Avaliação da eficiência de um algoritmo que utiliza inteligência artificial para agrupar pessoas de perfis similares, que utilizam da caminhada como transporte no dia a dia, para que se locomovam com mais segurança

Gustavo Oliveira Melo, Jair Angeluci Neto, Leonardo Elis da Silva

Departamento de Engenharia e Tecnologia – Ciência da Computação
Universidade Anhembi Morumbi (UAM) – São Paulo, SP – Brasil

Resumo. *A segurança de locais públicos é algo bastante preocupante, ainda mais quando se está desacompanhado, gerando medo e insegurança. O agrupamento de pessoas é uma solução a se considerar com o objetivo de reduzir este problema, e a utilização da tecnologia pode promover um grande auxílio nesse processo. Este artigo descreve a utilização de um algoritmo de Aprendizagem de Máquina não supervisionada para identificar grupos similares, para que possam se locomover por lugares públicos com maior segurança e conforto, havendo afinidade entre as pessoas de um grupo.*

Palavras-chave: Clusterização, K-means, Perfil similar, Segurança

Abstract. *The safety of public places is a matter of great concern, especially when unaccompanied, generating fear and insecurity. The clustering of people is a solution to consider in order to reduce this problem, and the use of technology can promote a great help in this process. This article describes the use of an unsupervised Machine Learning algorithm to identify similar groups, so that they can move around public places with greater security and comfort, with an affinity between the people of a group.*

Keywords: Clustering, K-means, Similar profile, Safety

1. Introdução

No mundo, pode-se observar que aproximadamente 30% da população tem medo da violência. Já no Brasil, os números são bem mais expressivos, pois segundo os dados da Fundação Getúlio Vargas Social (2017), 68% das pessoas sentem esse medo de violência, o que coloca o país em segundo lugar no mundo.

Mais especificamente, referindo-se a caminhar pelas ruas no Brasil, o cenário se mostra bastante significativo, e a situação fica ainda mais grave quando se trata de mulheres ou a caminhada está sendo feita durante a noite. Ainda, conforme a Fundação Getúlio Vargas Social (2017), no Brasil, 61% das pessoas moradoras de pequenas cidades sentem medo de andar sozinhos na rua à noite. Em relação a grandes centros urbanos, esse número aumenta para 75%. No mundo, referindo-se a gênero, 24% dos homens sentem esse medo, enquanto as mulheres são 35%. No Brasil o número é bem maior: 60% dos homens e 76% das mulheres. Essa diferença entre os gêneros acontece porque as mulheres têm uma percepção maior de vulnerabilidade do que os homens. Além disso, alguns tipos de crimes violentos, como assédio e agressão sexual, favorecem para o aumento do medo das mulheres (FACULTY OF SECURITY, 2013).

Esse medo, sentido ao caminhar em lugares públicos, é reduzido quando as pessoas estão andando em grupo ou, simplesmente, não estão sozinhas (DE SOUSA GUEDES, 2016). Também é importante observar que, de acordo com Tuan (2005), um dos motivos para o estabelecimento de conexões e laços entre as pessoas é o medo. Como exemplo, é citado que algumas cidades italianas medievais eram compostas de núcleos familiares reforçados e essas conexões entre as pessoas eram feitas pela necessidade e pelo medo.

Tendo em vista as informações apresentadas, e considerando que a segurança das rotas é um elemento bastante importante na decisão do caminho, sendo o terceiro fator mais importante nos dias de semana e primeiro nos fins de semana (AMIRGHOLY et al., 2017), a utilização da tecnologia como auxílio para gerar um agrupamento de pessoas pode ser uma alternativa para minimizar esse problema, levando em consideração similaridades entre as pessoas, com o objetivo de promover maior conforto, confiança e segurança.

Diante disso, este trabalho tem como objetivo utilizar um algoritmo de Machine Learning (K-means) para realizar o agrupamento de pessoas que utilizam da caminhada como transporte no dia a dia, bem como analisar a eficiência de um algoritmo que utiliza inteligência artificial para agrupar pessoas de perfis similares. Esse agrupamento é feito com base em semelhanças de perfil entre as pessoas, já que essa similaridade promove uma comunicação mais natural, maior grau de concordância com o outro e as relações são mais satisfatórias (BYRNE, 1971 apud SILVEIRA; HANASHIRO, 2009).

2. Fundamentação Teórica

2.1. Aprendizagem de máquina

Tendo em vista que o objetivo deste trabalho é utilizar a Inteligência Artificial para melhorar a caminhada das pessoas por locais públicos, a pesquisa sobre os fundamentos teóricos, tais como a definição de Inteligência Artificial e as técnicas utilizadas para a implementação, é de grande relevância.

