visitas

A frequência de visita ao cliente tem um impacto social e monetário no mesmo. Com o intuito de aumentar esta frequência, foi definido como 38 o número ideal de atendimentos diários realizados pelos vendedores. Todos os clusters que ficaram com uma quantidade inferior de 38 clientes tiveram suas periodicidades de visitas alteradas como semanais. Já os clusters que possuem quantidades superioes a 38 PDV's, as visitas foram alteradas para quinzenais nos clientes excedentes.

Na Tabela 12, é apontada uma redução de 52,5% de PDV's quinzenais, enquanto a de semanais cresceu 56,3%. Porto Alegre foi a sala de vendas que mais teve alterações: aumentou em 63% a quantidade de PDV's semanais após a implentação da clusterização.

Tabela 12: Comparação da frequência de visitas aos clientes após a clusterização

		Pré Clusterização			**Pós Clusterização**	
Cidade Atendida	**PDV**	**Semanais**	**Quinzenais**	**Mensais**	**Semanais**	**Quinzenais**
Sapucaia	2043	901	817	325	1459	584
Porto Alegre	3814	2088	1602	124	3417	397
Gravataí	2261	1227	792	242	1717	544
Total	**8118**	**4216**	**3211**	**691**	**6593**	**1525**

Fonte: Autor, 2023

Outro dado importante a ser analisado é a quantidade de visitas mensais que cada sala de vendas teve de acréscimo. Segundo a Figura 11, as salas de vendas obtiveram adições superiores a 1.000 visitas mensais. Enquanto Porto Alegre conseguiu o maior acréscimo, sendo 2.782 visitas mensais, Sapucaia obteve o maior crescimento percentual de visitas: 25,9% frente ao cenário inicial.

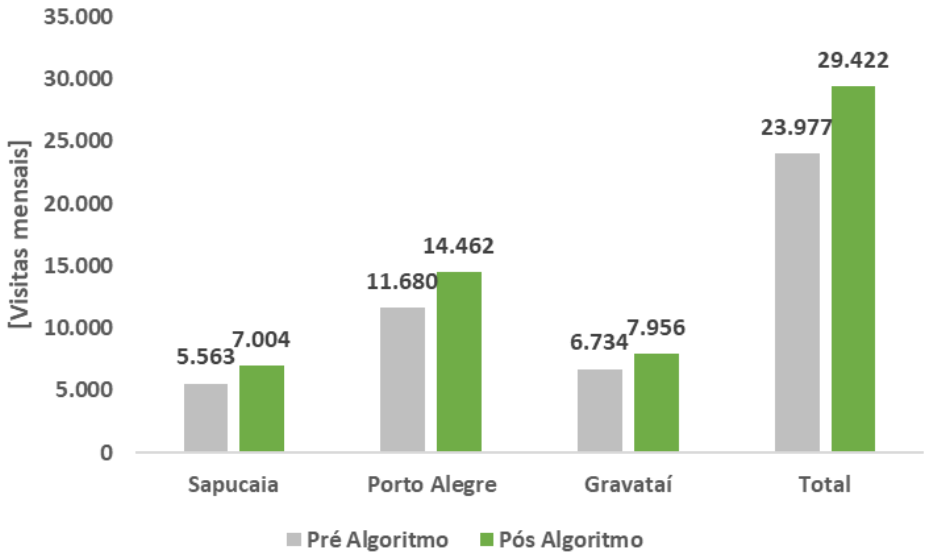


Figura 11: Total de visitas mensais a clientes em Porto Alegre, Sapucaia e Gravataí
Fonte: Autor, 2023

O resultado que mais se destacou na aplicação da técnica de *clustering* foi a redução da distância total dos PDV's em relação aos centróides de seus respectivos clusters: 69,1% do tamanho original. Em outros trabalhos, a aplicação de métodos de segmentação de dados também gerou reduções na jornada de equipes de atendimento, como por exemplo a redução de 23% do tempo de re-estabelecimento de energia em uma empresa elétrica no Irã (Yousefi e Hadi-Vencheh, 2023).

