Universidade Federal de São Carlos
Centro de Ciências Exatas e de tecnologia
Departamento de Estatística

Análise do Desempenho de Jogadoras de Handebol do Campeonato Mundial Feminino Juvenil

Júlia Beltramini

Trabalho de Conclusão de Curso

Universidade Federal de São Carlos
Centro de Ciências Exatas e de Tecnologia
Departamento de Estatística

Análise do Desempenho de Jogadoras de Handebol do Campeonato Mundial Feminino Juvenil
Trabalho de Conclusão de Curso

Júlia Beltramini
Orientadora: Maria Sílvia de Assis Moura

Trabalho de Conclusão de Curso a ser apresentado como parte dos requisitos para obtenção do título de Bacharel em Estatística.

São Carlos
27 de Novembro de 2021

Júlia Beltramini

Análise do Desempenho de Jogadoras de Handebol do Campeonato Mundial Feminino Juvenil

Trabalho de Conclusão de Curso

Este exemplar corresponde à redação final do trabalho de conclusão de curso devidamente corrigido e defendido por Júlia Beltramini e aprovado pela banca examinadora.

São Carlos, 26 de novembro de 2021.

Banca Examinadora

- Dra. Maria Silvia de Assis Moura
- Dr. Pedro Pedro Ferreira Filho
- Dra. Daiane Aparecida Zuanetti

Agradecimentos

Agradeço a minha família, meus pais e minhas irmãs, por me acompanharem e auxiliarem durante toda essa jornada.

A minha orientadora, professora Maria Silvia, por toda a paciência, sugestões e auxílio desde a elaboração do projeto, execução e finalização. Agradeço a banca avaliadora por aceitarem participar e suas contribuições para a melhora desse trabalho.

Agradeço aos meus amigos e amigas que trilharam essa jornada junto comigo, que proporcionaram momentos inesquecível, de muita descontração e parceria. Em especial a Graze, ao Félix e ao Ed, sem vocês essa jornada não seria tão especial quanto foi.

Resumo

A análise de desempenho de atletas vem se tornando cada vez mais relevante para diversas modalidades, no handebol não é diferente. Para realizar a classificação de dados a análise de agrupamentos é uma ótima opção. Serão utilizadas nesse trabalho, algumas técnicas de agrupamento, como métodos hierárquicos e métodos não hierárquicos, com o objetivo de agrupar e analisar o desempenho de jogadoras de handebol do juvenil que participaram do campeonato mundial. Entre os métodos supracitados, os método de Ward e o algoritmo de K-médias obtiveram resultados muito semelhantes e condizente com o nosso objetivo, quando utilizamos as cinco primeiras dimensões obtidas pela análise de componentes principais. Por fim, os quatro grupos obtidos foram caracterizados observando esses cinco primeiros componentes principais.

Palavras-chave: *Análise de agrupamento, Desempenho, Handebol feminino.*

Sumário

Lista de Tabelas

Lista de Figuras

Capítulo 1

Introdução

O Handebol é um esporte coletivo, regido pela IHF (International Handball Federation), a qual é responsável pela definição das normas, do tempo de jogo, tipo de bola entre outras coisas. A IHF foi fundada em 1946 na Dinamarca (Greco e Romero, 2011). O primeiro campeonato mundial masculino de handebol foi realizado em 1938, contudo apenas em 1957 foi disputado o primeiro campeonato mundial feminino (Menezes, 2011).

A modalidade esportiva Handebol tem cada equipe composta por seis jogadores de linha e um goleiro, esses são os jogadores titulares. Também integram a equipe sete jogadores suplentes, são os jogadores que ficam na reserva esperando uma substituição durante a partida. A modalidade é praticada em uma quadra com dimensões de 20 metros de largura e 40 metros de comprimento, entre duas equipes que tem como objetivo a realização de gols e impedir que a equipe oponente consiga marcar gols (Menezes, 2011). As possíveis posições que um jogador pode assumir são: armador central, goleiro, meias, pivô e pontas.

As posições que os atletas podem atuar tem as seguintes funções: o armador central é responsável por organizar as jogadas de ataque, o goleiro realiza a defesa do gol (impossibilitando que a equipe adversária faça gol), os meias são responsáveis por estabelecer um ritmo mais equilibrado e defensivo para a equipe, o pivô tem como uma de suas funções abrir espaço na defesa adversária para possibilitar infiltrações e arremessos dos jogadores de seu time e os jogadores de ponta são responsáveis por se posicionarem rapidamente para acelerar um contra-ataque.

Em esportes de alto nível a utilização de análise de desempenho está se tornando relevante (Donatelli, 2017). No Handebol é muito utilizado a análise de jogo, observando jogada a jogada. Uma análise mais ampla, considerando todos os jogos de um cam-

peonato, por exemplo, pode ser utilizada como uma forma de mostrar jogadores que se destacam durante uma competição, possibilitando recompensar esses atletas com possíveis contratações ou patrocínios, já que essas análises mostrarão os melhores jogadores. Portanto, uma análise que classifique esses jogadores permite valorizar os melhores atletas, assim favorecendo a categoria profissional e proporciona o crescimento da modalidade, logo ajudando na visualização do Handebol.

Para encontrar os jogadores que se destacam positiva ou negativamente no Handebol, podemos utilizar uma análise de desempenho dos atletas, levando em consideração suas características, por exemplo, número de conversão de arremessos em gols, número de punições de dois minutos, entre outros aspectos, etc. Uma combinação dessas características dificilmente é realizada para obter uma classificação dos melhores atletas.

Para comparar o desempenho desses atletas de handebol, em relação a suas características, será analisado o desempenho de jogadoras no Campeonato Mundial Feminino Juvenil no ano de 2018.

Os dados utilizados nesse trabalho correspondem a informações das atletas dos 24 países que participaram do Campeonato Mundial Feminino Juvenil (Sub-18) no ano de 2018, os quais foram diponibilizados pela IHF [1]. O banco de dados será criado a partir da compilação de tabelas com informação pessoais, como a posição que joga, a altura, o peso e com informações das quantidades de arremessos, gols, quantidade de assistências, quantidade de cartões recebidos, entre outras. Aqui serão utilizadas as características supracitadas para classificar o desempenho das jogadoras aplicando algumas técnicas de agrupamento.

Na literatura existem diversas técnicas de agrupamento, ou *clustering*, em que as técnicas de agrupamento hieráquico e o método não-hierárquicos são os algoritimos que tem um maior destaque (Donatelli, 2017). Neste trabalho, essas técnicas serão usadas, juntamente, com a técnica de componentes principais, com o objetivo de descrever a formação dos índices criados pela análise de componentes principais (ACP). A criação dos índices tem a intenção de reduzir de 35 variáveis para um número menor de índices, ou seja utilizando a informação das 35 variáveis de maneira mais concisa. Os índices serão utilizados como entrada para o método de agrupamento, então os agrupamentos das jogadoras serão realizados pelos valores dos índices criados pela ACP, para que seja

[1] https://archive.ihf.info/en-us/ihfcompetitions/worldchampionships/womensyouthworldchampionships/2018womensyouth(u18)worldchampionship/teaminfo.aspx

possível analisar o desempenho das jogadoras que têm um comportamento semelhante.

Além disso, queremos observar o grau de concordância entre os métodos de agrupamento para, no final, se for possível, criar um indicador composto que seja capaz de ranquear as jogadoras e os grupos gerados de forma a identificar a correspondência que há com o desempenho das jogadoras durante o campeonato Mundial Feminino Juvenil de 2018, utilizando a técnica de componentes principais (Commission *et al.*, 2008).

Com base no que foi apresentado, propomos este trabalho com o objetivo de comparar o desempenho de jogadoras de handebol do campeonato de 2018 do "Women's Youth (U18) World Championship", tendo em vista classificar as melhores jogadoras desse campeonato. Para possibilitar essa comparação, será criado um banco de dados com informações de todas as jogadoras dos 24 países que participaram do campeonato, contendo informações pessoais e de participação nos jogos do respectivo campeonato.

Então, o trabalho é organizado da seguinte forma: no Capítulo 2 são apresentadas técnicas de agrupamentos, hierárquicos e não-hierárquico, e também informações sucintas sobre análise de componentes principais; no Capítulo 3 é descrito informações sobre o banco de dados; no Capítulo 4 mostramos os resultados práticos da aplicação desses métodos e, por fim, no Capítulo 5 apresentamos algumas considerações finais sobre o trabalho de graduação.

Capítulo 2

Metodologia

Esse capítulo apresenta as metodologias estatísticas aplicadas na análise do desempenho de jogadoras de handebol do campeonato de 2018 do "Women's Youth (U18) World Championship", tendo em vista classificar as melhores jogadoras desse campeonato. Para agrupá-las e classificá-las serão utilizadas técnicas de agrupamentos. A Seção 2.1, trata de medidas de similaridade e medidas de distância. Nas Seções 2.1.1 e 2.1.2 são apresentados métodos de agrupamentos hierárquicos e K-Médias. A Seção 2.2, exibe a técnica de componentes principais.

2.1 Análise de Agrupamento

Com a intenção de identificar indivíduos semelhantes (no caso desse trabalho, jogadoras de handebol), objetos, produtos, entre outros, em grupos homogêneos e com características parecidas, a análise de agrupamento é a técnica mais utilizada (Hair *et al.*, 2009). Essa técnica pode ser utilizada para identificar, em banco de dados, padrões de comportamento e também para agrupar variáveis, sendo ela mais usada para agrupar objetos (Barroso e Artes, 2003).

Existem vários procedimentos de agrupamento, ou *clustering*, os dois maiores grupos são: método hierárquico e não-hierárquico, e grande parte dos algoritmos de agrupamento podem ser classificados nesses dois grandes grupos. Outros exemplos de procedimentos de agrupamento são: *fuzzy*, grafo-teórico, redes neurais e modelos evolucionários, entre outros. Contudo, os métodos mais populares são os hierárquicos e não-hierárquicos. Além disso, os métodos não-hierárquicos também podem ser chamados de métodos de partição.

A análise de agrupamento tem como objetivo maximizar a homogeneidade dos in-

divíduos dentro do mesmo grupo e, simultaneamente, maximizar a heterogeneidade dos indivíduos entre grupos diferentes. Portanto, jogadoras de grupos distintos são mais diferentes do que jogadoras que estão no mesmo grupo.

A análise de agrupamentos não é uma técnica de inferência e detém como objetivo quantificar as características estruturais de um conjunto de dados. Além disso, as únicas questões que devem ser observadas nas variáveis e indivíduos que sejam utilizados nas análises de agrupamento são a representatividade da amostra e a multicolinearidade entre variáveis.

Para efetuar os cálculos necessários para realizar os agrupamentos definimos que os dados com n indivíduos e p variáveis formam uma matriz de dados $X_{n\times p}$, em que x_i e x_j, $i, j = 1, ..., n$ e $i \neq j$, são linhas diferentes da matriz $X_{n\times p}$, portanto x_i e x_j são vetores com dimensão p. Então, as colunas da matriz correspondem as variáveis e as linhas correspondem aos indivíduos.

A maioria dos métodos de agrupamento necessita de uma medida de similaridade, para quantificar o quanto cada indivíduo é parecido ou não com os demais. Essas medidas de similaridade podem ser medidas de distância, medidas de correlação ou medidas de associação (Hair *et al.*, 2009). Como dependem do tipo de dados, não usaremos a medida de associação nesse trabalho, já que é para dados qualitativos e os dados do trabalho são métricos, ou seja, quantitativos.

Segundo Barroso e Artes (2003) e Hair *et al.* (2009), medidas de distância são as mais usadas em dados quantitativos, porque representam a proximidade de observações. O segundo autor considera que esses tipos de medidas originam perfis de centroide de grupos sobre as variáveis de agrupamento mais úteis do que os resultados de medidas de correlação.

As medidas de similaridade são representadas por uma matriz de similaridade $D_{n\times n}$, ou seja, é uma matriz que pode mostrar quão parecidas as observações são ou quão distantes estão as observações entre si,

$$D = \begin{bmatrix} d_{11} & d_{12} & \dots & \dots & d_{1n} \\ d_{21} & d_{22} & \dots & \dots & d_{2n} \\ \vdots & \vdots & \ddots & & \vdots \\ \vdots & \vdots & & \ddots & \vdots \\ d_{n1} & d_{n2} & \dots & \dots & d_{nn} \end{bmatrix}. \tag{2.1}$$

A matriz é simétrica e pode conter tanto medidas de similaridade como medidas de dissimilaridade. É uma medida de dissimilaridade, por exemplo, se d_{ij} for uma medida de distância, porém com o subterfúgio de que com uma relação inversa é possível converter a distância em uma medida de similaridade, logo $d_{ij} = \max_{i,j}\{d_{ij}\} - d_{ij}$ é a medida de similaridade (Härdle e Simar, 2015), considerando que, quanto maior a distância menor a semelhança entre os indivíduos.

Na literatura são apresentados alguns tipos de medidas de distância, sendo elas:

- **Euclidiana:** é a medida comumente mais usada, porém tende a formar grupos hiperesféricos, descrita como

$$d_{ij} = \left(\sum_{\ell=1}^{p} |x_{i\ell} - x_{j\ell}|^2 \right)^{1/2};$$

- **City-block:** se as variáveis forem altamente correlacionadas pode conduzir a agrupamentos espúrios, calculada da seguinte forma

$$d_{ij} = \sum_{\ell=1}^{p} |x_{i\ell} - x_{j\ell}|;$$

- **Minkowsky:** essa medida tem dois casos particulares, sendo eles a distância Euclidiana, quando $m = 2$, e a distância de Manhattan (city-block), quando $m = 1$. É definida como

$$d_{ij} = \left(\sum_{\ell=1}^{p} |x_{i\ell} - x_{j\ell}|^m \right)^{1/m};$$

- **Mahalanobis:** é a medida que realiza uma padronização em relação à estrutura de covariância entre as variáveis, definida por

$$d_{ij} = \big(x_i - x_j\big)^T S^{-1} \big(x_i - x_j\big),$$

na qual, S é a matriz de variância e covariância, x_i e x_j são os vetores das variáveis i e j, em que $i \neq j$ (Johnson *et al.*, 2002) , (Xu e Wunsch, 2005).

Para escolher a melhor medida de distância deve-se observar que: utilizar diferentes medidas e mudanças na escala podem resultar em diferentes agrupamentos, então é indicado comparar as diversas medidas e comparar também com resultados prévios ou padrões

teóricos; se as variáveis forem correlacionadas ou multicolineares, a distância Mahalanobis é a mais adequada já que pondera de acordo com a variância todas as variáveis.

2.1.1 Métodos de Agrupamento Hierárquico

Os métodos de agrupamento hierárquico podem ser divididos em dois grupos, sendo eles: aglomerativo e divisor, nos quais seus procedimentos são contrários um do outro. O aglomerativo inicia com os grupos contendo apenas um indivíduo e durante o procedimento vai agrupando os mesmos. Enquanto o divisor, como o próprio nome diz, vai dividir os grupos com o passar do procedimento. O método se inicia com um agrupamento contendo todos os indivíduos da amostra, e durante o procedimento esse único grupo vai ser dividido em *clusters* menores, essa diferença pode ser vista na Figura 2.1.

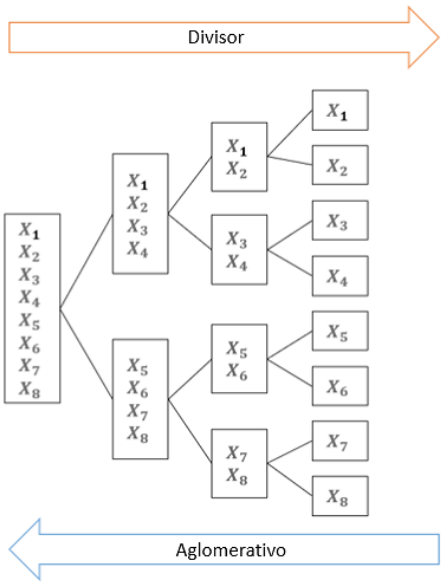

Figura 2.1: Ilustração dos métodos de agrupamento hierárquico.
Fonte: Produzida pela autora.