Segundo Andreas Kaplan e Michael Haenlein (2019), Inteligência Artificial pode ser definida como a habilidade do sistema de interpretar dados externos corretamente, aprender a partir desses dados e usar esses aprendizados para atingir objetivos e tarefas específicas através de adaptação flexível.

A Inteligência Artificial encontra-se em estágio evolutivo, visto que projetos envolvendo o assunto vêm sendo cada vez mais comuns e é crescente a presença de IA no dia-a-dia das pessoas e empresas, além de haver metas ainda não alcançadas com relação às suas produções (GESSINGER; HAMMES; COLLING, 2019).

Aprendizagem de máquina pode ser definida como uma técnica de aprendizado baseado na experiência, ou seja, a máquina adquire conhecimento à medida que vai efetuando tarefas, possibilitando a resolução de um problema específico (GESSINGER; HAMMES; COLLING, 2019), o que se adequa aos objetivos deste trabalho, que procura utilizar um algoritmo que possa interpretar as informações recebidas para realizar o agrupamento dos dados, de acordo com as semelhanças entre eles, utilizando a aprendizagem não-supervisionada.

Além disso, Machine Learning é uma sub-área, que estuda técnicas computacionais para adquirir novos conhecimentos, novas habilidades e novos meios de organizar o conhecimento já existente. Ela é considerada uma subcategoria muito importante de Inteligência Artificial, pois a capacidade de aprender é essencial para um comportamento inteligente (GESSINGER; HAMMES; COLLING, 2019).

Existem dois tipos de técnicas utilizadas na aprendizagem de máquina: a aprendizagem supervisionada e a aprendizagem não supervisionada. Enquanto a aprendizagem supervisionada treina um modelo a partir de dados conhecidos de entrada e saída para poder prever resultados, a aprendizagem não-supervisionada encontra padrões ocultos com base em dados de entrada (BUENO, 2019).

2.1.1. **K-means**

Com o objetivo de analisar as respostas de cada pessoa e agrupá-las de acordo com seus perfis, o algoritmo de agrupamento utilizado neste trabalho é o K-means, algoritmo de aprendizagem não supervisionada usado no campo de machine learning (PIMENTEL; DE FRANÇA; OMAR, 2003). A escolha para a utilização desse algoritmo deve-se ao fato de que o K-means é simples e eficiente, realizando cálculos simples de maneira rápida, ocasionando baixa quantidade de armazenamento de informações a serem processadas (DE CASTRO; DO PRADO, 2002).

O algoritmo K-means tem como objetivo encontrar o melhor agrupamento entre diferentes dados, de maneira que cada dado seja pertencente ao grupo do centróide mais próximo do mesmo (PIMENTEL; DE FRANÇA; OMAR, 2003).

O funcionamento do algoritmo K-means se dá pelos seguintes passos: recebe um grupo de dados; define aleatoriamente os K centróides; calcula distância de todos os dados para cada centróide; inclui cada dado no grupo mais próximo; calcula novos centróides de acordo com a média dos vetores de cada grupo; recalcula as distâncias dos dados para os novos centróides; realoca cada dado no grupo mais próximo; repete os dois passos anteriores até que os dados já estejam nos grupos dos seus vetores médios mais próximos (PIMENTEL; DE FRANÇA; OMAR, 2003).

Para realizar o cálculo das distâncias, foi utilizada a Distância euclidiana, que possui a seguinte fórmula:

$$E(x, y) = \sqrt{\sum_{i=0}^{n} (x_i - y_i)^2}$$

Onde x é um dado da base de dados, y é um dado do centróide e n é o número de tuplas.

2.2. **Trabalhos Similares**

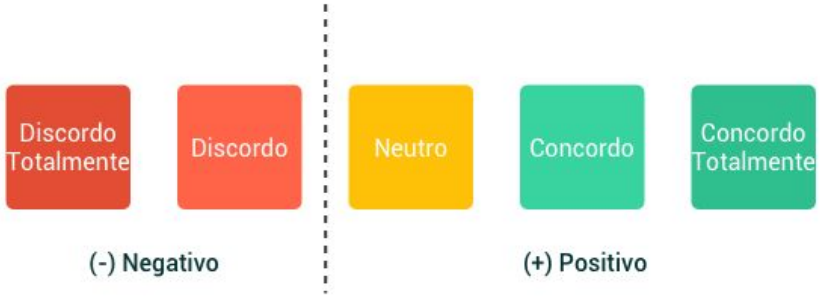
É importante que, no contexto deste trabalho, entenda-se que as características pessoais são de grande importância na formação de um grupo, promovendo maior conforto, uma comunicação mais natural e relações mais satisfatórias (BYRNE, 1971 apud SILVEIRA; HANASHIRO, 2009).