5 Conclusões e Trabalhos Futuros

O objetivo principal deste trabalho foi avaliar metodologias para segmentação geográfica de clientes através de métodos de segmentação estatística de dados. Tendo em vista a complexidade de atender todos os pontos de venda visando com a menor distância percorrida, a clusterização buscou diminuir a distância total entre todos estes pontos.

Para esta clusterização, foram estudados algoritmos para a segmentação de dados e avaliados os algoritmos KM, KMC++ e SVM. A partir disso, foi feita uma triagem e o KM foi o que obteve resultados mais consistentes entre os três. Assim, definiram-se quais seriam as peculiaridades de cada cenário em que seria aplicado o algoritmo.

Durante o trabalho, abordou-se três cidades atendidas pelas equipes de vendas estavam com visitas mal distribuídas geograficamente e quantitativamente. Tendo em vista isto, foram alcançados números expressivos de redução. O que mais se destacou foi a diminuição da distância dos clientes em relação aos pontos centrais de seus clusters: 69%. Quanto menores forem as distâncias que os vendedores percorrerem, menor serão os custos envolvidos de transporte.

Além de custos de transporte, o impacto ambiental que o trabalho pode atingir é notável, pois menos gasolina será necessária para completar os trajetos das visitas, evitando assim a combustão de CO_2 para a atmosfera. Também deve ser pontuado o possível impacto na saúde dos colaboradores da empresa, porque menor será a distância necessária para os mesmos atenderem.

Em termos de vendas e fidelidade ao cliente, é possível alcançar um aumento de 25,9 % de visitas de atendimento sem que fosse necessário realizar prospecção de clientes ou aumentar custos honorários para trabalhadores. Além disso, a produtividade discrepante de visitas entre os colaboradores da empresa foi nivelada, o que implica um padrão mais ajustado da forma que é feito o atendimento de vendas.

Ao analisar-se visualmente os clusters gerados pelo algoritmo, notou-se que a sua efetividade foi melhor em cenários com maiores quantidades de PDV's e clusters. Nestes casos, foi analisada pouca sobreposição de regiões de visita como também clusters poucos dispersos. Já em clusters menores, a ocorrência de PDV's distantes do ponto médio da região foi mais recorrente. Atribuí-se isso ao fato da definição arbitrária de K máximo e K mínimo.

5.1 Trabalhos futuros

O *K-means* teve como principal fator de escolha a fácil aplicação e a velocidade com dimensão elevada de informações. Mesmo criando clusters bem definidos e bem distribuídos, em certos casos tornaram-se necessários ajustes no cluster em que certo PDV foi alocado. Existem já algoritmos de *K-means* que, ao invés de utilizar coordenadas geográficas, usam o tempo médio de trânsito e estadia em cada cliente para calcular o cluster. Além de ter uma maior precisão, visto que barreiras geográficas serão contornadas, maior será o nível de informação para construção de modelos e tendências de atendimento.

Outro ponto que seria interessante explorar é a quantidade ideal de vendedores necessários para atender o cenário atual. Tendo em conta que o trabalho foi montado pensando em um número já definidos de clusters, não foi possível analisar qual a quantidade de mínima de clusters a que atenderia para obter a maior produtividade possível.

O algoritmo SVM é uma metodologia de aprendizado superviosionado que funciona muito bem com dados já classificados. Sendo assim, uma cenário que poderia ter sido testado é a utilização do SVM após a aplicação do *K-means,* visando que o SVM aprendesse a classificar e assim corrigisse pequenos desvios gerados pelo KM.

Por último, teria sido interessante ter sido calculado o impacto gerando no tempo de trajeto dos vendedores após a aplicação da segmentações geográfica. Para isso, um trabalho futuro seria a utilização de algum algoritmo heurístico parar calcular a melhor rota antes e depois da aplicação do *clustering,* possibilitando dessa maneira o comparativo.

