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 cabeça 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 200 de forma pseudoaleatória (usando a função *slice_sample* do *dplyr*).

	PESO	ALTURA
Mínimo	55.00	160.0
1o Quartil	65.00	170.0
Mediana	73.00	175.0
Média	73.44	175.3
3o Quartil	80.00	180.0
Máximo	146.00	198.0

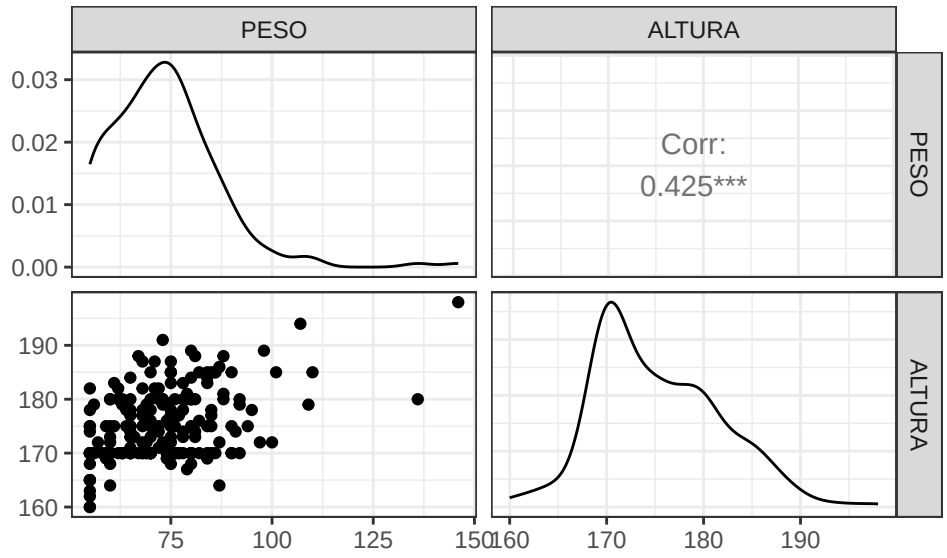

Figure 1: Gráfico de pares das variáveis

Pelo diagrama de dispersão, pode-se desconfiar de uma possível relação linear entre as variávies Peso e Altura. E, com isso, temos um coeficiente de correlação de 0.425. A distribuição do Peso aparenta ser uma distribuição assimétrica para a esquerda enquanto a Altura aparenta ser uma mistura de distribuições normais.

Materiais e Metodos

Para o agrupamento dessa amostra de $n = 200$ extraída vamos utilizar a clusterização baseada em modelos que assume que os dados são gerados por uma mistura de distribuições probabilísticas, onde cada cluster corresponde a uma dessas distribuições. O objetivo é estimar os parâmetros das distribuições para identificar os clusters nos dados. Algoritmos como o *Expectation-Maximization* (EM) são usados para ajustar *Gaussian Mixture Models* (GMMs) e determinar esses parâmetros. Essa abordagem permite uma representação flexível dos clusters e equilibra a complexidade do modelo com a qualidade do ajuste. Além disso, critérios como *Bayesian Information Criterion* (BIC) e *Akaike Information Criterion* (AIC) ajudam a avaliar a qualidade da clusterização, nesse projeto, será usado o BIC como critério.

Finite Mixture Models (FMMs) assumem que os dados são gerados por uma mistura de um número limitado de distribuições probabilísticas diferentes, cada uma representando um cluster. Cada componente da mistura tem parâmetros específicos, como média e variância, que descrevem seu comportamento. O objetivo é estimar tanto a proporção de cada componente quanto seus parâmetros, geralmente usando o algoritmo *Expectation-Maximization* (EM). Isso ajuda a identificar e modelar a estrutura subjacente dos dados em clusters distintos.

O algoritmo EM é uma técnica iterativa usada para encontrar os melhores parâmetros em modelos de mistura finita. Ele começa com estimativas iniciais dos parâmetros e, em seguida, alterna entre duas etapas: na primeira, chamada Expectation, calculamos as probabilidades de cada ponto de dados pertencer a cada cluster; na segunda, chamada *Maximization*, atualizamos os parâmetros dos clusters com base nessas probabilidades. Esse processo é repetido até que os parâmetros se estabilizem e não mudem mais.

O algoritmo EM funciona da seguinte forma: no passo E, definimos parâmetros iniciais e calculamos as probabilidades de pertencimento de cada ponto a cada cluster e no passo M, usamos essas probabilidades para atualizar os parâmetros, como médias e variâncias, maximizando a probabilidade dos dados observados.