A identificação de grupos similares de pessoas pode ser uma tarefa bastante complicada para o ser humano, já que esse processo pode envolver muitas informações (PIMENTEL.; DE FRANÇA; OMAR, 2003). Considerando isso, Pimentel, De França e Omar (2003), utilizaram o K-means para agrupar estudantes com características similares, levando em conta o grau de confiança de cada aluno nos tópicos de uma disciplina, a fim de personalizar o ensino presencial, permitindo que o professor possa fazer uso de uma pedagogia diferenciada adequada**.** Esse grau de confiança foi obtido a partir de um questionário com perguntas baseadas na escala Likert.

Já Quinteiro (2011), utilizou o K-means para agrupar pessoas que gostam de música e que possuem uma conta ativa no Facebook. Esse agrupamento foi feito com o intuito de aprimorar as estratégias de marketing para obter um melhor direcionamento de produtos e serviços.

Para complementar e definir as características do usuário, foi levado em consideração informações demográficas das pessoas, além de características relacionadas a frequências de modo, aquisição e audição dos vários tipos de música. As características também foram obtidas a partir de um questionário contendo perguntas baseadas na escala Likert.

Figura 1. Escala Likert



Além disso, a técnica de validação utilizada neste trabalho foi inspirada no projeto desenvolvido por Matte (2020), que explorou técnicas de otimização de desempenho do algoritmo K-means.

Tendo em vista os trabalhos similares citados, esse trabalho reforça a importância de se agrupar pessoas por meio de técnicas de clusterização e utiliza uma metodologia parecida com as citadas, considerando características semelhantes.

3. Metodologia

Este trabalho envolve uma pesquisa descritiva, visto que a pesquisa feita neste trabalho envolve questionário e a necessidade de estabelecer relações entre as diferentes pessoas entrevistadas de acordo com suas características.

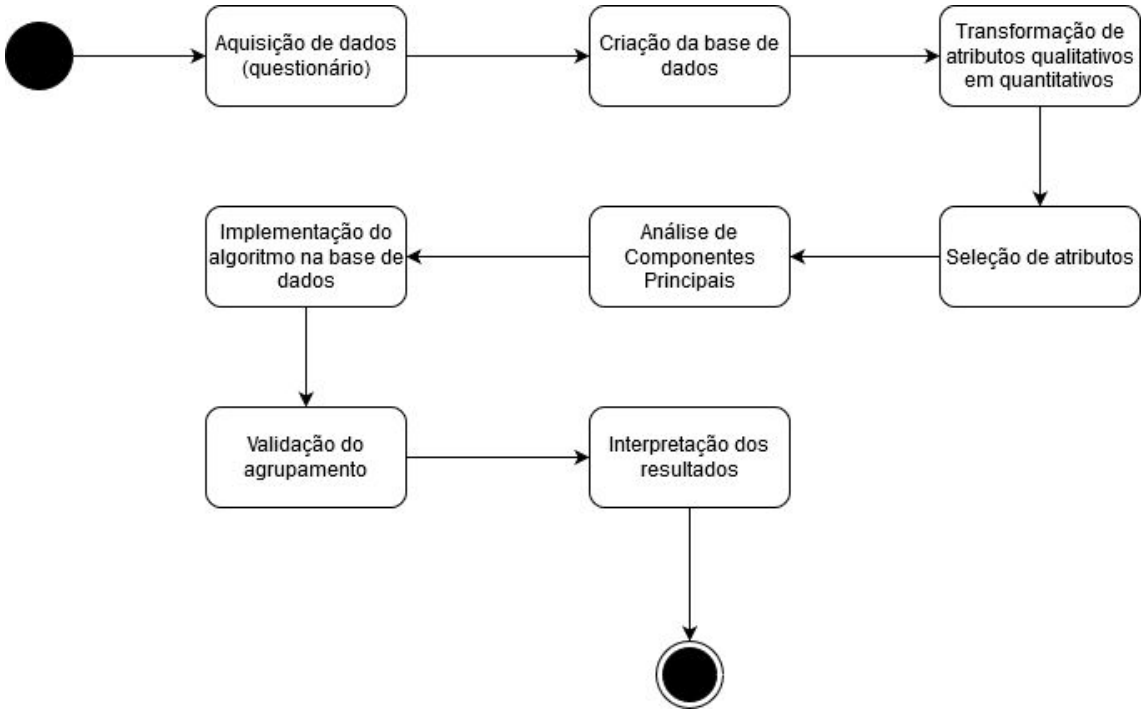
Quanto aos procedimentos técnicos, são abordados:

- Levantamento: foi realizada um questionário para levantar dados e, a partir da análise desse levantamento, foi implementada a base de dados.
- Experimental: com os dados obtidos do levantamento, realizou-se a manipulação dos dados e, posteriormente, uma avaliação dos resultados.

A análise dos dados foi feita de forma quantitativa, analisando dados estruturados e estatísticos para formar uma base de dados que possa ser analisada pelo algoritmo.

Para alcançar o objetivo geral desta pesquisa, foram estruturados os procedimentos metodológicos conforme representado na Figura 2.

Figura 2. Fluxograma da metodologia



A Figura 2 representa a sequência de etapas seguidas neste trabalho, indicando o que foi feito durante cada fase de construção. Para a realização do questionário, foi utilizado o Google Forms, para que fosse respondido online e permitisse maior alcance. Assim como Pimentel, De França e Omar (2003) e Quinteiro (2011), o questionário foi elaborado com perguntas na escala Likert e com perguntas demográficas. Para a construção da base de dados foi utilizado um arquivo CSV.

Depois, na fase de pré-processamento, foi feita a transformação de atributos qualitativos em quantitativos, tal como Matte (2020). Após isso, foi realizada a seleção dos atributos considerados mais relevantes para o agrupamento. Ainda na fase de

pré-processamento, foi empregada a Análise de Componentes Principais (PCA), que está presente no trabalho de Quinteiro (2011).

Para a implementação do algoritmo (K-means) foi utilizado a linguagem de programação Python juntamente com a biblioteca Scikit-learn, pela facilidade de implementação de diferentes funções e ferramentas.

Por fim, na fase de validação e interpretação dos resultados, assim como Matte (2020), foi aplicado um índice de validação interno, o índice silhueta.

3.1. **Estratégia de Validação**

Para a validação da pesquisa foi utilizado um índice de validação interno. Esse tipo de índice avalia o agrupamento com base nas informações dos próprios dados, sem nenhum conhecimento externo ao agrupamento (NALDI, 2011). Há diversos índices internos e especificamente para este trabalho foi utilizado o índice silhueta.

O índice silhueta mede a qualidade do agrupamento a partir da proximidade entre os elementos de um mesmo grupo e da distância entre esses elementos e os elementos do grupo mais próximo (NALDI, 2011). Ou seja, o índice silhueta avalia a qualidade do agrupamento levando em conta a compactação do grupo e a separação entre os grupos (MATTE, 2020).

Para cada objeto x_i pertencente a um grupo C_a, seu índice silhueta é calculado por:

$$silhouette(\mathbf{x}_i) = \frac{b(\mathbf{x}_i) - a(\mathbf{x}_i)}{max\{a(\mathbf{x}_i), b(\mathbf{x}_i)\}}$$

Onde a(x_i) é a média das distâncias de x_i a cada um dos outros objetos pertencentes ao grupo C_a e b(x_i) é o menor valor entre as médias das distâncias de x_i a todos os objetos pertencentes aos outros grupos que sejam diferentes de C_a (NALDI, 2011).

O valor da silhueta de cada objeto estará no intervalo [-1,1]. Se a silhueta estiver próxima a 1, então o objeto foi agrupado no grupo correto. Já se a silhueta estiver mais próxima a -1, então é bem provável que o objeto tenha sido agrupado no grupo errado (MATTE, 2020).

A partir do cálculo da silhueta de cada objeto, é possível calcular o índice silhueta para a base de dados como um todo, apenas fazendo a média das silhuetas de todos os objetos (NALDI, 2011). Esse cálculo pode ser representado pela seguinte fórmula:

$$ASWC(\pi) = \sum_{i=1}^{n_o} \frac{silhouette(\mathbf{x}_i)}{n_o}$$

O resultado da silhueta média irá determinar a qualidade do agrupamento. Se o resultado for menor ou igual a 0.25, então nenhuma estrutura substancial foi encontrada. Uma estrutura de agrupamento fraca e possivelmente artificial é encontrada quando esse

valor estiver entre 0.26 e 0.50. Se a silhueta média estiver entre 0.51 e 0.70, tem-se uma estrutura razoável. Por fim, quando o resultado pertencer ao intervalo de 0.71 a 1.00, a estrutura de agrupamento é consistente (FADEL; SEMAAN; BRITO, 2014).