Ambos os métodos podem ser representados por um dendrograma, também conhecido como Diagrama de Árvore, em que o eixo das abscissas representa os agrupamentos, os nodos representam os pontos onde os agrupamentos se fundem ao longo do eixo das ordenadas que mostra a distância (ou dissimilaridade) que essa fusão sucede. Com esse estilo de gráfico, ilustrado na Figura 2.2, se busca identificar a ocorrência de grandes saltos, já que esses representam a junção de indivíduos heterogêneos.

O algoritmo do método aglomerativo é dado pelos seguintes procedimentos:

1. Alocar todas as observações em grupos individuais;

2. Usar a matriz de similaridade e combinar os dois grupos mais parecidos, ou seja, com menor valor de distância, em um novo grupo;

3. Calcular a nova matriz D;

4. Repetir o passo 2 e 3 um total de $n-1$ vezes, até que seja formado apenas um grupo com todas as observações.

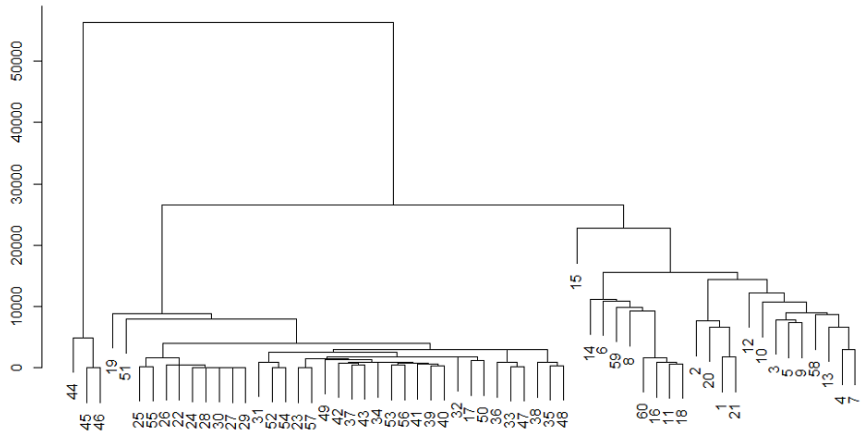

Figura 2.2: Exemplo da ordem de agrupamento através do dendrograma.
Fonte: Produzida pela autora.

Para calcular a nova matriz de similaridade entre a união dos grupos A e B com relação ao grupo C temos os métodos algomerativos (Härdle e Simar, 2015), os mais utilizados são:

- **Método de ligação simples:** a similaridade é encontrada pela menor distância entre as observações de um grupo com as do outro, ou seja,

$$d(C, A+B) = min\{d(C, A), d(C, B)\}.$$

  Um problema desse algoritmo é que, para agrupamentos mal delineados, esse método pode gerar longas e sinuosas cadeias, divergindo do objetivo de agrupamentos mais compactos;

- **Método de ligação completa:** a similaridade é definida pela maior distância entre as observações de um grupo com as do outro, ou seja,

$$d(C, A+B) = max\{d(C, A), d(C, B)\}.$$

Esse algoritmo elimina o problema de cadeias que o método de ligação simples tem;

- **Método de ligação média:** a similaridade é dado pela média de todas as observações em um grupo com todas as observações do outro, ou seja,

$$d(C, A+B) = \frac{n_A}{n_A+n_B}\, d(C,A) + \frac{n_B}{n_A+n_B}\, d(C,B),$$

em que, n_A e n_B são, respectivamente, os números de indivíduos nos grupos A e B. Esse método é pouco afetado por observações atípicas e resulta em agrupamentos com pouca variação interna;

- **Método centroide:** a similaridade é encontrada pela distância entre os centroides, que corresponde a distâncias entre os grupos, ou seja,

$$d(C, A+B) = \frac{n_A}{n_A+n_B}\, d(C,A) + \frac{n_B}{n_A+n_B}\, d(C,B) - \frac{n_C}{n_C+n_A+n_B}\, d(A,B),$$

em que, n_A, n_B e n_C são, respectivamente, os números de indivíduos nos grupos A, B e C. Depois de todo novo agrupamento é calculado um novo centroide, esse método é muito usado na física e não é fortemente afetado por valores atípicos;

- **Método de Ward:** a similaridade é calculada pela soma dos quadrados dentro dos grupos feita sobre todas as variáveis. É uma técnica muito influenciável por valores atípicos e os grupos são feitos com o intuito de minimizar o aumento na soma total de quadrados em todas as variáveis em todos os grupos, portanto não reúne indivíduos com menor distância, mas reúne aqueles que resultam em grupos com menor variância dentro dos grupos. Esse é um método indicado para quando se deseja grupos com tamanhos semelhantes.

Não existe um método de seleção padrão e objetivo para a regra de parada e para que o pesquisador defina o número de grupos. Então, para definir quantos grupos devem ser utilizado na análise, foram criados alguns critérios de parada. Contudo, eles tem alguns problemas, já que, na maior parte das vezes são técnicas muito complexas e ocorre naturalmente, durante o processo, um aumento na heterogeneidade com a redução do número de grupos. Portanto, deve-se notar durante o procedimento um aumento significativo da heterogeneidade entre os grupos. Algumas dessas regras de parada:

- **Raiz quadrada do desvio padrão médio (RMSSTD):** a heterogeneidade é medida pela raiz quadrada da variância do novo grupo resultante da união de dois agregados. Se a união de dois grupos gera um grande aumento na RMSSTD, indica que se está unindo dois grupos com características distintas, mostrando que o resultado anterior era uma possível solução final;

- ***pseudo-F*:** calcula a heterogenidade pela seguinte fórmula,

$$pseudo - F = \frac{tr[B/(K-1)]}{tr[W/(n-K)]},$$

em que B é a matriz da soma de quadrados entre os grupos, W é a matriz da soma de quadrados dentro dos grupos, K é o número de grupos e n é o número de indivíduos. Quanto maior o $pseudo - F$ se considera que mais eficiente é a divisão na redução da heterogeneidade dentro dos grupos. E quando se tem um decréscimo no valor do $pseudo - F$, devido a um aumento no número de grupos, mostra que o agrupamento pode não ser vantajoso;

- ***R-Squared* (RS):** mede a heterogeneidade da seguine forma

$$RS = \frac{SQG}{SQT},$$

em que a soma de quadrados total entre os grupos é SQG e SQT é a soma de quadrados total. Quanto maior o valor do RS, lembrando que ele varia entre 0 e 1, indica que os grupos formados são mais heterogêneos, logo o agrupamento anterior é melhor porque diminui a variância e faz os grupos ficarem mais homogêneos.

Os métodos de agrupamento hierárquico tem suas vantagens e desvantagens, algumas dessas vantagens são resultar em soluções simples e possibilitar uma análise mais rápida, porque, como deixa a definição do critério de parada para o pesquisador em uma etapa posterior, resulta em todas as soluções possíveis , ou seja, permite analisar uma vasta gama de soluções, mesmo sendo abrangentes. Algumas de suas desvantagens são a ineficiência em analisar grandes amostras, devido a um grande aumento na exigência de armazenamento dos dados, e por serem considerados míopes (Lattin *et al.*, 2011), já que o indivíduo uma vez agrupado não pode ser realocado.

2.1.2 Métodos de Agrupamento Não-Hierárquico

Os métodos não-hierárquicos ou métodos de partição, exigem um conhecimento prévio do número, K, de grupos que deseja-se formar ao final do procedimento. Eles procuram as melhores configurações do particionamento dos dados nesses K grupos.

MacQueen *et al.* (1967) define K-Médias como sendo um procedimento que é iniciado com K grupos, sendo esses compostos por um único ponto aleatório e, a cada passo do algoritmo, é adicionado um novo ponto no grupo cuja a média seja mais próxima desse novo ponto. Uma nova média do grupo é calculada cada vez que se acrescenta um novo ponto no grupo. Esse procedimento é um dos mais conhecidos e utilizados, quando se refere a métodos de agrupamento.

As maiores diferenças entre os agrupamentos hierárquicos e os não-hierárquicos são que o primeiro não relaciona indivíduos durante a execução do algoritmo e não necessita de um conhecimento prévio da quantidade de grupos, já o segundo realiza realocações dos indivíduos durante a aplicação do algoritmo e necessita da quantidade de grupos que serão formados anteriormente a sua execução (Härdle e Simar, 2015).

A representação gráfica de árvore hierárquica não é realizada para o método de K-Médias. Os pontos médios dos grupos são utilizados como centros de gravidade.

O algoritmo de K-Médias é dado da seguinte forma:

1. Dividir os indivíduos em K grupos iniciais;

2. Realocar cada indivíduo no grupo com centroide mais similar a ele. A cada novo indivíduo alocado calcular novamente o valor do centroide, tanto do grupo que ele foi incluso quanto do grupo que ele era originário;

3. Repetir o passo 2 até que seja finalizado as reatribuições (Johnson *et al.*, 2002).

O metódo K-Médias realiza, como já foi dito, realocações dos indivíduos nos grupos. Portanto, se um indivíduo que já está em um grupo durante o procedimento ficar mais semelhante a outro grupo, ele pode ser realocado para esse grupo ao qual ele é mais semelhante (Hair *et al.*, 2009).

O procedimento K-Médias tem como base a minimização da soma de quadrados da partição, definida por:

$$SQDP = \sum_{\ell=1}^{p} SQD_{(\ell)},$$

em que $SQD_{(\ell)}$ é a soma de quadrado dentro do grupo da variável ℓ. Um agrupamento (partição) é considerado bom quando minimiza o $SQDP$, que é a soma de quadrados da partição quando se considera todas as variáveis simultaneamente.

Os métodos de agrupamento não-hierárquicos tem suas vantagens e desvantagens, uma dessa desvantagens é que não se tem garantia de um agrupamento ótimo, mesmo utilizando uma solução inicial não-aleatória, visto que a escolha dos centroides iniciais influencia muito os resultados. Outra desvantagem é que necessita do conhecimento prévio da quantidade de grupos. Algumas de suas vantagens é poder analisar grandes conjuntos de dados, já que não necessita realizar o cálculo da matriz completa de similaridade. Outra vantagem dos métodos de K-Médias é que a medida de distância escolhida, observações *outlier* e o uso de variáveis que não são significativas ou são inadequadas não impactam tanto os resultados.

Uma maneira para definir a quantidade de grupos é realizar várias soluções com diversos números de grupos e, com o subterfúgio da soma de quadrados da partição, observar as vantagens de um número de grupos maior. Para efetuar essa comparação pode-se utilizar o índice G, como propõe Barroso e Artes (2003), em que:

$$G = \frac{SQDP_{(k)} - SQDP_{(k+1)}}{SQDP_{(k+1)}}.$$

Procura-se o menor valor de K que estabilize o índice G, ou seja, o valor de K cujo o acréscimo de mais um grupo não reduza drasticamente o valor do $SQDP$. Sabendo que $SQDP_{(k)}$ é a soma de quadrados dentro dos grupos.

Também podem ser usadas, para obter a quantidade de grupos, as mesmas regras de paradas dos métodos hierárquicos, explicadas na Seção 2.1.1.

Uma abordagem que consegue utilizar as vantagens de ambos os métodos, hierárquicos e não-hierárquicos, e se tem a intenção de utilizar na continuação desse trabalho de graduação é resultado de uma combinação dos métodos. Inicialmente, para estabelecer o número de grupos, identificar observações atípicas e caracterizar os centros dos grupos, é aplicado uma técnica hierárquica. Em seguida, elimina-se as observações atípicas, utiliza-se os resultados do método hierárquico e seus centroides como valores iniciais do método não-hierárquico. Assim, se obtêm as vantagens de ambos os métodos.

2.2 Componentes Principais

Análise de componentes principais (ACP) é frequentemente utilizada em banco de dados com uma grande quantidade de variáveis. Na maior parte dos estudos a ACP é uma etapa intermediária.

O objetivo da ACP é reduzir a dimensionalidade do problema, mas com a menor perda da quantidade de informações possível. Por intermédio de combinações lineares das variáveis originais, a metodologia tenta explicar a variância e covariância de todo o conjunto de dados. Essa técnica pode ser usada quando se tem a intenção de descartar informações que possam ser consideradas redundantes (Johnson *et al.*, 2002). No âmbito deste trabalho, essa metodologia vai auxiliar na construção de um indicador único, se esse for possível, que possibilitará ranquear as jogadoras em estudo e os grupos obtidos pelas técnicas de agrupamento.

Assim, almeja expressar as relações entre as variáveis da melhor maneira possível, através de combinações lineares ótimas. Então essas combinações lineares não correlacionadas são os componentes principais.

É definida a ACP por meio de três passos:

1. Obter a matriz de variância e covariância das variáveis do banco de dados:

$$\Sigma = \begin{bmatrix} \sigma_1^2 & & & & \\ \sigma_{12}^2 & \sigma_2^2 & & & \\ \vdots & \vdots & \ddots & & \\ \vdots & \vdots & & \ddots & \\ \sigma_{1p}^2 & \sigma_{2p}^2 & \dots & \dots & \sigma_p^2 \end{bmatrix}; \tag{2.2}$$

2. Para evitar o efeito de diferentes escalas, padronizar a matriz de variância e covariância, chegando assim, na matriz de correlação:

$$R = \begin{bmatrix} 1 & & & & \\ \rho_{12} & 1 & & & \\ \vdots & \vdots & \ddots & & \\ \vdots & \vdots & & \ddots & \\ \rho_{1p} & \rho_{2p} & \dots & \rho_{(p-1)1} & 1 \end{bmatrix}, \tag{2.3}$$

na qual:

$$\rho_{\ell m} = \frac{\sigma^2_{\ell m}}{\sqrt{\sigma^2_\ell \sigma^2_m}}, \quad \ell, m = 1, \ldots, p \quad \text{com} \quad \ell \neq m;$$

3. Calcular os autovalores e seus autovetores relacionado à matriz de correlação.

A análise de componentes principais realiza a transformação das p variáveis originais $(X_1, \ldots, X_p)$ em p variáveis $(Y_1, \ldots, Y_p)$, chamadas de componentes principais, que são combinações lineares não correlacionadas das variáveis $X_1, \ldots, X_p$. Sendo Y_1 o componente que mais explica a variabilidade total dos dados, Y_2 é o segundo componente que mais explica variabilidade e assim sucessivamente, até o último componente Y_p.

Os autovalores são ordenados de modo que $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0$. A ℓ-ésima componente principal é definida por:

$$Y_\ell = e_{\ell 1} X_1 + e_{\ell 2} X_2 + \ldots + e_{\ell p} X_p,$$

em que $\ell = 1, \ldots, p$, X é o vetor das p variáveis originais e e_i é o i-ésimo autovetor associado. Com os autovalores obtém-se informações sobre a variância de cada componente e a contribuição de cada um na explicação da variabilidade total dos dados, ou seja, o autovalor corresponde a variância do componente principal. Já para saber a contribuição de cada variável dentro de cada componente principal se observa os valores dos autovetores.

Os componentes gerados são ortogonais entre si, significando que são não correlacionados e reafirmando a intenção de descartar informações redundantes.

Capítulo 3

Campeonato Mundial Feminino Juvenil (Sub-18) no ano de 2018

Este capítulo tem como objetivo contextualizar sobre o jogo de handebol e explicar o conjunto de dados e o tratamento que será realizado nele, para a aplicação do mesmo que será realizada no Capítulo 4, com o intuito da análise de desempenho das atletas, da modalidade de handebol.

3.1 Contextualização do jogo

Para realizar uma contextualização do jogo de handebol é interessante conhecer algumas regras do Handebol e todo o contexto sobre o qual estamos imergindo, com o intuito de nos aproximarmos e compreendermos, mais adiante, os dados utilizados.

Uma partida de handebol tem duração total de 60 minutos, os quais são compostos de dois tempos de 30 minutos, possuindo um intervalo de 10 minutos entre eles.

A modalidade é realizada em uma quadra cuja dimensão é de 20 metros de largura por 40 metros de comprimento, dividida ao meio por uma linha central. Cada lado contém uma área ao redor do gol (baliza) demarcada pela linha dos seis metros, em que apenas o goleiro pode estar dentro, e os demais jogadores não podem invadi-la e/ou pisar na linha quando a bola está em jogo. Também há a linha dos nove metros, que é a linha tracejada, e a linha dos sete metros, que fica posicionada entre as linhas de seis e nove metros, onde as jogadoras se posicionam para cobrar as faltas de $7m$. A Figura 3.1 ilustra essas demarcações da quadra.