REFRÊNCIAS

ALRASHIDI, M.; RAHMAN, S. Short-term photovoltaic power production forecasting based on novel hybrid data-driven models. **Journal of Big Data**, v. 10, n. 1, 1 dez. 2023.

ASHUTOSH BHARDWAJ. **Silhouette Coefficient**. Disponível em: <https://towardsdatascience.com/silhouette-coefficient-validating-clustering-techniques-e976bb81d10c>. Acesso em: 19 mar. 2023.

ASSOCIAÇÃO BRASILEIRA DA INDÚSTRIA DE ALIMENTOS. **ABIA**. Disponível em: <https://www.abia.org.br/downloads/Infograficoanual2020v5.pdf>. Acesso em: 18 mar. 2023.

CERVIERI JÚNIOR, O.; RODRIGUES TEIXEIRA JUNIOR, J.; GALINARI, R.; et al. O setor de bebidas no Brasil_P_.pdf. [s.d.].

GAJA, H.; LIOU, F. Defects monitoring of laser metal deposition using acoustic emission sensor. **International Journal of Advanced Manufacturing Technology**, v. 90, n. 1–4, p. 561–574, 1 abr. 2017.

HAITIAN WEI. **Measuring Clusters** . Disponível em: <https://medium.com/@haataa/how-to-measure-clustering-performances-when-there-are-no-ground-truth-db027e9a871c#id_token=eyJhbGciOiJSUzI1NiIsImtpZCI6IjVkZjFmOTQ1ZmY5MDZhZWFlZmE5M2MyNzY5OGRiNDA2ZDYwNmIwZTgiLCJ0eXAiOiJKV1QifQ.eyJpc3MiOiJodHRwczovL2FjY291bnRzLmdvb2dsZS5jb20iLCJuYmYiOjE2NzkyODU0NzMsImF1ZCI6IjIxNjI5NjAzNTgzNC1rMWs2cWUwNjBzMnRwMmEyamFtNGxqZGNtczAwc3R0Zy5hcHBzLmdvb2dsZXVzZXJjb250ZW50LmNvbSIsInN1YiI6IjExMTYwMzEwNjc0Mzc0ODUzMTUzNiIsImVtYWlsIjoiZ3VpbGhlcm1la2F1ZXJkZW5hZGFsQGdtYWlsLmNvbSIsImVtYWlsX3ZlcmlmaWVkIjp0cnVlLCJhenAiOiIyMTYyOTYwMzU4MzQtazFrNnFlMDYwczJ0cDJhMmphbTRsamRjbXMwMHN0dGcuYXBwcy5nb29nbGV1c2VyY29udGVudC5jb20iLCJuYW1lIjoiR3VpIERlIE5hZGFsIiwicGljdHVyZSI6Imh0dHBzOi8vbGgzLmdvb2dsZXVzZXJjb250ZW50LmNvbS9hL0FHTm15eFp6OU90aHBYRVAxcURPbFBzSVhyLTNXV0tNcmNvTmI0VHlFckZpSmc9czk2LWMiLCJnaXZlbl9uYW1lIjoiR3VpIiwiZmFtaWx5X25hbWUiOiJEZSBOYWRhbCIsImlhdCI6MTY3OTI4NTc3MywiZXhwIjoxNjc5Mjg5MzczLCJqdGkiOiIwMGE5MDI0MjcwNDJlZDFiNTQzOWY2YzM3ZmU1NWJkNGM1ZTc4YmZkIn0.daHmDTWOQxRAnrv3zhStvL672ccyqY9xajihLtsCupstZZPQRAdQO2PLVha6NzLyArDz-HtLvMSLrDQ-_36MvuSC2Mg8lihcQxZD1fDE7BcQiRKBNNLz48AshTwqLelRgjl6sdlyPd1fAoFuBUyrvEgFLRRN0FzhY94wkYsqLxZRYLoLIajlV0HeVkIn-UWTUuX-QEFwCyXow-JeX1gocvwnklV2k48UugIqAmWAM0vlDMKcI2PLNxGVxVToSLatRHfCMOustbl8OXh0RnRcxDo65x3nFZuAXf6YFdFS3KunPABQhpZYvjCC5sW12b03EhXgSeI2Rx7FqGgW_s4jsA>. Acesso em: 19 mar. 2023.