4. Desenvolvimento

Com o objetivo de realizar o agrupamento, foi criado um questionário para a coleta de diferentes perfis, que foram analisados pelo algoritmo. Para a construção do questionário, foi utilizada a ferramenta Google Forms, aplicativo de administração de pesquisas fornecido pelo Google. O formulário contém 19 perguntas baseadas na escala Likert, variando de 1 a 5, e foram obtidas 333 respostas. A partir dessas respostas, foi criada uma base de dados CSV, com 333 dados (cada um representando um perfil), com 19 atributos cada (cada atributo representando a resposta de 1 pergunta).

Após a montagem da base de dados, realizou-se uma análise dos atributos para que fossem identificados os melhores para o agrupamento, a partir de testes feitos aplicando o algoritmo K-means a cada atributo individualmente, conforme a Tabela 1.

Tabela 1. Índice silhueta de cada atributo

Atributo	Número de grupos (K)	Maior Índice silhueta	Número de grupos (K)	Segundo maior Índice
Idade	5	1	4	0.921
Escolaridade	6	1	5	0.937
Estado civil	4	1	3	0.986
Fumar	5	1	4	0.927
Conversar	5	1	4	0.961
Animais de estimação	5	1	4	0.946
Custo	5	1	4	0.977
Tempo	5	1	4	0.969
Conforto	5	1	4	0.992
Segurança	5	0.997	4	0.995
Qualidade visual	5	1	4	0.944
Frequência - trem	5	1	4	0.987
Frequência - metrô	5	1	4	0.979
Frequência - ônibus	5	1	4	0.927
Frequência - caminhada	5	1	4	0.918
Frequência - bicicleta	5	1	4	0.9688
Frequência - manhã	5	1	4	0.940
Frequência - tarde	5	1	4	0.934
Frequência - noite	5	1	4	0.885

Para cada atributo foi selecionado o número de K como a quantidade de opções de resposta no questionário. Por exemplo, para a pergunta sobre escolaridade, haviam seis respostas possíveis (Ensino Fundamental ou menos, Ensino Médio incompleto, Ensino Médio completo, Ensino Superior incompleto, Ensino Superior completo e Pós-graduação). Já para o estado civil, apenas quatro (Solteiro(a), Casado(a),

Divorciado(a) e Viúvo(a)). Para todas as outras perguntas, haviam cinco opções de resposta.

Todos os atributos obtiveram 1 como resultado do índice silhueta, menos segurança, e, considerando isso, fez-se necessária mais uma execução para definir os melhores atributos. No próximo passo, foi considerado o segundo maior valor do índice silhueta para cada atributo individualmente. Os atributos selecionados (em vermelho) foram aqueles que obtiveram as melhores médias de Índice silhueta (considerando o maior e o segundo maior índice): Estado civil, Conforto, Segurança, Frequência - trem e Frequência - metrô.

Outra estratégia utilizada foi a Análise de Componentes Principais (PCA). A PCA é uma técnica da estatística multivariada que converte um conjunto de atributos originais em um outro conjunto de atributos denominados de componentes principais. As componentes principais devem conter a menor perda de informação possível, considerando a variação total dos dados originais (VARELLA, 2008).

O uso da PCA reduz a dimensionalidade dos dados, gerando um conjunto de novas variáveis que não são correlacionadas umas com as outras (GOTELLI; ELLISON, 2016). Além de melhorar a visualização dos dados, a PCA pode ajudar muito quando é utilizado o aprendizado não supervisionado em uma base de dados (ZHU; IDEMUDIA; FENG, 2019), que é o caso deste trabalho. Isso pode ser explicado pelo fato de que essa técnica permite a eliminação de redundância e características menos importantes (QUINTEIRO, 2011).

4.1. **Resultados**

Para os testes do algoritmo, foram utilizadas todas as técnicas de pré-processamento descritas anteriormente. Os resultados dos testes são apresentados na Tabela 2.