Durante uma partida de handebol existem fases de jogo, sendo elas: ataque, defesa

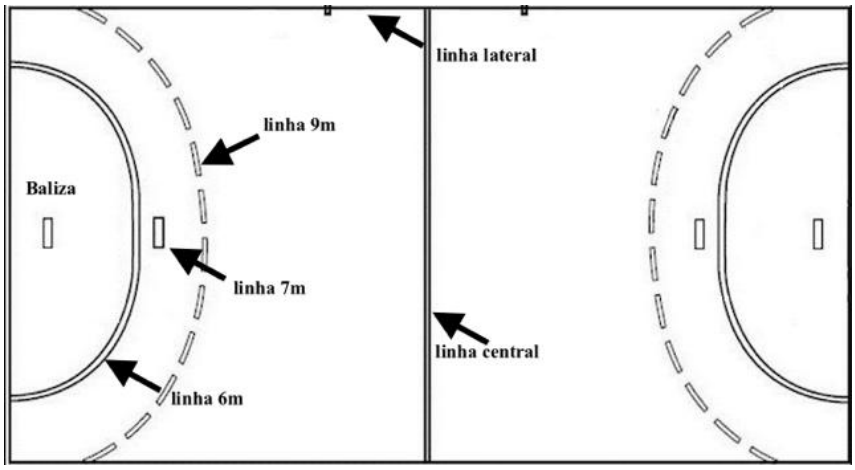

Figura 3.1: Representação das demarcações da quadra da modalidade de handebol.
Fonte: Produzida pela autora.

e transição (defensiva ou ofensiva). O ataque consiste em avançar na direção do gol adversário permanecendo com a posse de bola, buscando efetuar o gol tentando arremessar dos melhores locais. A defesa corresponde a dificultar os arremessos do time adversário procurando recuperar a posse de bola, evitando o avanço e gol do adversário. A transição defensiva consiste em um retorno rápido do ataque para o posicionamento de defesa, tentar recuperar a posse de bola e encaminhar os atacantes a locais da quadra de maior dificuldade para arremessar. Já na transição ofensiva tem a intenção de induzir a situações de contra ataque, do qual pretende-se manter a posse de bola e, rapidamente, arremessar de locais de quadra mais favoráveis aos atacantes.

No ataque existem duas linhas ofensivas, em que as seis jogadoras de linha (todas as jogadoras do time exceto a goleira) se dividem, sendo a primeira linha ofensiva a mais próxima do próprio gol e da linha central da quadra, e é composta pelos postos específicos de armadora esquerda/*left back* (A), armadora central/*center back* (B) e armadora direita/*right back* (C). E a segunda linha ofensiva, que é mais distante do próprio gol (próxima a linha de 6 m da área e gol do adiversário), é composta pelos postos específicos do ponta direta/*right wing* (D), pivô/*pivot* (E) e ponta esquerda/*left wing* (F). A Figura 3.2 ilustra os pontos específicos (posição) de cada jogadora no ataque.

As posições das jogadoras são:

- **Goleira** (*Golkeeper*): fica posicionada na área do gol, em que apenas ela pode ficar. Tem permissão para, durante uma tentativa de defesa, tocar na bola com qualquer parte do corpo. Não tem permissão de levar a bola que está fora da área do gol para dentro da área, contudo tem permissão de sair da área sem a posse da bola,

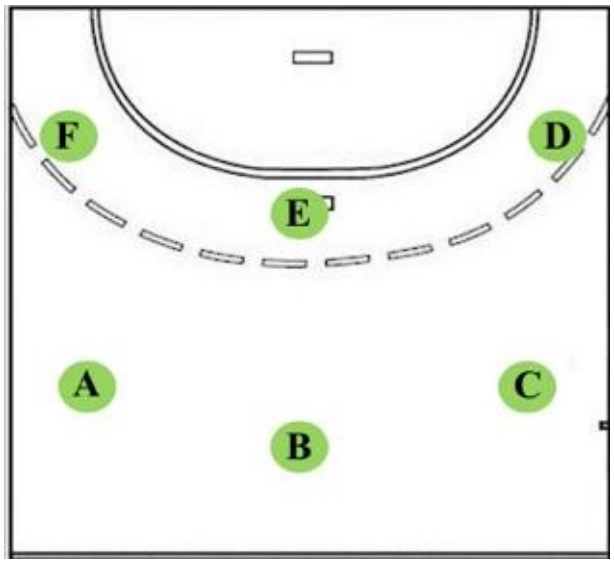

Figura 3.2: Representação dos pontos específicos da modalidade de handebol.
Fonte: Produzida pela autora.

passando a respeitar as mesmas regras que as demais jogadoras. As habilidades necessárias são conseguir antever possíveis pontos de ataque do time adversário, seu reflexo deve ser rápido, ser capaz de armar contra-ataques e, além disso, deve saber se comportar como jogadora de linha.

- **Armadoras** (*Back*): na maioria das vezes os times tem três armadoras (esquerda, central e direita), comumente estão em pontos específicos da primeira linha ofensiva. Por causa disso, são as atacantes com maior espaço de atuação e possibilidade de deslocamento. São exigidos dessas jogadoras um abrangente conjunto de ações motoras, como bons passes, deslocamento, recepção, arremessos, fintas e mudanças de direção das trajetórias da bola. São as responsáveis pela elaboração do jogo ofensivo, como o posicionamento da pivô e solicitam das demais jogadoras do ataque a execução de certos meios táticos ofensivos para que a equipe continue com a posse de bola;

- **Pontas** (*Wing*): tem como atuação as áreas extremas da quadra, próxima as linhas laterais, assim estando na segunda linha ofensiva. Desempenham a função de iniciar a movimentação rápida da bola, gerando ataques com maior velocidade, e ajudam as armadoras em situações de desequilíbrio ofensivo. Devem possuir bom deslocamento (com e sem a bola), bons passes, arremessos, recepções, dribles e fintas;

- **Pivô** (*Pivot*): Sua capacidade de deslocamento sem a bola deve ser boa (para dificultar a movimentação das defensoras, assim criando chances de arremesso para

as demais jogadoras) e também deve ser boa com a posse de bola (objetivando atrair a defensora, dando assim continuidade ao jogo ofensivo). Além disso, deve saber circular entre as armadoras com o intuito de causar um desequilíbrio no sistema defensivo do time adversário. Seu posicionamento pode ser de costas para o gol ou lateralmente em relação ao gol. A primeira posição proporciona uma situação de bloqueio frontal da defensora, facilitando arremessos de maior distância, já o segundo posicionamento facilita o bloqueio lateral, o giro após receber a bola e facilita a infiltração das armadoras na defesa do time adversário.

Os tipos de punições que uma jogadora pode sofrer, definida pela IHF (2016), são as seguintes:

- **7 metros**: uma jogadora da equipe que sofreu a agressão, ao realizar a cobrança da penalidade, posiciona-se na linha dos sete metros, frente a frente com a goleira, enquanto as demais jogadoras (de ambos os times) devem estar fora (para trás) da linha dos 9 metros. Um exemplo de ação punida com a aplicação dos sete metros é quando ocorre alguma agressão/infração em que há uma chance clara de gol, próximo a área dos seis metros;

- **2 minutos**: corresponde a penalização em que a atleta ficará excluída da partida por dois minutos, deixando o time em desvantagem numérica (não podendo ser substituída nesse período). Exemplos de ações que resultam nessa punição são: quando a jogadora comete uma infração contra uma adversária que está correndo com velocidade elevada ou cometidas com alta intensidade; faltas contra o pescoço, cabeça de outra jogadora, entre outras situações;

- **Cartão amarelo**: uma advertência (comunicada pelo árbrito mostrando o cartão amarelo) é uma punição adequada para ações e/ou condutas antidesportivas que devam ser sancionadas progressivamente, em que a ação do infrator é exclusiva ou principalmente dirigida ao corpo da jogadora adversária;

- **Cartão vermelho**: é a desqualificação da atleta que será, sempre, para todo o restante de tempo do jogo, devendo essa se retirar imediatamente da quadra e área de substituição, e acaretará uma exclusão de dois minutos para a equipe. Exemplos do que gera esta punição é a terceira exclusão (2 minutos) de uma mesma jogadora ou uma atleta que ataca sua adversária de modo perigoso para sua saúde;

- **Cartão azul**: corresponde a uma desqualificação (cartão vermelho) com acréscimo, depois do término da partida, de um relatório escrito, com o intuito de viabilizar que as autoridades responsáveis decidam as medidas posteriores a serem tomadas. Exemplos de ações que resultam nesse tipo de punição são: uma conduta antidesportiva extrema, ações perigosas, imprudentes, premeditadas ou maliciosas, sendo as duas últimas não relacionadas de nenhuma maneira com a situação do jogo.

As penalidades podem ser aplicadas isoladamente, combinadas ou sancionadas progressivamente (para a própria jogadora ou equipe) dependendo da infração cometida. Um exemplo disto é que cada jogadora tem a possibilidade de uma advertência, e até três para cada equipe, sendo que posteriormente a punição passa a ser de exclusão de dois minutos ou desclassificação.

Os treinadores buscam jogadoras que tenham uma fluidez posicional de jogo, que elas tenham o conhecimento de outras posições, além da posição que jogam, assim percorrendo a quadra toda, devido ao dinamismo que o jogo ofensivo apresenta.

A importância dessas habilidades é devido a colaborarem na antecipação de jogadas, em função da atleta saber as possíveis ações da adversária, portanto saber resolver diversos problemas de diferentes situações de jogo. Logo, todas as jogadoras devem desenvolver o máximo de características específicas de cada uma das posições de jogo para intervir de forma astuta e multifacetada nas diversas situações ofensivas (Menezes, 2011).

Um exemplo disso, são atletas que jogam na posição de ponta e que apresentam características da posição de armadora, no qual sabem jogar na primeira linha ofensiva. Então, as atletas atualmente, precisam saber se posicionar e jogar em outras áreas da quadra que não a correspondente a sua posição de jogo. Contudo, é importante que essas jogadoras melhorem competências específicas da posição que habitualmente atuam.

3.2 Conjunto de dados

O conjunto de dados a ser estudado é sobre jogadoras de handebol do campeonato mundial feminino juvenil do ano de 2018 e foi disponibilizado pela IHF, em seu *site*[1]. Este campeonato ocorreu em Kielce, na Polônia, sendo a sétima edição do torneio, datado de 7 a 19 de agosto de 2018. A seleção da Rússia foi a grande campeã da competição, em

[1]https://archive.ihf.info/en-us/ihfcompetitions/worldchampionships/womensyouthworldchampionships/2018womensyouth(u18)worldchampionship/teaminfo.aspx

uma final contra a Hungria.

No *site* são disponibilizadas tabelas contendo informações das 24 equipes que fizeram parte do campeonato. Essas tabelas contêm alguns dados sobre quantidade de gols marcados por cada jogadora, quantidade de cartões (punições) que cada jogadora recebeu, quantidade de partidas que participou, além de disponibilizar outras tabelas com algumas informações pessoais de cada jogadora, como: nome, posição que joga, idade, altura (em cm), peso (em kg), entre outras informações.

No banco de dados apresentado foram contabilizadas diversas variáveis, em torno de 35, em cada um dos nove jogos realizados, sendo que times que foram eliminados antes das semifinais tem uma quantidade menor de jogos realizados, por se tratar de um campeonato eliminatório. As informações são de todas as jogadoras que foram convocadas para suas respectivas seleções, logo fizeram parte do time oficial que participou do campeonato. Essas variáveis possibilitam informar a quantidade de assistências, roubadas de bola, ataques realizados, conversão de arremesso e faltas cometidas, feitas pelas jogadoras.

Para a aplicação, não serão consideradas na análise as atletas que jogam na posição de goleira e jogadoras com um número inferior a cinco partidas jogadas. Lembrando que as seleções participaram, durante todo o campeonato, de no máximo nove jogos e no mínimo de sete partidas.

Essas atletas são retiradas devido a obterem muitas variáveis com valores iguais a zero, causando uma influência irrelevante na análise e/ou uma atuação diferente no jogo comparado as demais posições e jogadoras. Como é o caso das goleiras, devido a dificuldade de comparar goleiras com jogadoras de posições diferentes.

A utilização desse banco de dados é coerente, visto que o objetivo desse estudo é definir o desempenho das jogadoras durante todo o campeonato. Portanto, ter um único valor para cada variável de cada atleta é correto, correspondendo, então, a soma de todas as partidas jogadas em um único valor.

No Apêndice A, está contido uma tabela com as observações de todas as variáveis da seleção da Angola, com o intuito de ilustrar como é o banco de dados que será utilizado na aplicação.

Pela Tabela A.1, observa-se que a maior parte das variáveis são discretas, já que mostram a quantidade de vezes que certa ação foi executada. Nota-se, também, que no time da Angola nenhuma jogadora recebeu cartão vermelho e no banco de dados não foi considerado a punição por cartão azul, já que nenhuma das seleções participantes do

campeonato recebeu este cartão. O fato de não ter sido aplicado o cartão azul é muito comum em grandes competições e, como o evento que está sendo estudado é mundial, a não ocorrência dessa punição é desejável, por se tratar de uma punição para atitude antidesportiva extrema do jogador, como foi explicado na Seção 3.1.

Por fim, uma breve explicação das variáveis que compõem o banco de dados. Essas variáveis são:

- **Name**: corresponde ao nome da atleta;
- **Position**: corresponde a posição que a atleta joga;
- **DOB**: é a data de nascimento da atleta;
- **Age**: corresponde a idade da atleta em anos;
- **Club**: corresponde ao clube pelo qual a atleta joga;
- **cm**: corresponde a altura em centímetros (cm) da atleta;
- **kg**: corresponde ao peso em quilogramas (kg) da atleta;
- **IM**: corresponde ao número de jogos internacionais que a atleta participou;
- **IG**: corresponde ao quantidade de gols que a atleta fez em jogos internacionais;
- **G**: corresponde ao número de jogos que a atleta estava na seleção, integrava o time podendo participar da partida ou ficar no banco de reserva durante a partida durante o campeonato;
- **P**: corresponde ao número de jogos que a atleta jogou pela seleção, número de vezes que entrou ativamente no jogo durante o campeonato;
- **Goals**: é a quantidade total de gols feitos pela jogadora durante todo o campeonato;
- **Shots**: é a quantidade total de arremessos feitos pela jogadora durante todo o campeonato;
- **Av.g**: corresponde a média de gols da atleta por jogo participado;
- **Wing_g (gW)**: é o número de gols que a atleta realizou nas áreas das posições de pontas (área extrema da quadra) durante todo o campeonato;

- **Wing_s (sW)**: corresponde a quantidade de arremessos que a atleta realizou nas áreas das posições de pontas durante toda a competição;

- **7m_g (g7m)**: é o número de gols que a atleta realizou na linha dos sete metros durante todo o campeonato;

- **7m_s (s7m)**: corresponde a quantidade de arremessos que a atleta realizou na linha dos sete metros durante toda a competição;

- **9m_g (g9m)**: é o número de gols que a atleta realizou antes da linha dos nove metros durante todo o campeonato;

- **9m_s (s9m)**: corresponde a quantidade de arremessos que a atleta realizou antes da linha dos nove metros durante toda a competição;

- **6m_g (g6m)**: é o número de gols que a atleta realizou na linha dos seis metros durante todo o campeonato;

- **6m_s (s6m)**: corresponde a quantidade de arremessos que a atleta realizou na linha dos seis metros durante toda a competição;

- **Break_g (gBk)**: é o avanço (*Breakthrough*) correspondente a quantidade de gols realizados quando a jogadora se infiltra na defesa adversária;

- **Break_s (sBk)**: corresponde a quantidade de arremessos realizados quando a jogadora se infiltra na defesa adversária;

- **Fastbreak_g (gFb)**: é o número de gols que a atleta realizou em um contra-ataque rápido durante todo o campeonato. *Fastbreak* corresponde a um contra-ataque rápido, em que o time coloca a bola para dentro da quadra e chega para a posição de gol o mais rápido possível, com a intenção de não possibilitar que a defesa adversária consiga se organizar;

