Agrupando Jovens para a Prática Segura de Esportes Coletivos na Escola Preparatória de Cadetes do Exército (EsPCex)

Trabalho 1 (ME921) - Malcolm dos Reis - 187642

Introdução

Todo jovem brasileiro do sexo masculino, no ano em que faz 18 anos, deve se apresentar para a Junta de Serviço Militar mais próxima da sua residência para realizar o alistamento para o Serviço Militar obrigatório, como previsto por Lei na Constituição da República Federativa do Brasil de 1988. Os jovens que não se apresentarem deixam de possuir alguns direitos civis, como obter passaporte, se matricular em uma universidade pública e trabalhar para alguma instituição pública. Portanto, todos os cidadãos que hoje, usufruem de todos os direitos civis no Brasil, em algum momento já realizaram o alistamento obrigatório.

Hoje, o Governo Federal possui um Programa de Dados Abertos que vêm colocando a disposição do público dados relevantes de várias instituições públicas, dentre elas, o Exército Brasileiro.

Com isso, foi possível obter dados sobre as caracteristicas dos jovens que realizaram o alistamento obrigatório em todo o Brasil no ano de 2022. Dentre essas informações temos o estado e a cidade de nascimento, peso, altura, tamanho da circunferencia da cabea e o tamanho do calçado desses jovens, se eles foram dispensados do serviço militar obrigarório ou não, entre outras informações sobre o cidadão que se apresentou. Esse banco de dados contém 1020927 observações e pode ser baixado através desse site.

No Brasil, os jovens que residem em cidades que não possuem uma Junta de Serviço Militar com estrutura suficiente para colocar esse programa em prática, geralmente, são logo dispensados desse serviço militar obrigatório. Entretanto, na cidade de Campinas, cidade do interior de São Paulo, tem uma grande área dedicada a atividades militares como batalhões de logística e de infantaria, por exemplo. Dentro dessa área tem a EsPCex (Escola Preparatória de Cadetes do Exército) que, além de realizar o curso de cadetes para os jovens concursados, também faz uma espécie de programa militar com aqueles jovens que irão fazer o serviço militar obrigatório nessa cidade.

Nesse sentido, na EsPCex tem uma extensa área de prática esportiva a qual nela está inclusa uma quadra poliesportiva, sendo possível a prática de vários esportes coletivos como futsal, handbol, basquete e vôlei.

Nesse sentido, sabendo que todos os jovens participantes do programa do serviço militar obrigatório podem, em determinados horários, podem fazer o uso da quadra poliesportiva, seria muito bom separar esses jovens por estrutura corporal para que os times envolvidos na prática naquele momento tenham um tamanho alinhado para evitar que, por exemplo, pessoas muito altas e com peso muito acima joguem um jogo de muito atrito físico, como handebol e futebol, com uma pessoa de altura bem menor com um peso muito inferior.

Portanto, vamos fazer um agrupamento desses jovens para que eles possam praticar os esportes coletivos nas quadras poliesportivas da EsPCex com segurança.

Como dito anteriormente, os dados podem ser baixados através desse site, mas para realizar o serviço obrigatório na EsPCex o jovem tem que residir na cidade de Campinas, então vamos utilizar apenas os dados de quem reside nesta cidade.

O método adotado para ler esse banco de forma mais otimizada foi ler o arquivo (.csv) em *chunks*, que envolve dividir a leitura dos dados em partes menores e já fazer as manipulações necessárias em vez de

carregar todo o conjunto de uma vez. Nesse caso, a cada bloco o banco filtrava apenas as informações as quais o município de residência fosse "Campinas" e o estado de residência fosse "SP" e, depois disso, selecionar apenas as colunas que continham as variáveis de peso e altura de cada jovem. Desse modo, não é necessário baixar o banco de dados, mas sim, apenas colocar o link de acesso a ele dentro do parâmetro *file* da função *read_csv_chunked* usada para fazer a leitura do banco no modo descrito.

Assim, ao final da leitura de todos os blocos tínhamos o peso e a altura de todos os jovens que se apresentaram para o serviço militar obrigatório na cidade de Campinas. Nesse momento, não estamos considerando se o jovem que se apresentou foi dispensado ou não do serviço militar, iremos fazer os agrupamentos supondo que todos que se apresentaram irão fazer parte desse programa e que todos eles irão utilizar as quadras poliesportivas da instituição para a prática de esportes coletivos.

Todo o processo de tratamento dos dados e de clusterização foi feito utilizando a linguagem R de programação e os códigos produzidos estão disponíveis nesse projeto do GitHub

Após esse procedimento, temos 5052 observações com 5576 células vazias. Com isso, para executar a clusterização vamos eliminar as linhas que contém essas informações faltantes, sobrando 2264 observações, e, para esse trabalho de agrupamento, vamos retirar uma amostra de tamanho 100 de forma pseudoaleatória (usando a função *slice_sample* do *dplyr*).