Tabela 2. Resultados

Número de grupos (K)	Variância Explicada (PCA)	Índice Silhueta
5	74,49%	0.6195934820581069
20	74,49%	0.7539490826527104
35	74,49%	0.8163691229222759
50	74,49%	0.8549240769827527
65	74,49%	0.8844035138683003

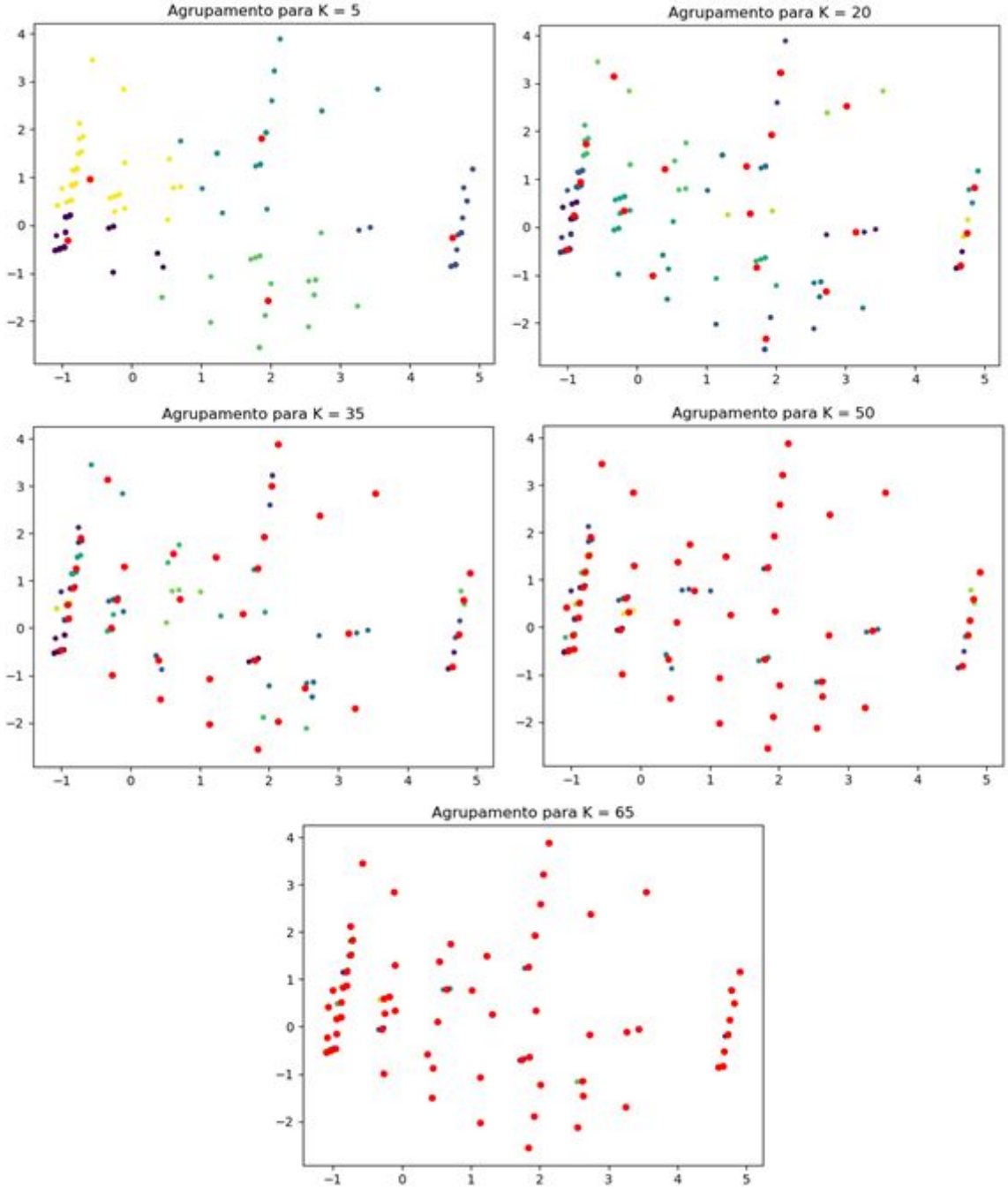
Foram feitos cinco testes, com K = 5, 20, 35, 50 e 65. Em todos os testes foram selecionados os atributos mais relevantes e aplicada a Análise de Componentes Principais (PCA). A PCA, considerando apenas os dois primeiros componentes principais, obteve uma variância explicada de cerca de 74,49%, ou seja, os dois novos eixos conseguem explicar cerca de 74,49% da variância de todos os dados.

De acordo com os testes, pode-se observar que para K = 5, tem-se uma estrutura razoável, já que o índice silhueta está entre 0.51 e 0.70. Para K = 20, 35, 50 e 65, os valores dos índices de silhueta foram considerados satisfatórios, pois os índices estão entre 0.71 e 1.00. Porém com K = 65, foram obtidos 28 grupos com apenas 1 dado, o

que não condiz com a proposta do trabalho que é agrupar pessoas. A mesma coisa acontece com K = 50, obtendo 17 grupos com apenas 1 dado e com K = 35, 5 grupos com apenas 1 dado foram gerados.