- **Fastbreak_s (sFb)**: corresponde a quantidade de arremessos que a atleta realizou em um contra-ataque rápido durante toda a competição;

- **Assists (Ass)**: é o número de passes (assistências) que uma atleta realiza, durante todo o campeonato, para outra jogadoras efetuar o gol. Assistência corresponde ao passe de bola que em subsequência resulta em gol de outra jogadora da equipe;

- **Turnovers (Tur)**: corresponde a quantidade de vezes que a atleta provocou, durante todo o campeonato, a troca da posse de bola (do seu próprio time) para o time adversário, devido a penalidade, falta, por deixar passar a bola dos limites laterais da quadra ou por cometer alguma infração (um exemplo é quando a jogadora toca na bola com os pés sem ter a intenção);

- **Steals (Ste)**: corresponde a quantidade de vezes que a atleta roubou a bola do time adversário durante todo o campeonato. Não é permitido a roubada de bola do adversário usando as duas mãos;

- **7m_comm. (c7m)**: é o número de penalidades de sete metros que a atleta provocou durante todo o campeonato;

- **7m_rec. (r7m)**: corresponde a quantidade de penalidades de sete metros que a atleta sofreu durante toda a competição;

- **2m_pun. (p2m)**: corresponde a quantidade de penalidades de 2 minutos que a atleta realizou, provocou, durante toda a competição. É a quantidade de vezes que a atleta ficou dois minutos, como punição, fora da quadra;

- **2m_rec. (r2m)**: é o número de penalidades de dois minutos que a atleta sofreu durante todo o campeonato. É a quatidade de vezes que a jogadora sofreu uma falta que acarretou, como punição, para outra atleta ficar 2 minutos fora de quadra;

- **Yellow (Y)**: corresponde a quantidade de cartões amarelos (advertências) que a atleta recebeu durante todo o campeonato, ou seja, a jogadora realizou uma falta leve ou média;

- **Red**: corresponde a quantidade de cartões vermelhos (desqualificações) que a atleta recebeu durante todo o campeonato, ou seja, a jogadora foi suspensa da partida;

Capítulo 4

Resultados

Esse capítulo aborda o tratamento realizado com o banco de dados. A Seção 4.1, trata de todas as transformações e exclusões realizadas com o banco de dados. Nas Seções 4.2.1 e 4.2.2 são apresentados os resultados das análises de componentes principais das variáveis relacionadas a punições e posteriormente considerando todas as variáveis. A Seção 4.3 apresenta os resultados obtidos pelos diferentes métodos de agrupamento, uma validação do número de grupos através de métodos gráficos, e por fim uma caracterização dos grupos para entender quais as informações cada um deles expressa.

4.1 Análise dos Dados

Para esse trabalho, como dito na Seção 3.2, não serão considerados as jogadoras da posição de goleira e jogadoras com menos de cinco partidas jogadas.

As jogadoras que tinham informações faltantes referentes as variáveis peso e altura tiveram seus dados imputados considerando a mediana dessas características com relação a posição em que essas jogadoras atuam.

Do banco de dados algumas variáveis não serão consideradas, em razão de não contribuírem com muita informação, cinco dessas variáveis são: cm, kg, IM, IG e Red. Em relação as variáveis IM e IG serão excluídas da análise porque tem, respectivamente, 126 e 160 valores faltantes, logo têm mais de 30% e 40% de dados faltantes. Já a Red, foi retirada, pelo motivo de somente 6 jogadoras haverem recebido cartão vermelho durante toda a competição. Já as variáveis cm e kg não foram consideradas devido a apresentarem pouca variabilidade. Calculamos o IMC (Índice de Massa Corporal) com a intenção de utilizá-lo no lugar das variáveis de peso e altura, porém esse apresentou pouca variabilidade

entre as atletas, portanto também não será utilizado.

Além disso, todas as variáveis relacionadas as quantidades de gols, arremessos, passes e punições foram divididas pela quantidade de partidas que a jogadora participou efetivamente do jogo (P). Logo, se obtém as médias por jogo dessas respectivas variáveis.

Após realizada a exclusão, imputação dos dados faltantes e a divisão das variáveis pelo número de jogos que a atleta participou, supracitados, o banco de dados se resume a 30 variáveis e 271 jogadoras. Com os gráficos de barras representados na Figura 4.1, podemos perceber que a posição com maior quantidade de atletas é a *Center back* e a maioria das jogadoras tem entre 17 e 18 anos, como esperado já que os dados se referem a um campeonato de categoria *sub-18*. Também podemos notar que a quantidade de atletas por times ficou entre 8 e 14 jogadoras e dessas 120 participaram ativamente de somente sete jogos.

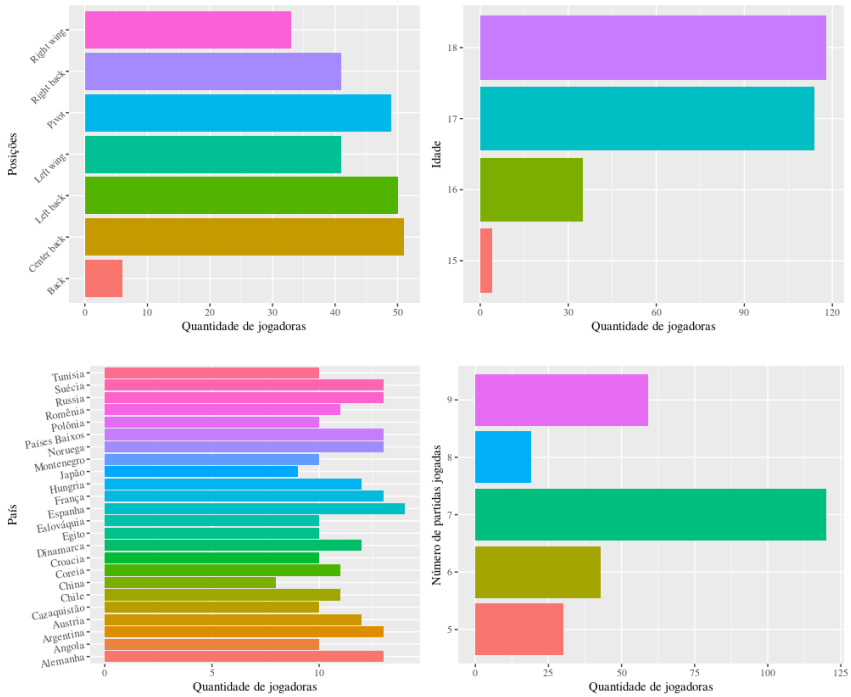

Figura 4.1: Gráfico de barras com as quantidades de jogadoras por posição, idade, país e quantidade de partidas jogadas.

4.2 Análise Componentes Principais

A técnica de componentes principais (ACP) foi aplicada, inicialmente, nas variáveis relacionadas com a punição que as jogadoras sofreram ou provocaram. Realizamos o ACP das variáveis de punição, separadamente, porque quando consideravamos todas as

variáveis juntas, a explicação das variáveis relacionadas as punições eram vistas somente apartir da décima dimensão, portanto consideramos mais vantajoso observarmos separadamente essas variáveis e posteriormente inclui-lás em um ACP com as demais variáveis.

Em seguida, foi aplicada novamente a técnica em todas as variáveis, substituindo as variáveis relacionadas a punição e considerando em seu lugar os componentes criados pela ACP das variáveis de punição. E como é mostrado na Seção 4.2.2, esses índices criados pela ACP das variáveis de punição são bem explicados até a quinta dimensão.

4.2.1 Análise Componentes Principais das Variáveis de Punição

Posteriormente ao tratamento dos dados, foi realizado uma ACP para as variáveis 7m_comm. (c7m), 7m_rec. (r7m), 2m_pun. (p2m), 2m_rec. (r2m) e Yellow (Y), que são variáveis relacionadas a punição que as jogadoras sofreram ou provocaram durante todo o campeonato.

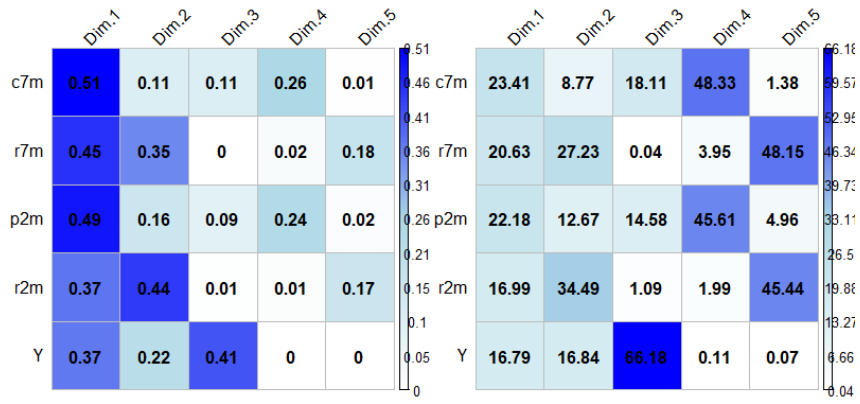

Figura 4.2: Matriz de correlação das variáveis com os componente (matriz à esquerda) e a matriz de contribuição das variáveis em cada um dos componentes (matriz à direita).

A ACP foi aplicada nas variáveis de punição e obtivemos os resultados apresentados na Figura 4.2. Através da matriz de correlação das variáveis com os componentes, é possível verificar o quanto cada variável é correlacionada com cada componente, podemos notar que todas as variáveis de punição são bem explicadas pelo primeiro componente, com valores maiores que 0.37. Enquanto isso, observamos que as variáveis ligadas as punições que atleta sofreu durante toda a competição, sendo essas as variáveis r7m e r2m, também têm grande parte de sua informação representada no segundo componente.

Podemos ver que utilizando só as duas primeiras componentes é possível obter uma boa informação de todas as variáveis.

Já em relação a matriz de contribuição das variáveis em cada um dos componentes, ela ilustra quanto de informação total cada variável contribui dentro de cada componente, ou seja, quanto de cada componente é explicado por cada variável. Notamos que as variáveis que mais contribuem para o primeiro componente principal são, respectivamente, c7m e p2m, contribuindo 23.41% e 22.18% para a variabilidade total desse componente. Essas variáveis informam a quantidade de infrações de sete metros e dois minutos que a jogadora provocou durante toda a competição. Porém, as variáveis relacionadas a quantidade de infrações que a atleta sofreu, r2m e r7m, contribuem com mais informação no segundo componente, portanto são as que mais contribuem dentro desse componente.

O terceiro componente expressa mais informação pela variável Yellow (Y), que contribui 66.18% para a variabilidade total desse componente. Por esse motivo poderia ser interessante utilizar o terceiro componente, porém, as duas primeiras componentes já foram capazes de captar bastante da informação desta variável. Além disso, esse componente só explica 12.5% da variabilidade total das atletas, como podemos ver pelo gráfico de cotovelo da Figura 4.3.

Após analisados as matrizes de correlação e de contribuição, um gráfico de cotovelo e o círculo unitário para o primeiro plano foram construídos. Os resultados obtidos podem ser observados na Figura 4.3.

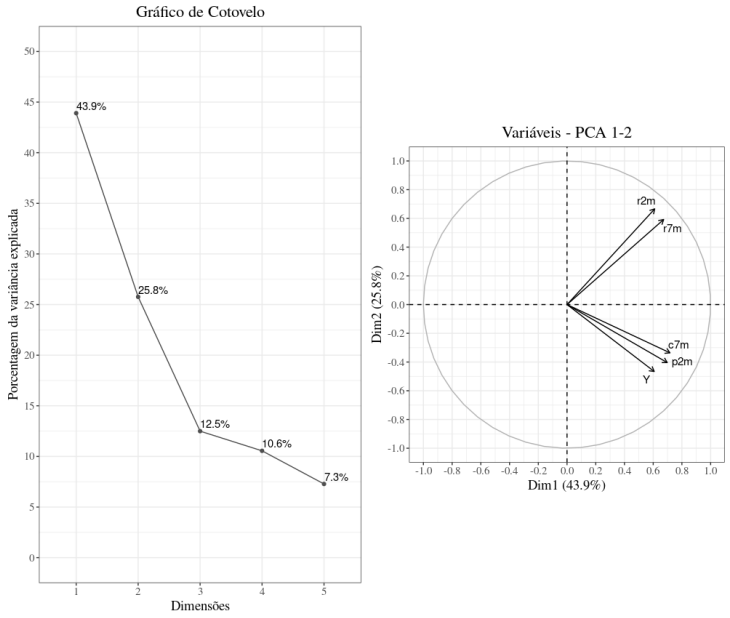

Figura 4.3: Gráfico de cotovelo para os componentes (gráfico à esquerda) e círculo unitário para o plano 1-2 (gráficos à direita).

Podemos perceber, pelo gráfico de cotovelo apresentado na Figura 4.3, que utilizando os dois primeiros componentes há uma explicação de quase 70% da variabilidade total dos dados, que é o suficiente para explicar a informação expressa pelas cinco variáveis relacionadas com a punição que as jogadoras sofrem ou provocam. Por meio dos resultados, somos capazes, pelo círculo unitário do primeiro plano, de observarmos que as cinco variáveis de punição são bem representadas no primeiro plano, já que se encontram próximas da circunferência unitária. Notamos que r2m e r7m estão fortemente correlacionados entre si, mas são independentes das variáveis c7m, p2m e Y, as quais, entre si apresentam uma correlação forte.

O primeiro componente podemos entender como um índice geral de agressividade da jogadora, porque está relacionado com a média de faltas realizadas por jogo, considerando todas as variáveis de punição. O índice é definido como um valor positivo, logo quanto maior os valores das variáveis observadas em cada jogadora maior será o valor do índice. Entretanto se esse valor observado das variáveis for pequeno, o valor do índice será menor e mais próxima do zero ela estará posicionada.

Com relação ao segundo componente podemos interpretar como um índice que mede o *fairplay* das atletas, porque percebemos um contraste entre dois grupos de variáveis. O primerio grupo, variáveis localizadas no primeiro quadrante, é composto por variáveis com pesos positivos: média de penalidades de sete metros e dois minutos que a atleta sofreu por jogo. Já as variáveis do segundo grupo, do quarto quadrante, a saber a média de punição de sete metros, dois minutos e cartão amarelo que a jogadora provocou por jogo, contribuem de forma negativa no índice. Portanto, quanto mais positivo for o valor do índice, maior *fairplay* houve no jogo da atleta, porque ela sofreu mais punições do que provocou. Contudo, quanto mais negativo for o índice, significa que a jogadora realizou mais infrações, logo foi punida por isso, prejudicando seu time que ficou em desvantagem a cada punição provocada.

Para ilustrar o comportamento dos índices de agressividade e *fairplay*, criados pela ACP, realizamos uma análise exploratória básica. Com os *boxplots* representados na Figura 4.4, podemos notar que o índice de agressividade, quando observamos as posições é, em geral, maior para as pivôs seguidas das atletas que jogam como *right back*. Essas são as posições mais agressivas devido a terem as maiores medianas e os maiores valores discrepantes. Quando levamos em consideração os time, a Polônia é a seleção considerada a mais agressiva.