O BIC é um critério que ajuda a decidir quantos clusters usar em modelos de clusterização. Ele é como um guia que nos diz se devemos escolher um modelo mais simples ou mais complexo, considerando a quantidade de dados e a complexidade do modelo. O objetivo é encontrar um equilíbrio entre o ajuste aos dados e a simplicidade do modelo, assim, escolhemos o número de clusteres em que, dentre os modelos, o BIC seja maior.

Resultados

Como esse projeto busca juntar militares que tenham similaridades físicas, faz mais sentido utilizarmos modelos esféricos, os quais $\Sigma = \sigma^2 I$ tem apenas um parâmetro. São esses os modelos “EII” e “VII”.

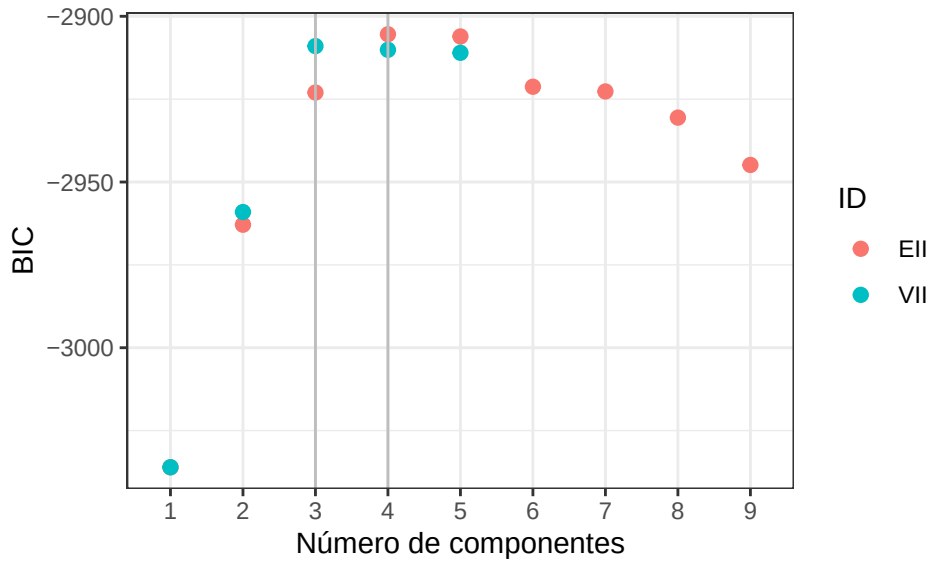

Figure 2: BIC por número de clusters

A primeira linha cinza indica o número de cluster ideal caso escolhêssemos o modelo VII e a segunda linha indica a Temos o BIC por número de clusteres dispostos na Figura 3 e, com isso, foi escolhido o modelo EII com 4 clusteres, sendo estes dispostos na Figura 4.

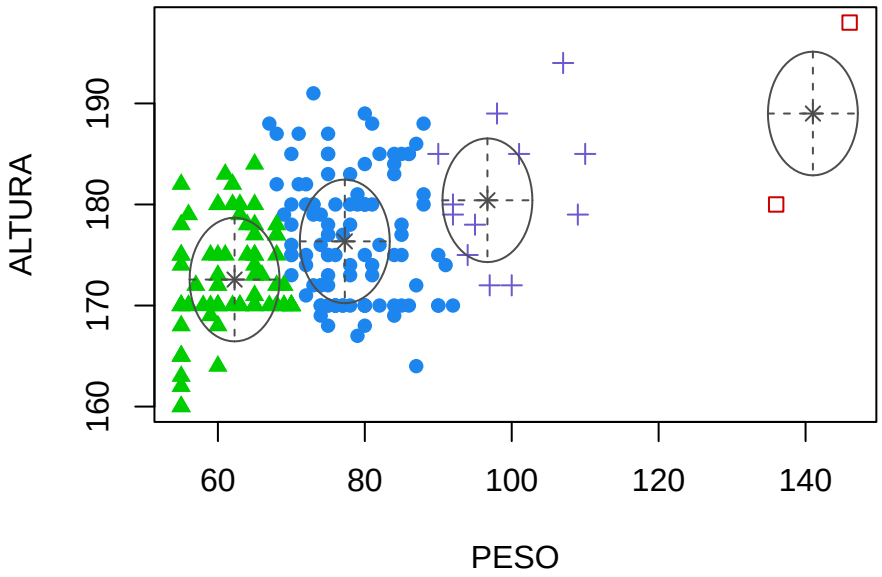

Figure 3: Dispersão dos dados já agrupados

Discussão

Portanto, através da clusterização baseada em modelos, utilizando o algoritmo EM e o critério Bayesiano de informação, conseguimos dividir os jovens que ingressaram no programa de serviço militar obrigatório em Campinas em 4 grupos, 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. Projeto do Github de Malcolm dos Reis sobre clusterização no serviço militar.
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.