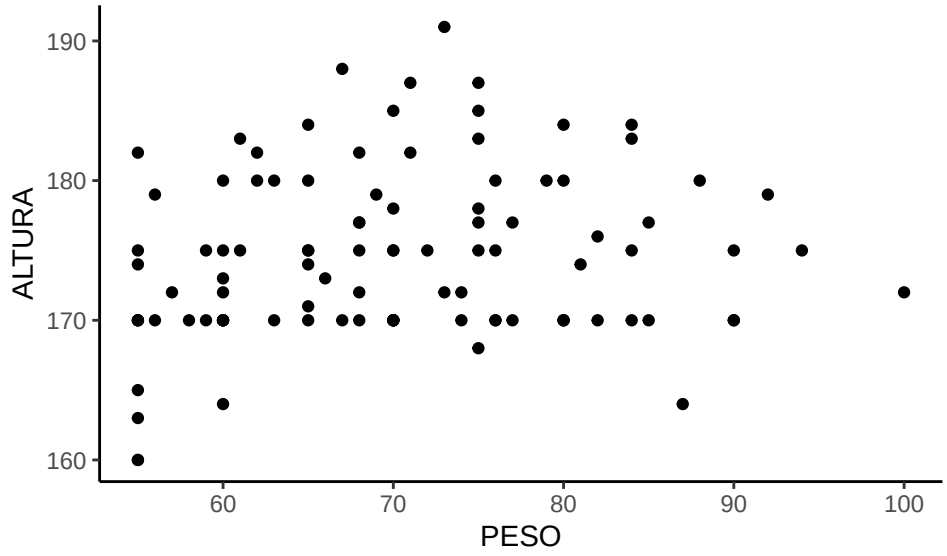

Figure 1: Diagrama de dispersão das variáveis

Materiais e Metodos

Para o agrupamento dessa amostra de $n = 100$ extraída vamos utilizar o agrupamento hierárquico com o método de ligação completa (*complete linkage*) e a quantidade de grupos criados será definido pelo método de Elbow. Esse modo supõe que nossa amostra ainda não possui nenhuma divisão, ou seja, ainda esses dados não foram rotulados.

O *complete linkage* é um método aglomerativo de agrupamento hierárquico, isto é, no passo $h = 0$ temos n grupos e a cada passo, encontramos o melhor par de grupos e fazemos a fusão destes até o último passo em que todas as n observações formam um único grupo. Desse modo, o que difere a ligação completa das outras é a maneira de escolher o melhor par de grupos.

A escolha do melhor par de grupos realizada pelo *complete linkage* considera os pontos mais distantes entre dois grupos de pontos e une os grupos o qual essa distancia é menor. Por exemplo, dados os grupos A, B e

C, temos d_{AB} como a distância dos pontos mais distantes de A e B, d_{AC} como a distância dos pontos mais distantes de A e C e d_{BC} como a distância dos pontos mais distantes de B e C. Se a menor distância dentro dessas consideradas for a d_{AB} então, neste passo, vamos unir o grupo A ao grupo B formando um grupo só e o grupo C se mantém como era no passo anterior.

Lembrando que a distância aqui considerada para essas ligações e a distância euclidiana que, sendo o ponto 1 com as coordenadas (x_1, y_1) e o ponto 2 com as coordenadas (x_2, y_2), é dada por:

$$d(Euc) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

A ligação completa, então, nos gera uma árvore hierárquica do passo 0 até o último em que todos os grupos inciais estão unidos. Para definir em qual passo vamos parar de fazer essas ligações foi utilizado o método de Elbow. Esse método para cada passo K, computa a variância total considerando o número de grupos existente. O gráfico dessa variância total a cada passo tende, em algum momento, apresentar um ponto de inflexão, e é nesse ponto que temos o número de grupos ideal para os dados e para o método escolhido.

Além disso, com esse método de Elbow podemos obter o número de grupos ideal para nossos dados antes mesmo de ver o dendograma (representação gráfica da hierarquia formada pelo agrupamento por ligação completa)

Resultados

Começando pelo método de Elbow, a partir do gráfico de pontos (Figura 2) que mostra a variância considerando o número de grupos existentes, é possível observar uma leve inflexão quando temos K = 5 grupos, como indicado pela linha vermelha. Essa leve inflexão foi considerada suficiente para a determinação do número de clusteres que iremos trabalhar.