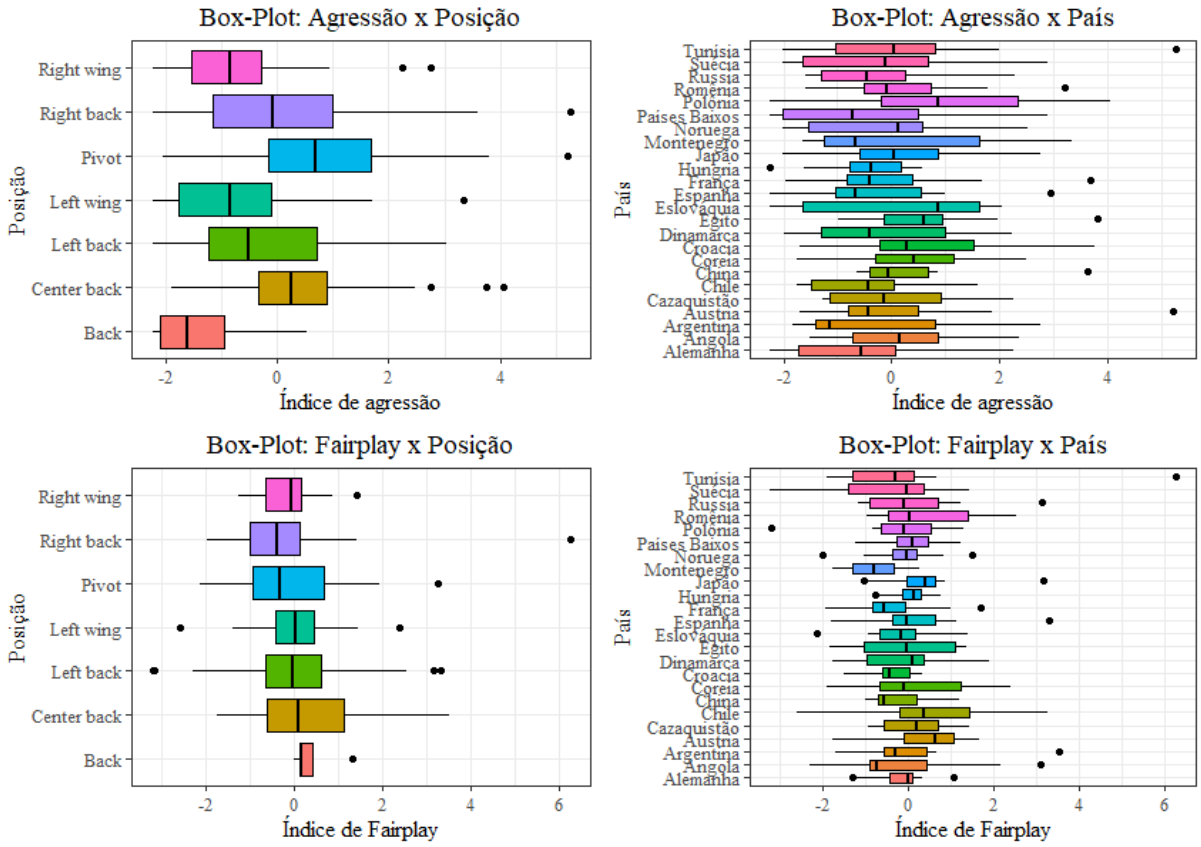

Figura 4.4: Índice de agressão e *fairplay*, referente as posições e países das jogadoras.

Podemos perceber que o índice de *fairplay*, quando separado pela posição que as atletas jogam, tem medianas muito próximas. Observamos que mesmo as jogadoras da posição pivô, que são extremamente agressivas, apresentam um índice de *fairplay* próximo de zero, logo são atletas que provocaram muitas penalizações porém também sofreram muitas penalidades. Então, a maioria das jogadoras sofre e provoca penalidades na mesma dimensão. Além disso, quando levamos em consideração o país a qual elas jogam, vemos uma intersecção das caixas em quase todas as seleções, também notamos que a seleção da Alemanha tem valores *outliers*, contudo não são valores tão altos se comparamos com os demais nove países que tem pontos *outliers*, maiores que os desse time.

4.2.2 Análise Componentes Principais com todas as Variáveis

Posteriormente à criação das variáveis de agressividade e *fairplay*, pelo uso da técnica análise de componentes principais, foi realizado uma ACP com essas duas variáveis e as demais variáveis quantitativas do banco de dados. Os resultados obtidos podem ser observados na Figura 4.5.

Notamos, pelo gráfico de cotovelo, que utilizando os cinco primeiros componentes há

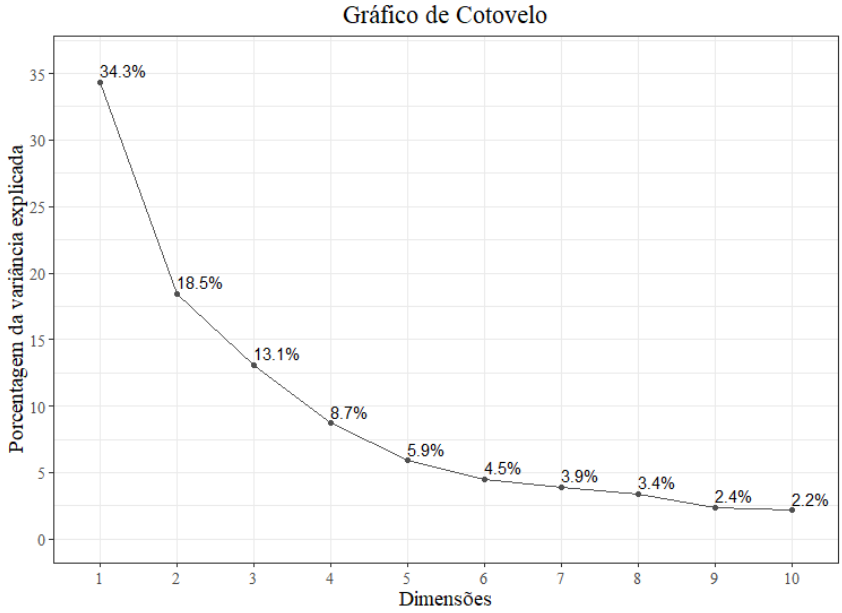

Figura 4.5: *Scree Plot* para os 10 primeiros componentes.

uma explicação de mais de 80% da variabilidade total dos dados, sendo esses componentes suficientes para uma boa representação das informações apresentadas pelas variáveis quantitativas presentes no banco de dados.

Posteriormente à análise do gráfico de cotovelo, foram construídas as matrizes de correlação e de contribuição, para os primeiros cinco componentes. Obtivemos os resultados apresentados na Figura 4.6.

Podemos perceber, pela matriz de correlação das variáveis com os componentes, que as variáveis *Goals*, *Shots*, g7m, s7m, g9m, s9m, gBk, sBk, Ass e Tur são bem explicadas pelo primeiro componente, com valores maiores que 0.43. No caso, do segundo componente, as variáveis mais explicadas são gW, sW, gFb e sFb com valores maiores que 0.70. Enquanto isso, observamos que as variáveis g6m, s6m e agre têm a maior parte da informação no terceiro componente. Já a quarta componente, é majoritariamente explicada por g7m, s7m, gBk e sBk, com valores maiores que 0.19. Por fim, a variável *fairplay* é majoritariamente explicada pelo quinto componente. Podemos ver que utilizando somente os cinco primeiros componentes é possível obter uma boa informação de todas as 19 variáveis.

Com relação a matriz de contribuição das variáveis em cada um dos componentes, notamos que as variáveis que mais contribuem para o primeiro componente são *Shots*, *Goals*, s9m, g9m, Tur, sBk, Ass, gBk, s7m e g7m, respectivamente, contribuindo entre 12.13% até 6.58% para a variabilidade total desse componente. Essas variáveis, na maior

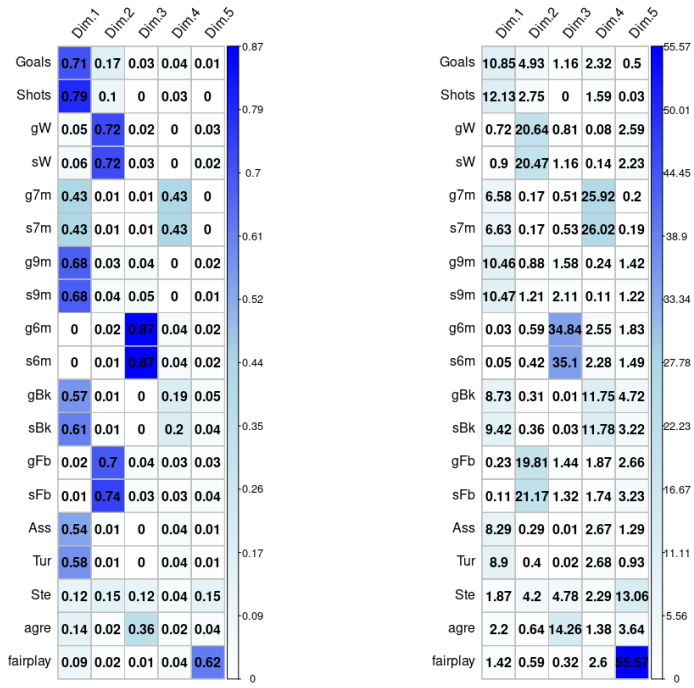

Figura 4.6: Matriz de correlação das variáveis com os componente (matriz à esquerda) e a matriz de contribuição das variável em cada um dos componentes (matriz à direita).

parte informam a média de arremessos e gols realizados pelas jogadoras, por partida, de diversas regiões da quadra. Com exceção das variáveis Tur (*Turnovers*) e Ass (*Assists*) que correspondem, na devida ordem, à média por jogo que a atleta foi responsável pela troca da posse de bola (por falta, penalidade ou por deixar a bola sair dos limites da quadra) e à média de assistências da jogadora que resultaram em gols de outra jogadora do time.

Já as variáveis s6m e g6m são as que menos contribuem para a informação do primeiro componente, respectivamente, 0.05% e 0.03%. Porém são as que contribuem 34.85% e 35.19% do terceiro componente, portanto as que mais contribuem para a informação do terceiro componente.

No segundo componente, observamos que as variáveis que mais contribuem para esse componente são sFb, gW, sW e gFb, nessa ordem. Contudo, as variáveis g7m, s7m, sBk e gBk são as que menos contribuem na informação da variabilidade total desse componente, no entanto são as que mais contribuem no quarto componente, sendo as que mais contribuem dentro do quarto componente.

A variável *fairplay* é a que mais contribui para a informação da variabilidade total do quinto componente, que contribui 55.57% da informação desse componente. Lembrando que esse componente só explica 5.90% da variabilidade total das jogadoras, como pode

ser visto pelo gráfico de cotovelo, na Figura 4.5.

Após analisados as matrizes de correlação e de contribuição, um gráfico com o círculo unitário para os planos 1-2, 1-3, 1-4, 1-5 e 2-3 foram construídos. Os resultados obtidos podem ser observados na Figura 4.7.

Por meio dos resultados, pelo círculo unitário do primeiro plano ($Dim1 \times Dim2$), observamos que as variáveis sW, gW, *Goals*, *Shots*, g9m, s9m, gBk, sBk, Ass e Tur são bem representadas no primeiro plano, já que se encontram próximos da circunferência unitária. Notamos que gBk, sBk, Ass e Tur estão fortemente correlacionadas entre si, mas são independentes de sFb e gFb, as quais, entre si exibem uma correlação forte. Vemos que as variáveis sW e gW são fortemente correlacionadas, contudo são independentes de *Goals* e *Shots*, sendo que esses são correlacionados entre si. Já no caso das variáveis s6m e g6m, são fortemente correlacionadas entre si, porém são as variáveis que estão mais distantes do círculo unitário, portanto são mal representadas por esse plano.

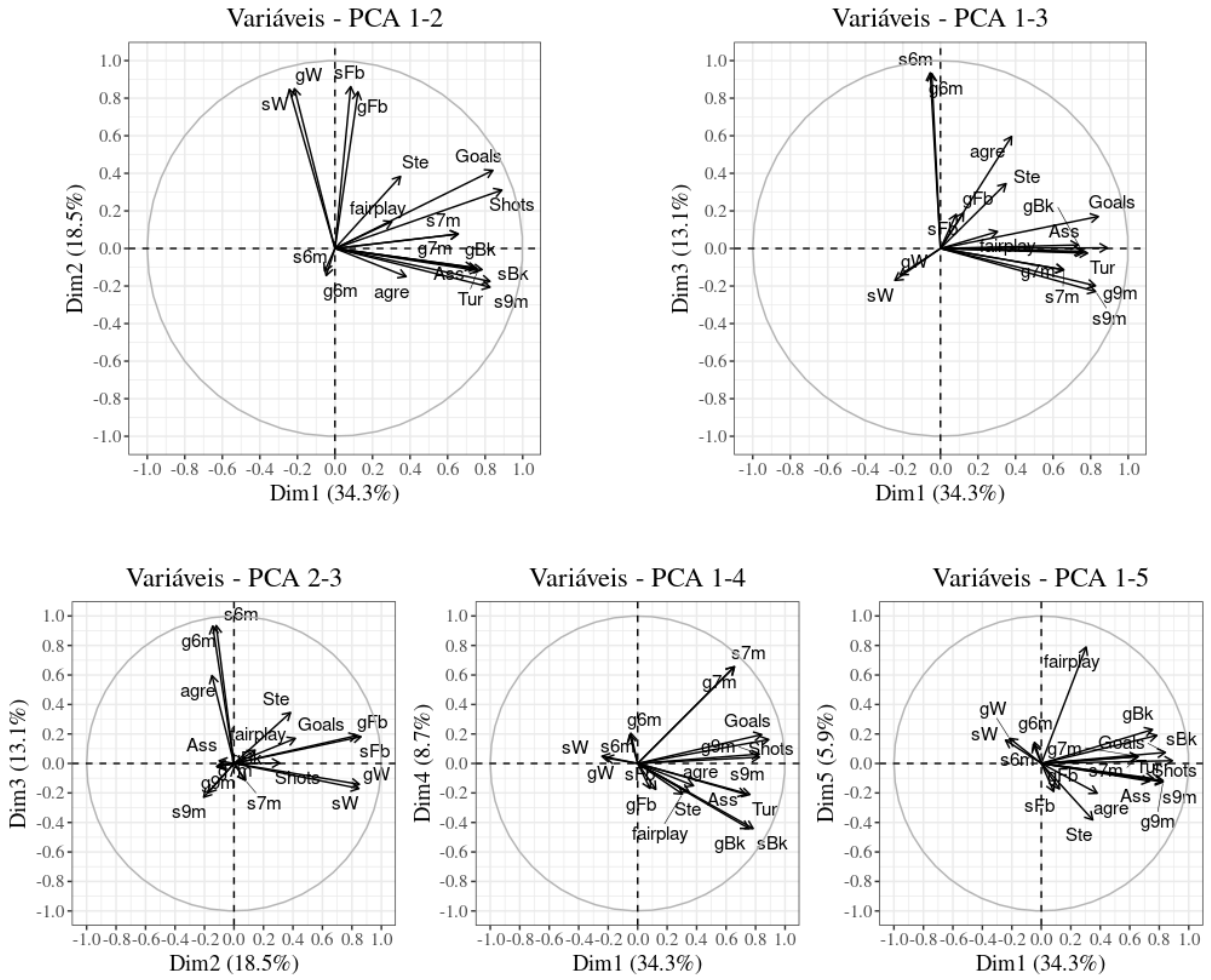

Figura 4.7: Círculo unitário para os planos 1-2, 1-3, 2-3, 1-4 e 1-5.

Podemos observar, pelo segundo plano ($Dim1 \times Dim3$), que as variáveis s6m e g6m, estão próximas do cículo unitário, logo são bem representadas no segundo plano. As variáveis agre e Ste são melhores representadas, em comparação ao primeiro plano, porém sW e gW passam a ser mal representadas nesse plano. Em relação as demais variáveis, que eram bem representadas no primeiro plano, continuam a ser bem representadas no segundo plano. Notamos que s6m e g6m são fortemente correlacionadas entre si, mas são independentes das variáveis Ass, *Shots*, gBk, sBk e Tur, as quais, entre si apresentam uma correlação muito forte, já que estão quase sobrepostos. Além disso, vemos que o índice de agressividade (agre) tem uma correlação negativa com as variáveis sW e gW.

Quando observamos, o círculo unitário do terceiro plano ($Dim2 \times Dim3$), que as variáveis g6m, s6m, gFb, sFb, gW e sW são bem representadas nesse plano. Percebemos também que g6m e s6m são fortemente correlacionadas entre si, assim como gW e sW, que são variáveis bem correlacionadas entre si. Contudo, g6m e s6m são independentes de gFb e sFb, as quais, entre si são fortemente correlacionadas. Com relação as demais variáveis, elas não são bem representadas nesse terceiro plano.

No quarto plano ($Dim1 \times Dim4$), percebemos que as variáveis g7m e s7m são as com melhor representação nesse plano, já que são as mais próximas do círculo unitário. Ainda notamos que essas variáveis são fortemente correlacionadas entre si, no entanto são independentes de gBk e sBk, as quais, entre si exibem uma correlação forte.