Diante dessa análise e considerando os testes feitos, pode-se afirmar que com a base de dados utilizada e todas as técnicas de pré-processamento empregadas, o agrupamento gerado para um K = 20 é o mais adequado. Já que além de obter um valor de índice de silhueta satisfatório, todos os grupos obtidos possuem mais de 1 dado, estando de acordo com a proposta do trabalho. Os gráficos que mostram os agrupamentos gerados para cada um dos testes podem ser vistos na Figura 3, onde cada ponto vermelho representa o centróide de um grupo.

Figura 3. Gráficos dos agrupamentos para K = 5, 20, 35, 50 e 65

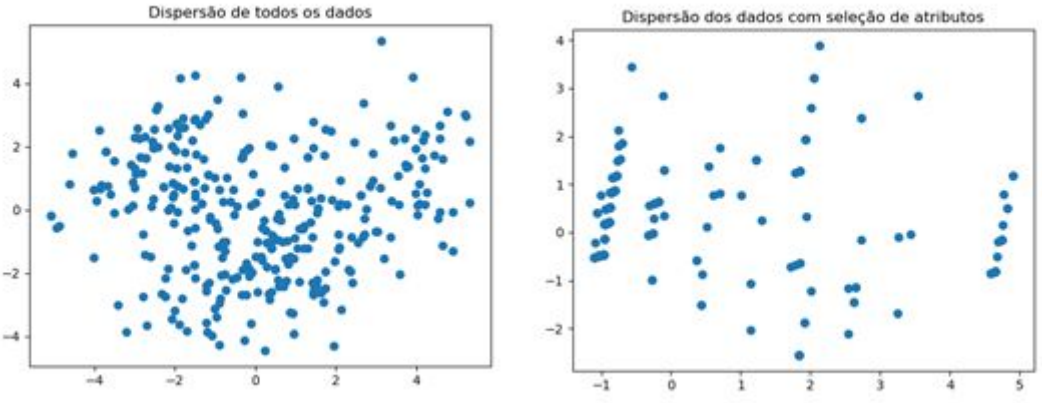


5. Conclusão

Pode-se observar que, utilizando-se todos os métodos para melhorar o desempenho do algoritmo, os resultados obtidos foram satisfatórios. De acordo com o Índice silhueta, obteve-se um agrupamento consistente, com um índice médio de 0.7858. É importante ressaltar que o projeto envolve a criação de grupos, portanto grupos com apenas 1 pessoa, por vezes formados pelo algoritmo, não correspondem com o objetivo do trabalho.

Para que fosse possível alcançar esses resultados, fez-se necessária uma seleção dos atributos mais adequados para a realização do agrupamento, visto que, com 19 atributos por dado o algoritmo poderia não obter um desempenho tão eficiente, considerando que, conforme as 333 respostas obtidas através do questionário, os dados estariam muitos próximos entre si, o que aumenta consideravelmente a possibilidade de criação de grupos fracos. Isso pode ser visto na Figura 4, onde é mostrada a dispersão dos dados considerando todos os atributos e a dispersão dos dados considerando apenas os atributos selecionados.

Figura 4. Gráficos de dispersão dos dados



Analisando os gráficos, pode-se concluir que o gráfico da esquerda, que apresenta a dispersão de todos os dados, mostra que os dados estão bem próximos e é complicado determinar grupos bem definidos. Em contrapartida, o gráfico que apresenta a dispersão dos dados com a seleção de atributos, apresenta dados um pouco mais separados e é mais fácil definir grupos, justificando um resultado satisfatório no agrupamento com a seleção dos atributos mais relevantes.

Além disso, uma redução de dimensionalidade dos atributos (PCA) foi utilizada para que o algoritmo pudesse analisar mais precisamente os atributos de cada dado em relação à execução do algoritmo K-means sem o uso deste procedimento.

Considerando a diminuição da quantidade de atributos para chegar-se a um resultado satisfatório, a construção de um novo questionário, com menor quantidade de perguntas, visando criar uma base de dados com menos atributos a serem analisados pelo algoritmo e, possivelmente, a diminuição de opções de resposta (de 1 a 3, por exemplo), é um aprofundamento necessário para este trabalho, permitindo uma análise mais profunda dos impactos da quantidade de atributos e quantidade de respostas diferentes obtidas nos agrupamentos realizados pelo algoritmo.