Com o número de grupos ideal já obtido, conseguimos o h, medida do corte feito no dendograma da hierarquia gerada pelo método *complete linkage* em cima desses dados, que está indicado pela linha vermelha no gráfico do Dendograma do cluster da Figura 3.

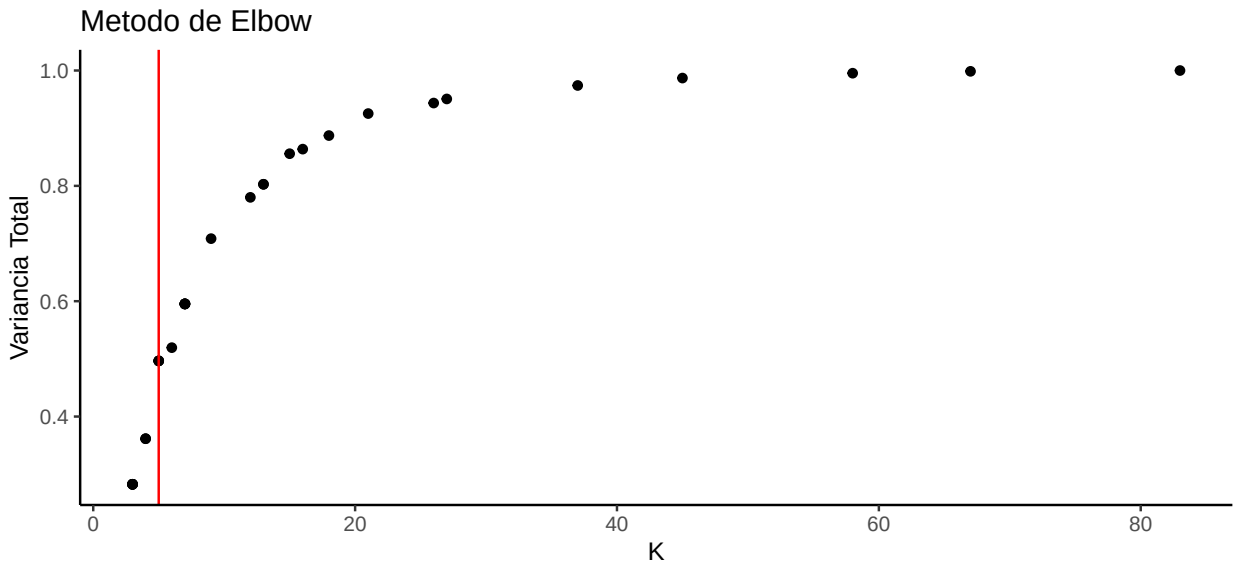

Figure 2: Método de Elbow

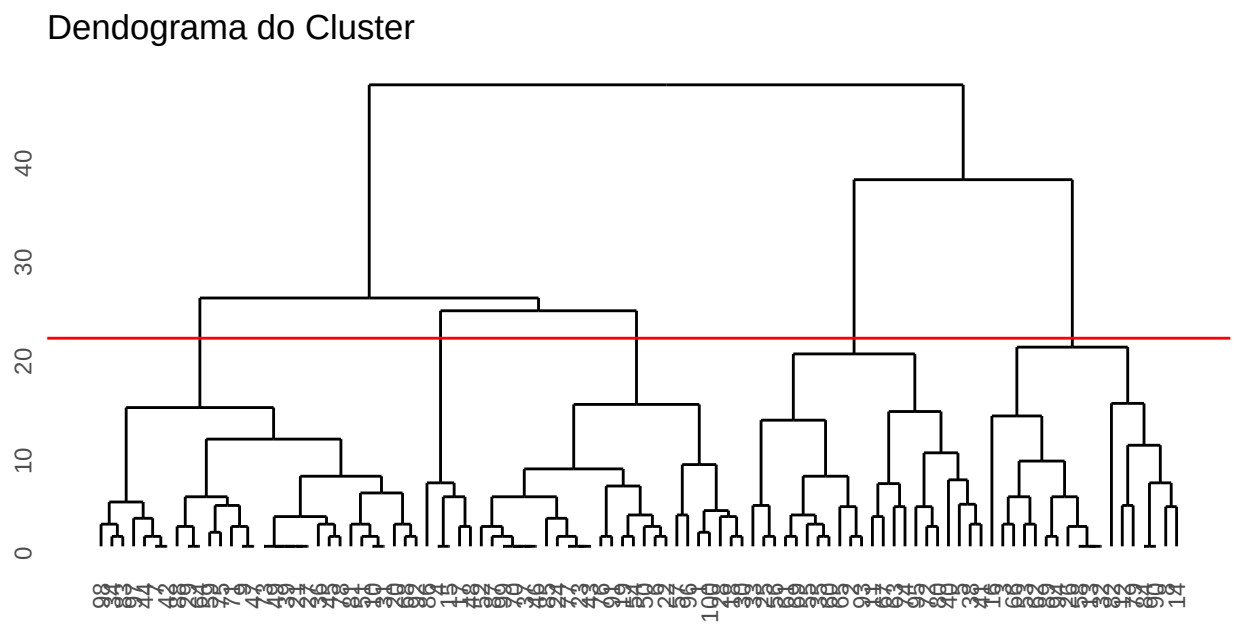

Figure 3: Dendograma do cluster (agrupamento feito por complete linkage)