O quinto, e último, plano ($Dim1 \times Dim5$) observamos que o índice *fairplay* é melhor representado nesse plano, em comparação aos outros planos. Vemos que o índice é independente das variáveis Ass, Tur, s9m e g9m, as quais, são fortemente correlacionadas entre si.

Então, podemos entender o primeiro componente como um índice de habilidade gerais das jogadoras (Índice 1), já que considera diversos tipos de arremessos e gols (*Goals*, *Shots*, g7m, s7m, g9m, s9m, gBk, sBk) além de variáveis de colaboração de jogo (Ass e Tur). Sendo definido como um valor positivo, quanto maior os valores observados nas variáveis maior o valor do índice. O segundo componente podemos interpretar como um índice que mede as habilidades na posição de ponta (Índice 2), porque está relacionado com tipos de gols e arremessos que são geralmente realizados por atletas que ocupam a posição de ponta em quadra. Percebemos que quanto maior os valores de gols e arremessos do tipo *wing* e *fastbreak* maior é o valor desse índice.

Com relação ao terceiro componente, podemos entender como um índice de habili-

dades na posição de pivô (Índice 3), devido considerar g6m, s6m e agre, que é um tipo de arremesso realizado com maior frequência por jogadoras da posição de pivô, e como visto na Seção 4.2.1 são as atletas mais agressivas.

Podemos interpretar o quarto componente como um índice de sete metros ou *breakthrough* (Índice 4), porque percebemos um contraste entre dois grupos de variáveis que são bem representadas nesse componente. Observando o quarto plano ($Dim1 \times Dim4$), vemos que o primeiro grupo, variáveis localizadas no primeiro quadrante: média de gols e arremessos de sete metros por jogo. Já as variáveis do segundo grupo, do quarto quadrante, reference a média de gols e arremesso de *breakthrough*, que contribuem de forma negativa para o índice. Portanto, quanto mais positivo for o valor do índice 4, maior a quantidades de gols e arremessos a jogadora fez de sete metros. Contudo, quanto mais negativo for o valor do índice 4, significa que a atleta realizou mais gols e arremessos de *breakthrough* por jogo.

O quinto e último componente, podemos entender como um índice de *fairplay* (Índice 5), tendo a mesma interpretação apresentada na Seção 4.2.1.

4.3 Resultados da Análise de Agrupamento

Em seguida, a realização da ACP com todas as variáveis quantitativas, foi realizado métodos de agrupamento hierárquicos e não-hierárquico. Os agrupamentos foram realizados pelas funções "*hclust*"e "*kmeans*", respectivamente, ambas contidas no R.

Como dito na Seção 2.1, os métodos hierárquico e não-hierárquico são os métodos mais populares. Começaremos realizando o método hierárquico, devido a facilidade de escolher a quantidade de grupos, ao qual pretendemos particionar as jogadoras. Isso deve-se ao dendrograma, que é uma visualização dos agrupamentos que facilita essa decisão, sendo essa a vantagem e causa de iniciarmos com esse método. Posteriormente, utilizamos as médias dos grupos dados pelo método hierárquico como centroides do algoritmo de K-Médias.

Para os agrupamentos, utilizamos todos os componentes da ACP como entrada (*input*), para não perdermos nenhuma informação sobre a variabilidade dos dados. No entanto, para facilitar na interpretação utilizaremos somente os cinco primeiros componentes, sendo esses os índices que serão analisados nas próximas Seções.

4.3.1 Agrupamento Hierárquico

Verificamos todos os tipos de medidas de distância da Seção 2.1 e a medida de distância final utilizada para a construção da matriz de distâncias foi a medida de *city block*. Optamos pela medida de distância de *city block*, por ela considerar o módulo das diferenças, consequentemente faz com que os pontos extremos não fiquem tão expressivos. Portanto, utilizamos essa medida, já que ela é menos sensível para pontos *outliers*, devido aos dados possuírem pontos extremos.

O método Ward, foi o agrupamento hierárquico utilizado por criar grupos homogêneos de forma e tamanhos semelhantes. Os resultados obtidos estão apresentados na Figura 4.8. O dendrograma indica que uma quantidade satisfatória de grupos é de quatro grupos. Realizamos o corte no peso igual a 60, sendo o peso proporcional a soma de quadrados entre esses grupos, já que notamos que não ocorre uma diminuição significativa na soma de quadrados entre os grupos quando cortamos em uma quantidade maior de grupos.

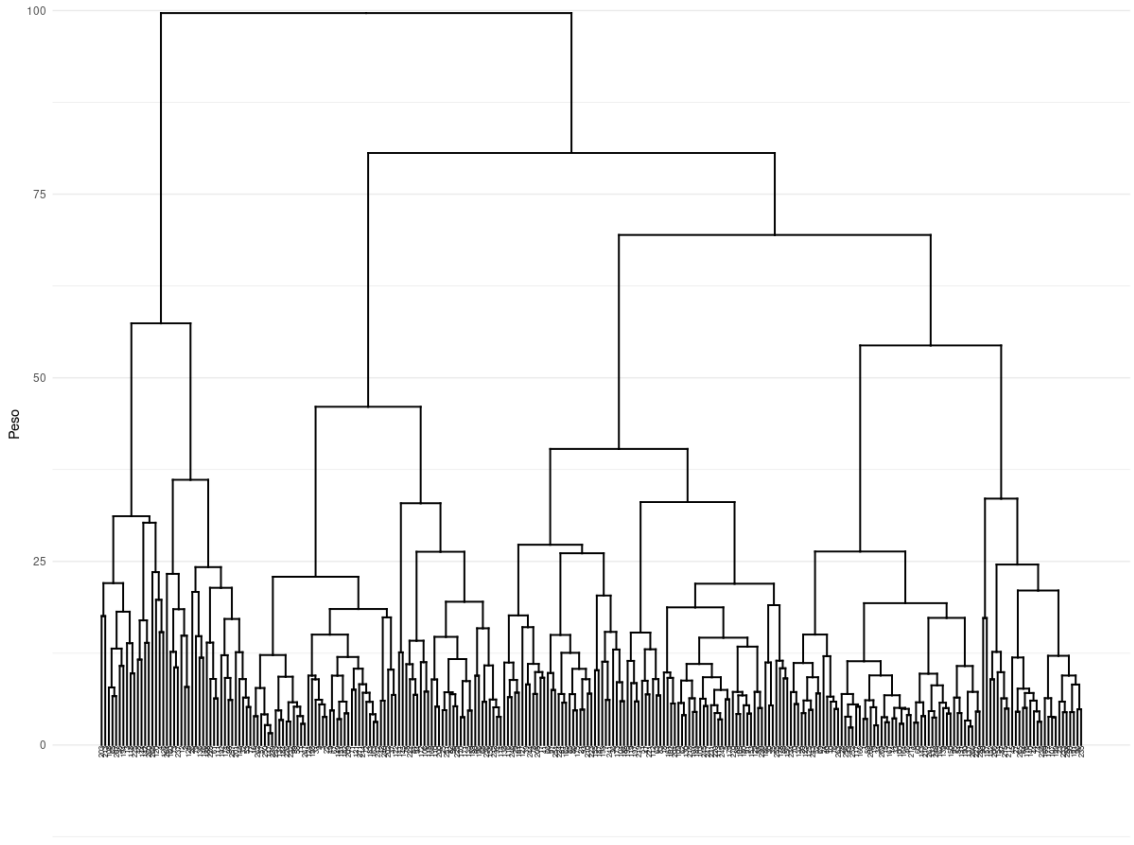

Figura 4.8: Dendrograma pelo agrupamento via método hierárquico Ward.

Isso também pode ser visto na Figura 4.9, gráfico de cotovelo, no qual um número interessante de partições é quatro, em razão de que ao acrescentarmos mais um grupo não temos uma queda significativa na soma de quadrado total dos grupos.

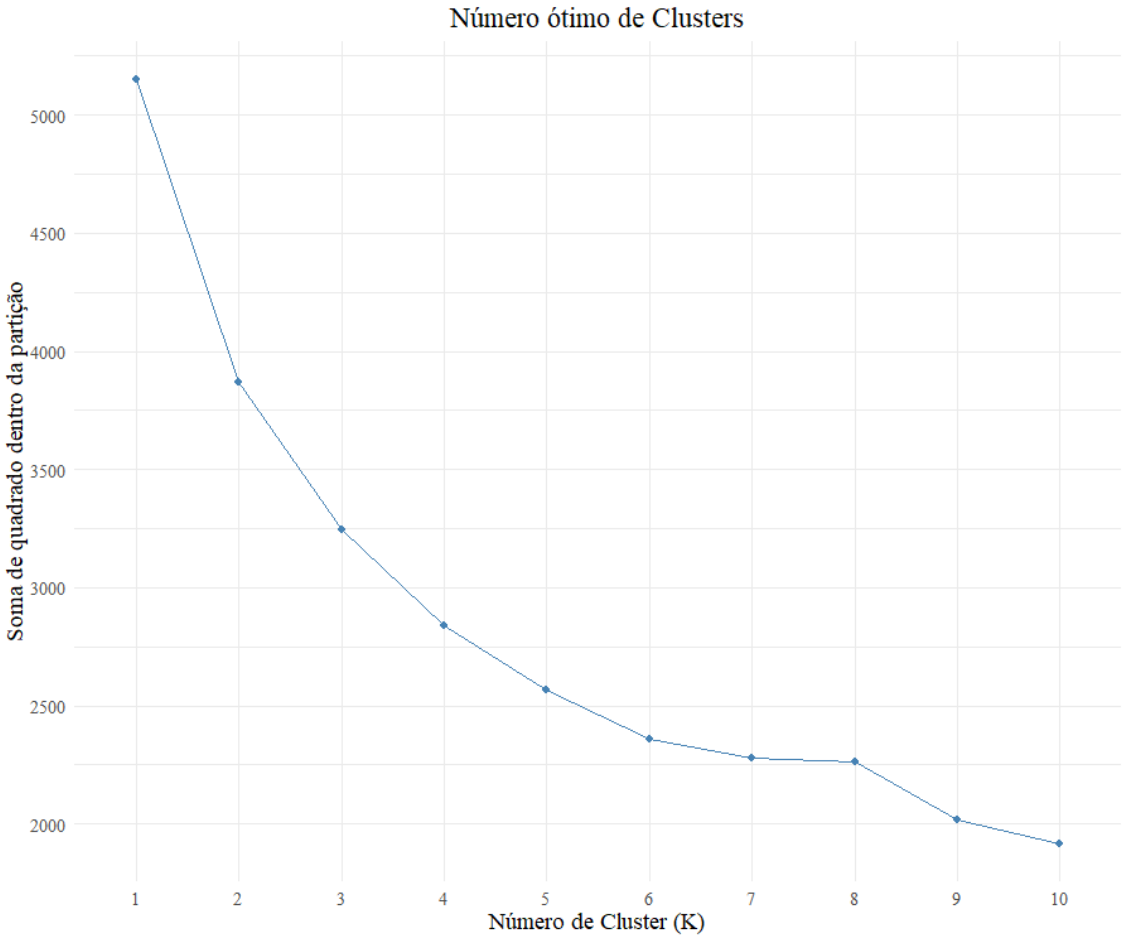

Figura 4.9: Gráfico para a validação da escolha de grupos.

Então, efetuamos o corte para quatro grupos, obtivemos as seguintes quantidades de cada uma das partições e a quantidade de jogadoras em cada posição nos grupos, esses resultados podem ser observados nas Tabelas 4.1 e 4.2.

Tabela 4.1: Quantidade de atletas em cada um dos grupos via método Ward.

Grupos	**1**	**2**	**3**	**4**
Quantidade de Jogadoras	81	42	69	79

Podemos perceber, pela Tabela 4.2, que o grupo 1 é composto, na maior parte, por jogadoras da posição de pivô (*pivot*), sendo composto por 45 jogadoras da posição pivô das 81 quem fazem parte desse grupo. Também, notamos que o grupo 3 é formado majoritariamente por jogadoras da posição de ponta (*wing*). No qual é composto por 60 jogadoras da posição ponta, tanto esquerda quanto direita, das 69 atletas que fazem parte

Tabela 4.2: Quantida de atletas por posição em cada um dos grupos via método Ward.

		Posições						
		Back	**Center back**	**Leaf back**	**Left wing**	**Pivot**	**Right back**	**Right wing**
Ward	**1**	5	6	11	4	45	8	2
	2	1	13	14	3	1	7	3
	3	0	1	1	33	2	5	27
	4	0	31	24	1	1	21	1

desse grupo.

Em seguida analisamos os cinco índices, criados pelo ACP, das jogadoras de cada um dos quatro grupos. Os resultados estão ilustrados na Figura 4.10. Os métodos de agrupamento foram utilizados para separar os índices criados na Seção 4.2.2, não utilizamos os métodos para separar as jogadoras. Logo, um bom agrupamento resulta em grupos compostos por jogadoras com valores semelhantes para os cinco índices criados.

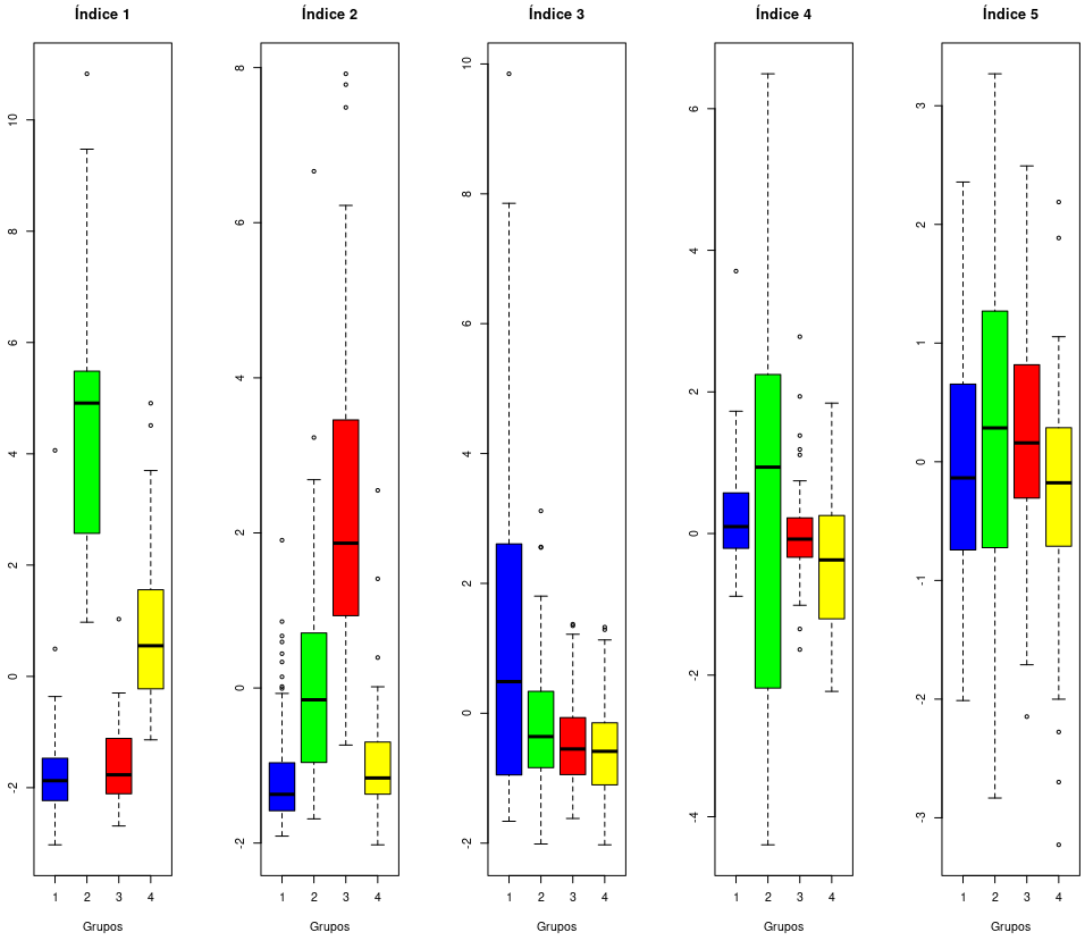

Figura 4.10: Índices dos grupos via método Ward.