IBM. **How SVM Works**. Disponível em: <https://www.ibm.com/docs/en/spss-modeler/saas?topic=models-how-svm-works>. Acesso em: 18 mar. 2023.

Investimento em TI chegará a R$345 bilhões até o fim de 2022 – IPNews – Comunicação Interativa. Disponível em: <https://ipnews.com.br/investimento-em-ti-chegara-a-r345-bilhoes-ate-o-fim-de-2022/>. Acesso em: 6 mar. 2023.

JASSIM, M. A.; ABD, D. H.; OMRI, M. N. Machine learning-based new approach to films review. **Social Network Analysis and Mining**, v. 13, n. 1, 1 dez. 2023.

JOSHI, S. **Types of Clustering Algorithms in Machine Learning With Examples**. Disponível em: <https://www.analytixlabs.co.in/blog/types-of-clustering-algorithms/>. Acesso em: 21 fev. 2023.

LUCENA, W. **Agrupamento hierárquico. Agrupamento hierárquico ou Hierarchical… | by William Lucena | Medium**. Disponível em: <https://medium.com/@will.lucena/agrupamento-hier%C3%A1rquico-329e30a9f32d>. Acesso em: 21 fev. 2023.

MONKEY LEARN. **Support Vector Machines (SVM) Algorithm Explained**. Disponível em: <https://monkeylearn.com/blog/introduction-to-support-vector-machines-svm/>. Acesso em: 18 mar. 2023.

OYEWOLE, G. J.; THOPIL, G. A. Data clustering: application and trends. **Artificial Intelligence Review 2022**, p. 1–37, 27 nov. 2022.

PIECH, C. **K Means:The Basic Idea**. Disponível em: <https://stanford.edu/~cpiech/cs221/handouts/kmeans.html>. Acesso em: 21 fev. 2023.

QIN, S.; HOU, X.; WEN, Y.; et al. Machine learning classifiers for screening nonalcoholic fatty liver disease in general adults. **Scientific reports**, v. 13, n. 1, p. 3638, 1 dez. 2023.

SCIKIT-LEARN. **Clustering Algorithms - Scikit-learn**. Disponível em: <https://scikit-learn.org/stable/modules/clustering.html>. Acesso em: 18 mar. 2023.

SHIPRA SAXENA. **How does SVM work**. Disponível em: <https://www.analyticsvidhya.com/blog/2021/03/beginners-guide-to-support-vector-machine-svm/>. Acesso em: 18 mar. 2023.

SURYA PRIY. **Clustering in Machine Learning**. Disponível em: <https://www.geeksforgeeks.org/clustering-in-machine-learning/>. Acesso em: 18 mar. 2023.

TABIANAN, K.; VELU, S.; RAVI, V. K-Means Clustering Approach for Intelligent Customer Segmentation Using Customer Purchase Behavior Data. **Sustainability 2022, Vol. 14, Page 7243**, v. 14, n. 12, p. 7243, 13 jun. 2022.

Understanding K-Means, K-Means++ and, K-Medoids Clustering Algorithms | by Satyam Kumar | Towards Data Science. Disponível em: <https://towardsdatascience.com/understanding-k-means-k-means-and-k-medoids-clustering-algorithms-ad9c9fbf47ca>. Acesso em: 18 mar. 2023.

YOUSEFI, A.; HADI-VENCHEH, A. Resiliency and reliability of the power grid in the time of COVID-19: An integrated ABC-K-means model for optimal positioning of repair crew. **Electric Power Systems Research**, v. 216, p. 109022, 1 mar. 2023.