A partir dessas considerações, temos cada cluster bem definido, como é mostrado no Diagrama de pontos que coloca os pontos em seus respsctivos grupos da Figura 4.

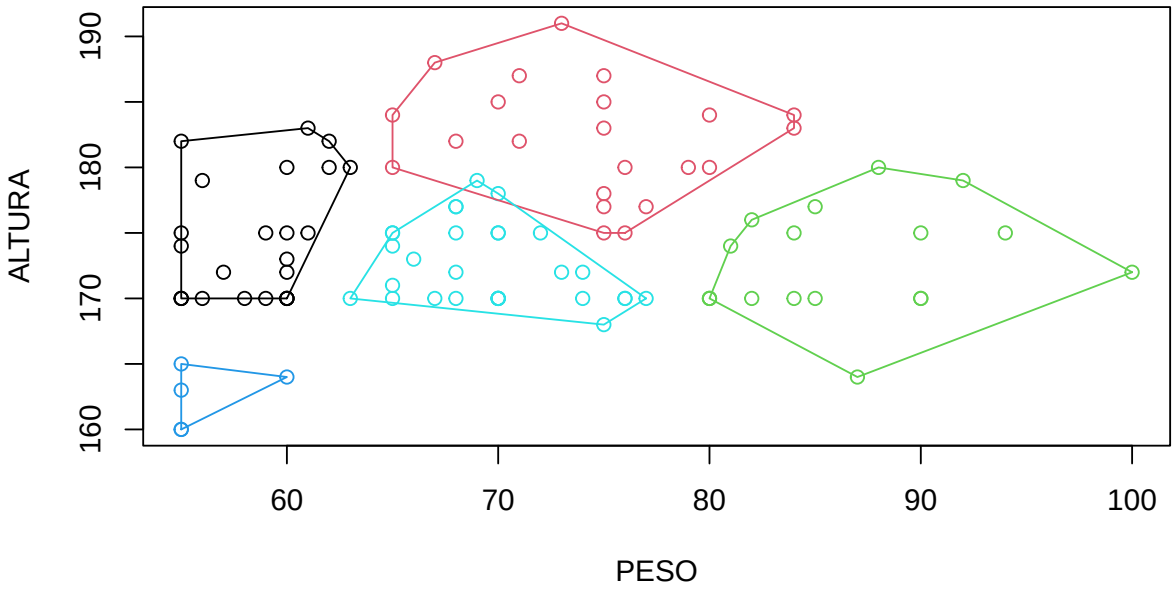

Figure 4: Diagrama de pontos apresentando os grupos obtidos por complete linkage

Discussão

Portanto, através do agrupamento hierárquico utilizando o método de ligação completa, conseguimos dividir os jovens que ingressaram no programa de serviço militar obrigatório em Campinas, fazer com que nas práticas de esportes coletivos nas quadras poliesportivas da EsPCex sejam mais seguras, diminuindo a disparidade de porte físico entre as equipes e evitando assim, possíveis lesões vindas do atrito entre pessoas com portes físicos muito diferentes.

Sabendo que estamos considerando o peso a altura dos jovens, uma outra alternativa para realizar essa separação é dividí-los pelo IMC (Índice de massa corporal). Entretanto, a clusterização é preferível porque ela considera os dados observados, diferente da separação feita pelo IMC onde a faixa do índice já é rotulado antes mesmo de saber sobre as observações. Caso a proporção de pessoas abaixo do peso (seguindo a tabela de IMC) seja grande, a separação por esse índice coloca todas essas pessoas no mesmo grupo, enquanto utilizando a metodologia apresentada isso não aconteceria.

Referências

1. Ludwig, Guilherme. Slides apresentados em aula na Universidade Estadual de Campinas em 2024.
2. B. S. Everitt, S. Landau, M. Leese e D. Stahl. Cluster Analysis, 5th edition. John Wiley & Sons, 2011.
3. Bouveyron C, Celeux G, Murphy TB, Raftery AE. Model-Based Clustering and Classification for Data Science: With Applications in R. Cambridge University Press; 2019.