Pela Figura 4.10, podemos observar que todos os grupos tem um comportamento semelhante para os índices 3, 4 e 5. As jogadoras dos diferentes grupos apresentam valores próximos para esses índices, mas ressaltamos que o grupo dois tem uma variabilidade

maior, em relação aos outros grupos, para os índices 4 e 5. Além disso, o grupo um tem uma variabilidade maior no índice 3, comparando aos demais grupos.

Notamos, no índice 1, que os grupos dois e quatro são bem diferentes dos grupos um e três. Podemos observar que o grupo dois contém as atletas com maiores valores para o índice de habilidades gerais, seguido do grupo quatro. Também, vemos que os grupos um e três apresentam uma grande interseção, então essas jogadoras possuem valores muito próximos para o índice 1.

Podemos ver que no índice 2 os grupos dois e três são os que mais diferem em relação aos outros grupos. Notamos que o grupo três contém as jogadoras com maiores valores para o índice de habilidade na posição de ponta, seguido pelo grupo dois. Portanto, acreditamos que as atletas do grupo dois são as mais completas, devido a serem as que mais se destacam nos índices 1 e 2.

Posteriormente, realizamos uma análise das variáveis originais das jogadoras, com o intuito de entender o perfil e as características das atletas que compõe cada um dos grupos formados. Esses resultados são ilustrados nas Figuras 4.11, 4.12 e 4.13.

Com a Figura 4.11, podemos observar uma distribuição semelhante para a média de steals, por jogo, em cada um dos grupos. Os grupo três e quatro estão com um comportamento idêntico para a média de gols, shots e steals. Porém, em relação à média de assistências e turnovers, por partida durante o campeonato, os grupos um e três são os com o comportamento mais semelhante e com menores valores nessas características, quando comparado com os outros grupos.

Além disso, notamos que os grupos um e dois apresentam ter o mesmo padrão de comportamento, em todas as variáveis. O grupo um tem as menores médias de quantidade de gols, shots, assistências, turnovers e steals, por jogo. Já o grupo dois tem as maiores médias por jogo, para todas as variáveis supracitadas.

Quando observamos as variáveis relacionadas à punição que as jogadoras podem provocar ou sofrer, Figura 4.12, aquelas alocadas no grupo dois tem uma maior quantidade de 2 minutos e 7 metros sofridos por jogo. Cremos que seja coerente esse comportamento, devido a acreditarmos que essas são as atletas que mais se destacaram nas partidas, como foi supramencionado. Logo, são as jogadoras mais requisitadas durante o jogo, visto que essas são as que possuem maior quantidade de gols, em diferentes posições em quadra, como mostramos na Figura 4.13. Essa característica, acreditamos que, as tornam atletas mais visadas, por consequência mais marcadas, possibilitando sofrerem mais penalidades.

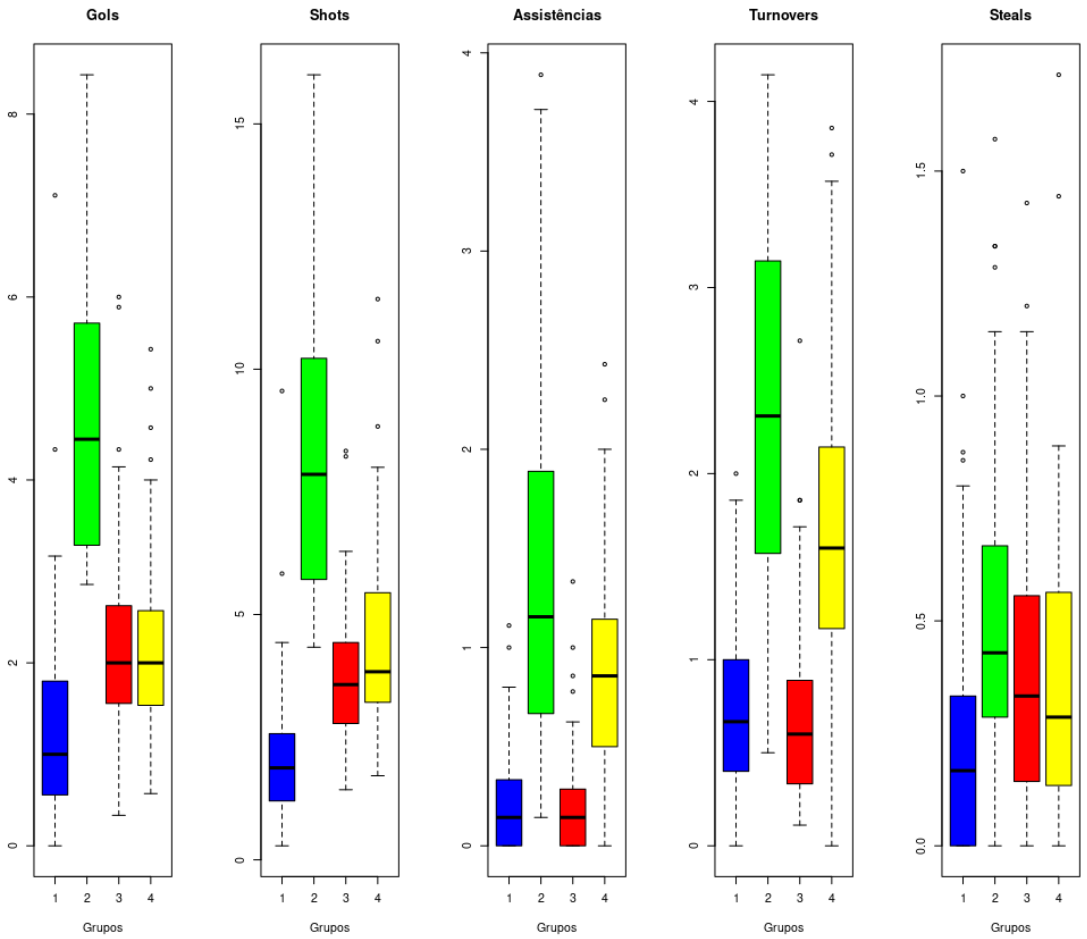

Figura 4.11: Algumas características dos grupos via método Ward.

Também, notamos que para a variável 2 minutos provocados ($p2m$), os grupos um e dois, tem uma distribuição semelhante e os maiores valores, ou seja, as jogadoras de ambos os grupos tem quantidades muito parecidas e maiores valores que os outros grupos para a quantidade de vezes que a atleta ficou dois minutos fora da quadra, resultados das penalidades provocadas de 2 minutos, por jogo.

Além disso, observamos que o grupo um é formado por jogadoras que também tem as maiores quantidades de punições de 7 metros provocados, logo são as atletas que mais realizaram faltas que acarretaram na punição de 7 metros. Essas características são coerentes com o que foi visto, anteriormente, sobre o índice de agressividade em relação às jogadoras da posição pivô, dado que atletas que jogam na posição de pivô são mais agressivas, ou seja, provocam mais punições de 2 minutos e 7 metros. Isso explica os comportamentos do grupo um que, como já foi dito, é formado majoritariamente por jogadoras pivô, como mostra a Tabela 4.2.

Observando os grupos três e quatro, podemos notar um comportamento muito se-

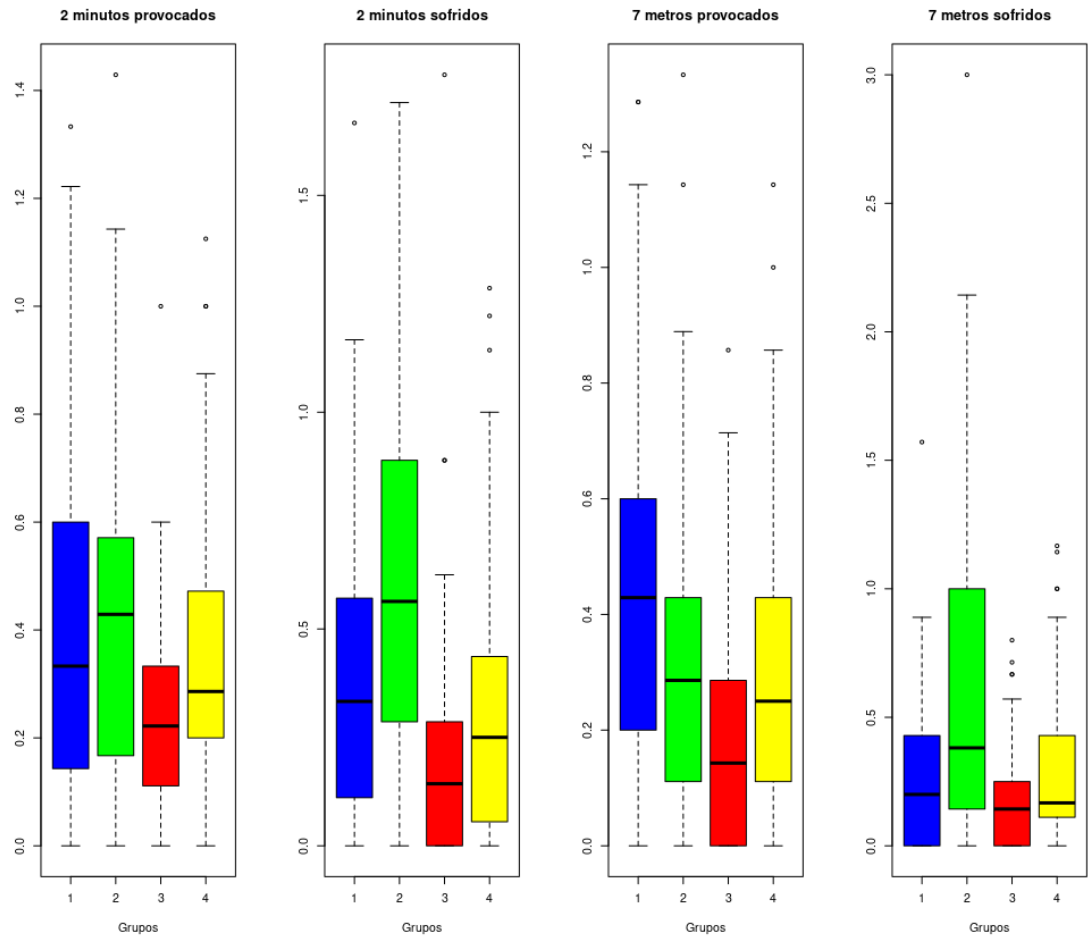

Figura 4.12: Variáveis relacionadas a punição dos grupos via método Ward.

melhante da distribuição de todas as variáveis relacionadas à punição. Em que o grupo 3 apresenta as menores quantidades de punições sofridas e provocadas por partida, seguido do grupo quatro. Portanto, são as jogadoras que menos provocaram e sofreram punições durante o campeonato.

Finalizamos observando algumas variáveis secundárias de cada jogadora: tipos de gols e os respectivos tipos de arremessos. Na Figura 4.13, representamos somente as variáveis *gols*, dos diferentes tipos, devido a elas serem muito correlacionadas com as respectivas variáveis de média de arremessos por jogo.

Podemos observar, pela Figura 4.13, que o grupo dois é composto por jogadoras que possuem vantagens em relação às demais, tanto em gols de 7 metros, gols de 9 metros e gols *breakthrough*, seguido pelas jogadoras do grupo quatro. Nos quesitos gols *wing*, gols de 6 metros e gols de *fastbreak*, podemos notar que o grupo dois é o segundo maior grupo em quantidades de gols, por jogo. Quanto ao número de gols de 6 metros o grupo dois é inferior ao grupo um, visto que o grupo um tem o maior destaque para gols de 6 metros.

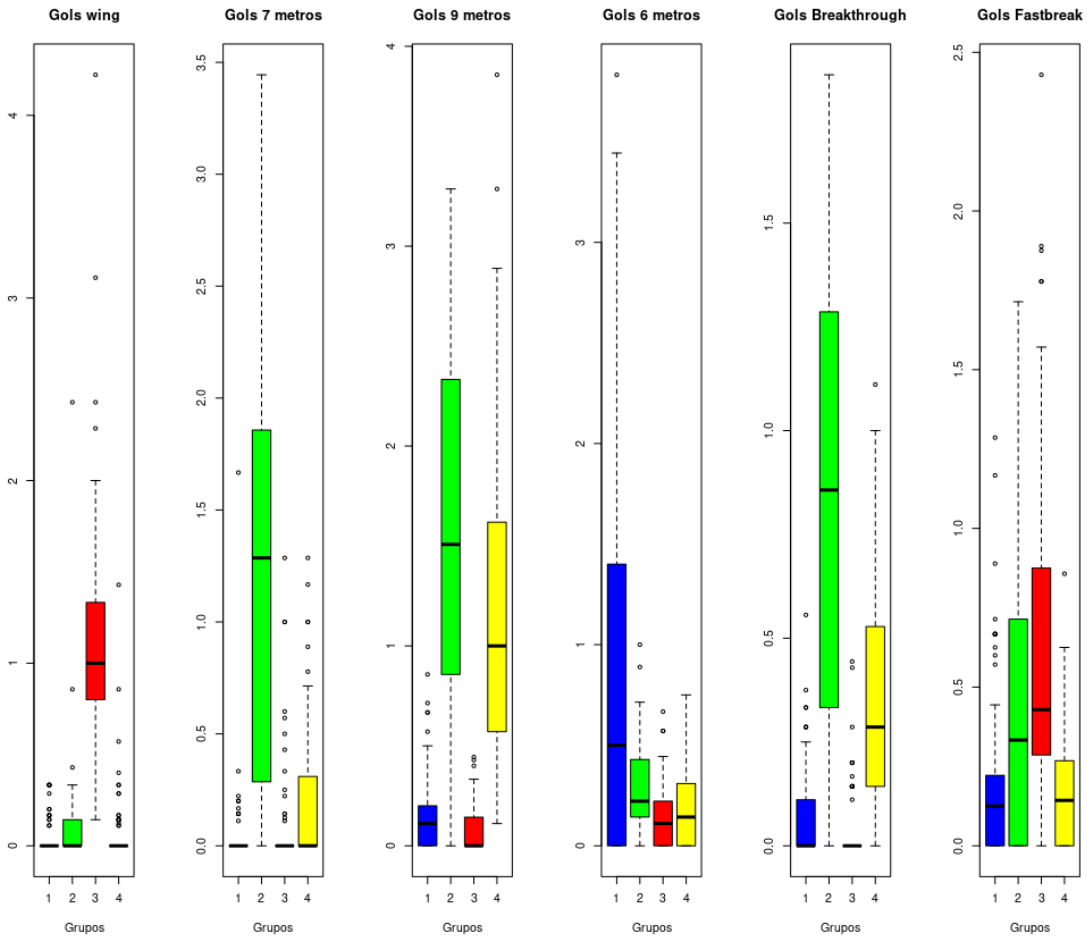

Figura 4.13: Variáveis relacionadas à média de gols, de diferentes áreas da quadra, dos grupos via método Ward.

Enquanto que para gols de *wing* e *fastbreak* as atletas do grupo três apresentam um maior destaque, logo o grupo três aloca jogadoras que por partida marcaram mais gols de *wing* e *fastbreak* que as demais atletas durante todo o campeonato.

Essas características podem ser explicadas pelas posições nas quais as atletas que compõem, majoritariamente, os grupos um e três atuam. Dado que o grupo um é em grande parte formado por jogadoras da posição pivô, como mostra a Tabela 4.2, é coerente elas terem uma quantidade maior de gols de 6 metros, por jogo, já que a linha dos 6 metros é a área que o pivô mais circula.

Observando o grupo três, acreditamos que faz sentido que esse seja o grupo com maior quantidade de gols *wing* e *fastbreak*, por jogo, visto que é majoritariamente composto por atletas que jogam na posição de ponta (direita ou esquerda). Comumente, as jogadoras de ponta são as atletas mais utilizadas em jogadas rápidas (*fastbreaks*).

Visto os resultados supracitados e sabendo que o grupo dois é formado por jogadoras

de todas as posições, acreditamos que as atletas do grupo dois sejam as mais completas em relação a habilidades e eficiência, ou seja, tenham uma maior fluidez posicional de jogo. Em consequência de serem jogadoras bem pontuadas nas áreas comuns de sua posição e também se destacarem nas demais áreas da quadra, as jogadoras do grupo dois realizaram uma quantidade elevada de gols dos tipos característicos de suas posições como também dos tipos característicos de outras posições que não a posição habitual que jogam.