Pretende-se também adaptar este projeto no futuro, criando um sistema que possa sugerir rotas de acordo com o perfil de um grupo de pedestres. Para isso, será necessária a implementação de ferramentas que possam identificar pessoas de uma mesma região, além de detectar as características de viagem, como por exemplo, rotas com maior quantidade de pontos turísticos, rotas com maior fluxo de pessoas, rotas mais rápidas etc.

6. Referências

AMIRGHOLY, Mahyar; GOLSHANI, Nima; SCHNEIDER, Craig; GONZALES, Eric J.; GAO, H. Oliver. An advanced traveler navigation system adapted to route choice preferences of the individual users. **International Journal Of Transportation Science And Technology**, [s.l.], v. 6, n. 4, p.240-254, dez. 2017.

BUENO, Andre Luis Cavalcanti. **Relaxamento Adaptativo da Sincronização Através do Uso de Métodos de Aprendizagem Supervisionada**. 2018. Tese de Doutorado. PUC-Rio.

DE CASTRO, ARMANDO ANTONIO MONTEIRO; DO PRADO, PEDRO PAULO LEITE. Algoritmos para reconhecimento de padrões. **Revista Ciências Exatas**, v. 8, n. 2002, 2002

DE SOUSA GUEDES, Inês Maria Ermida. **Medo do Crime: Emergência, Reações Emocionais e Discursos. Contributos para a Utilização de Multi-Metodologias**. 2016. Tese de Doutorado. Universidade do Porto.

FACULTY OF SECURITY. **International Yearbook**. Skopje, 2013.

FADEL, Augusto César; SEMAAN, G.; BRITO, J. Um estudo da aplicação de técnicas de combinação de agrupamentos. **Anais do XVII Simpósio de Pesquisa Operacional e Logística da Marinha**, v. 1, n. 1, p. 188-200, 2014.

FUNDAÇÃO GETÚLIO VARGAS SOCIAL. Centro de Políticas Sociais. **Percepções da Crise**. São Paulo, 2017.

GESSINGER, Joice; HAMMES, Laerson; COLLING, Juliane. INTELIGÊNCIA ARTIFICIAL ARTIFICIAL INTELLIGENCE, 2019.

GOTELLI, Nicholas J.; ELLISON, Aaron M. **Princípios de estatística em ecologia**. Artmed Editora, 2016.

KAPLAN, Andreas; HAENLEIN, Michael. Siri, Siri, in my hand: Who's the fairest in the land? On the interpretations, illustrations, and implications of artificial intelligence. **Business Horizons**, v. 62, n. 1, p. 15-25, 2019.

MATTE, Marcelo Kuchar. **Impacto do uso da desigualdade triangular para acelerar o algoritmo k-means**. 2020. 168 f. Dissertação (Mestrado) - Curso de Ciência da Computação, Unifaccamp, Campo Limpo Paulista, 2020.

NALDI, Murilo Coelho. **Técnicas de combinação para agrupamento centralizado e distribuído de dados**. 2011. Tese de Doutorado. Universidade de São Paulo.

PIMENTEL, Edson P.; DE FRANÇA, Vilma F.; OMAR, Nizam. A identificação de grupos de aprendizes no ensino presencial utilizando técnicas de clusterização. In: **Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)**. 2003. p. 495-504.

QUINTEIRO, José António Teixeira. **Segmentação de indivíduos no facebook que gostam de música: abordagem exploratória, recorrendo à comparação entre dois algoritmos, k-means e fuzzy c-means**. 2011. Tese de Doutorado. Instituto Superior de Economia e Gestão.

SILVEIRA, Nereida Salette Paulo da; HANASHIRO, Darcy Mitiko Mori. Similaridade e dissimilaridade entre superiores e subordinados e suas as implicações para a

qualidade da relação diádica. **Rev. adm. contemp.**, Curitiba, v. 13, n. 1, p. 117-135, Mar. 2009

TUAN, Yi-fu. **Paisagens do Medo**. São Paulo: Unesp, 2005. 374 p.

VARELLA, Carlos Alberto Alves. Análise de componentes principais. **Seropédica: Universidade Federal Rural do Rio de Janeiro**, 2008.

ZHU, Changsheng; IDEMUDIA, Christian Uwa; FENG, Wenfang. Improved logistic regression model for diabetes prediction by integrating PCA and K-means techniques. **Informatics in Medicine Unlocked**, v. 17, p. 100179, 2019.