4.3.2 Agrupamento Não-hierárquico

Para ralizarmos o agrupamento via algoritmo K-Médias, utilizamos a mesma quantidade de partições do método Ward. As médias dos grupos formados pelo método hierárquico foram usadas como centroides iniciais para o método não-hierárquico. Os resultados são apresentados nas Tabelas 4.3 e 4.4 e nas Figuras 4.14, 4.15, 4.16 e 4.17.

Tabela 4.3: Quantidade de atletas em cada um dos grupos via algoritmo K-Médias.

Grupos	**1**	**2**	**3**	**4**
Quantidade de Jogadoras	86	31	67	87

A Tabela 4.3 nos exibe que a maioria dos grupos tem tamanhos próximos. Podemos perceber, pela Tabela 4.4, que continuam com uma configuração muito semelhante aos grupos obtidos pelo método Ward. O grupo um continuou sendo composto majoritariamente por jogadoras pivô e o grupo três é formado na maior parte por atletas das posições de ponta (*left wing* e *right wing*).

Tabela 4.4: Quantidade de atletas por posição em cada um dos grupos via algoritmo K-Médias.

		Posições						
		Back	**Center back**	**Leaf back**	**Left wing**	**Pivot**	**Right back**	**Right wing**
K-Médias	**1**	4	7	12	7	44	8	4
	2	0	11	11	0	2	5	2
	3	0	2	1	32	2	5	25
	4	2	31	26	2	1	23	2

Para esse agrupamento, observamos através da Figura 4.14 que os índices dos quatro grupos foram muito semelhantes aos do agrupamento hierárquico. Continuamos a acreditar que o grupo dois contém as atletas mais completas.

Consequentemente o grupo dois tem uma quantidade maior de gols, *shots*, assistência, *turnovers* e *steals*, por partida, observados na Figura 4.15. Podemos notar, pela Figura 4.16, que o mesmo grupo sofre a maior quantidade de penalidades por jogo.

Além disso, é sempre um dos grupos que tem a maior vantagem, em relação aos demais, em todos os tipos de gols, sejam eles gols de 7 metros, 9 metros, *breakthrough* (com os maiores valores), e os segundos maiores valores para gols de *wing*, 6 metros e *fastbreak*. Tais informações podem ser visualizadas na Figura 4.17.

As análises anteriores são relavantes para conhecermos o perfil das jogadoras que compõem cada um dos grupos, proporcionados pelo método não-hierárquico. Como já foi dito obtemos resultados muito parecidos com os grupos formados pelo método Ward.

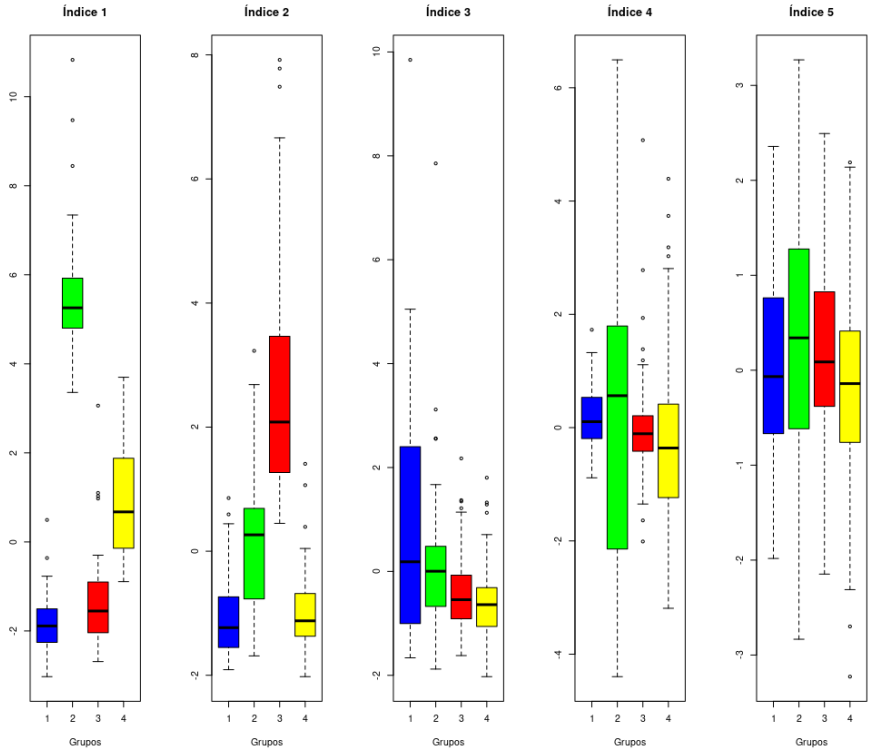

Figura 4.14: Índices dos grupos via algoritmo K-Médias.

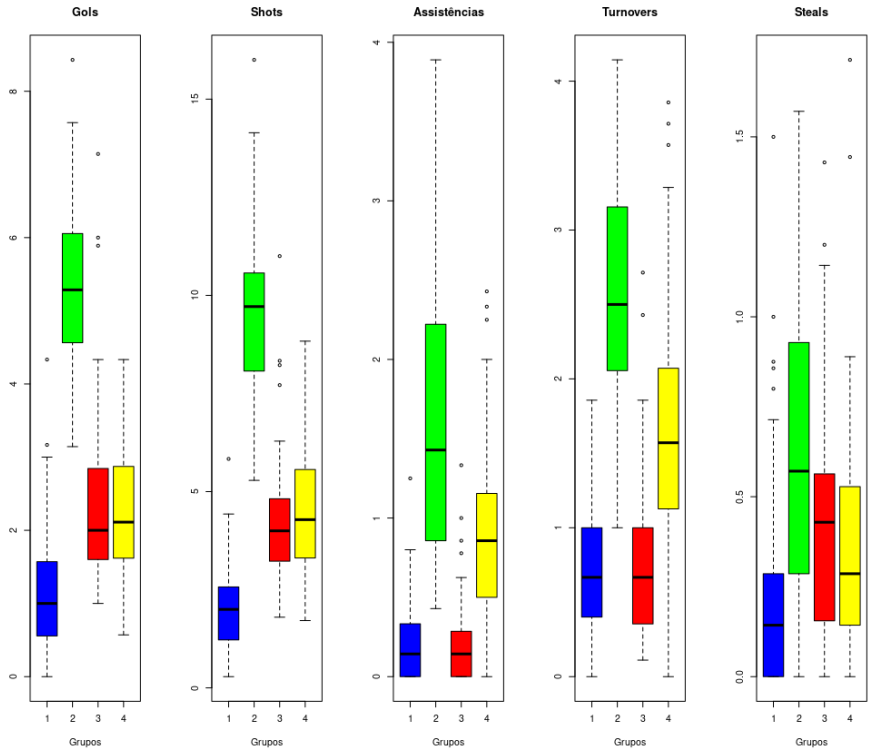

Figura 4.15: Algumas características dos grupos via algoritmo K-Médias.

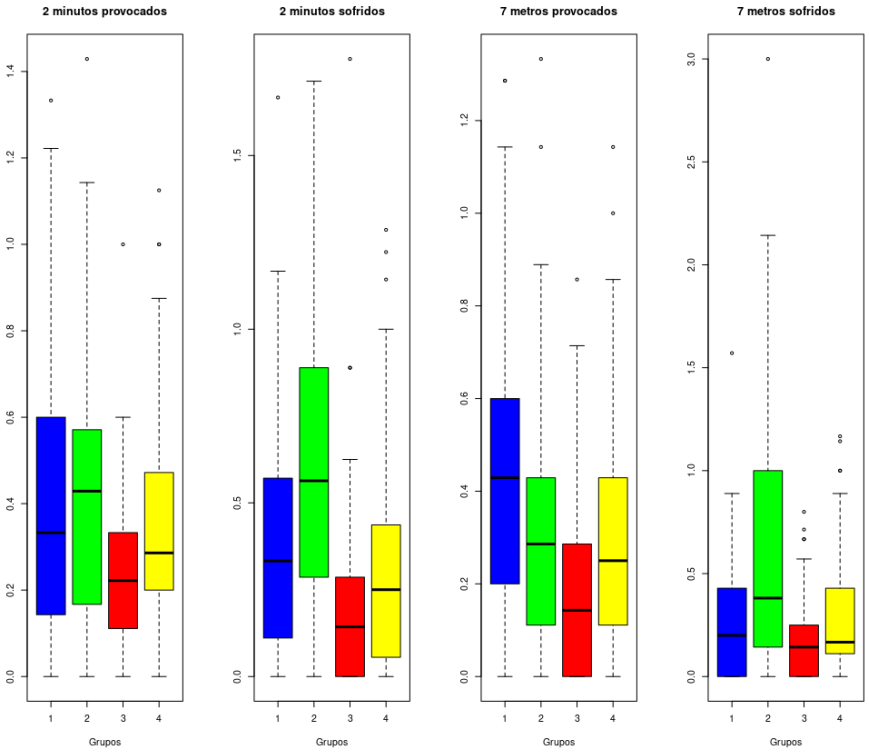

Figura 4.16: Variáveis relacionadas a punição dos grupos via algoritmo K-Médias.

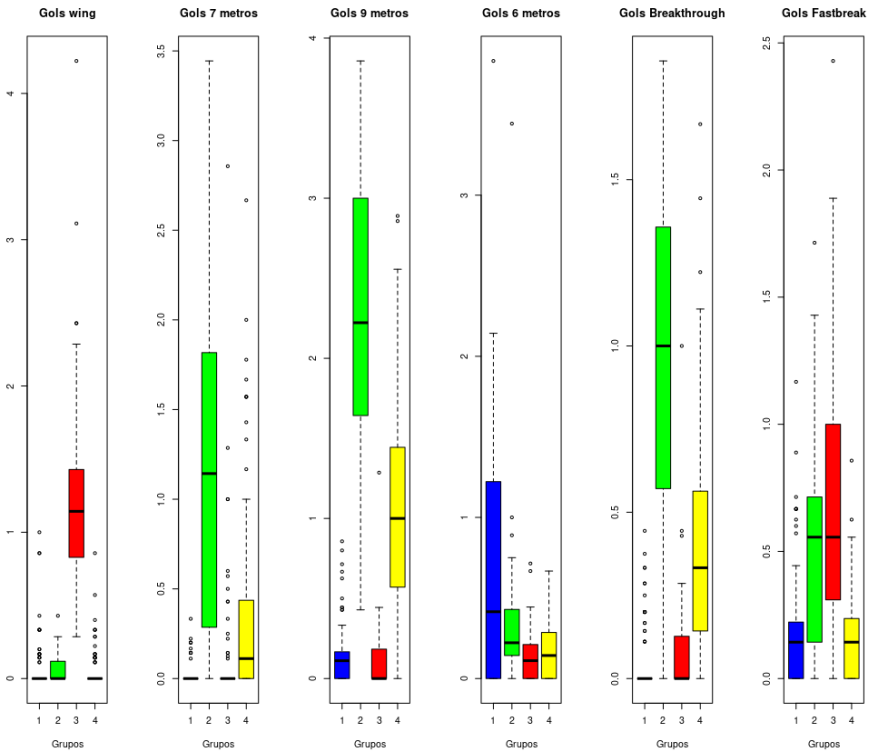

Figura 4.17: Variáveis relacionadas a média de gols, de diferentes aréas da quadra, dos grupos via algoritmo K-Médias.

4.3.3 Comparando os resultados dos Métodos

Por fim, analisamos a concordância entre os métodos utilizados, Ward e K-Médias, observando a proporção de jogadoras que foram designadas para os mesmos grupos nos diferentes procedimentos. Pela Tabela 4.5, podemos notar que os métodos se assemelham suficientemente bem, dado que os agrupamentos via método Ward e algoritmo K-Médias concordam em 88.19% na atribuição das jogadoras para os quatro grupos.

Tabela 4.5: Comparação dos agrupamentos via método Ward e K-Médias.

		Ward			
		1	**2**	**3**	**4**
K**-Médias**	**1**	77	0	6	3
	2	1	27	0	3
	3	1	2	63	1
	4	2	13	0	72

Capítulo 5

Considerações Finais

Exploramos, nesse trabalho de graduação, as metodologias de análise de componentes principais e de agrupamento, aplicamos os métodos no conjunto de dados das jogadoras que participaram do Mundial Feminino Juvenil de 2018 que foi sediado na Polônia. Nosso principal objetivo é definir grupos de jogadoras com desempenhos semelhantes, tendo como base os cinco primeiros componentes principais obtidos pela ACP realizada com as variáveis do banco de dados que foram consideradas mais informativas. Utilizamos métodos de agrupamento hierárquico e não-hierárquico para alcançarmos tal objetivo.

Da perspectiva teórica, a aplicação desses métodos viabilizou salientar as vantagens e desvantagens de cada procedimento, como ainda observar o grau de concordância entre eles. Pretendíamos realizar um índice composto, porém como os índices obtidos pelo método de componentes principais já dependiam de diversas variáveis optamos por não realizá-lo. Então, efetuamos somente a caracterização dos quatro grupos que foram obtidos, essa quantidade de grupos foi escolhida e verificada através do dendrograma e do gráfico de cotovelo.

Como resultado prático, testamos as medidas de distância e podemos concluir que a melhor medida de distância, para esses dados, foi a distância de *city blocks*, devido a presença de pontos extremos. E o método hierárquico e o algoritmo de K-Médias tiveram resultados muito parecidos, com uma concordência de respostas maior que 88%. O método Ward, foi o método que resultou os grupos mais homogêneos de forma e tamanhos semelhantes.

Exploramos mais detalhadamente o método Ward, obtivemos que o grupo dois é composto pelas jogadoras mais completas, sendo atletas que mais se destacaram em características comuns como: assistências, *turnovers* e *steals*. Essas atletas também são as

que tiveram uma melhor fluidez posicional de jogo, tendo um bom desempenho em características da sua posição habitual e em habilidades tipicamente caracteristicas de outras posições, sendo as que mais se destacam nos índices 1 (habilidades gerais das jogadoras) e 2 (habilidades na posição de ponta). Portanto, as jogadoras do grupo dois são as atletas, desse campeonato, que tem as características mais buscadas pelos treinadores em atletas com potencial profissional. No Apêndice B, mostramos quem são e algumas informações dessas jogadoras alocadas no grupo dois.

Caso haja alguém que se interesse em realizar análises semelhantes ou melhores que as deste trabalho, pode buscar usar uma variável relacionada ao tempo que cada jogadora ficou em quadra (essa informação só não foi utilizada por não constar no banco de dados) ou impor um número mínimo de gols realizados para cada tipo de gol, talvez obtenha informações mais interessantes. No lugar de dividir pelo número de jogos que a atleta participou, optaríamos por dividir pelo tempo em quadra que a jogadora teve durante todo o campeonato e filtraríamos somente as atletas que tem um valor mínimo de gols para cada tipo de gol.

Apêndice A

Tabela com os dados da Angola

A seguir veremos uma amostra do banco que contem observações das jogadoras da seleção Angolana.

Tabela A.1: Banco de dados.

Name	Position	DOB	Age	Club	(cm)	(kg)	IM	IG	G	P	Goals	Av.g	Shots	$Wing_g$	$Wing_s$	$7m_g$	$7m_s$	$9m_g$	$9m_s$	$6m_g$	$6m_s$	$Break_g$	$Break_s$	$Fastbreak_g$	$Fastbreak_s$	Assists	Turnovers	Steals	$7m_comm.$	$7m_ec.$	$2m_gun.$	$2m_ec.$	Yellow	Red
FREDERICO Paula Francisco	Goalkeeper	30.01.2000	18	Clube Desportivo 1° De Agosto	176	69	7		7	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SUCACUEXE Catarina Paulo S	Left wing	27.05.2002	16	Clube Desportivo Marinha Guerra	173	61			7	3	0	0	2	0	2	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	1	0	0	0	0
ESTEVES Edna Kiala	Right back	24.05.2000	18	Clube Desportivo Marinha Guerra	174	68			7	7	2	0,285	9	0	0	0	0	2	9	0	0	0	0	0	0	4	3	2	3	1	4	0	0	0	0
KIALA Luzia Santana	Right back	07.01.2000	18	Clube Desportivo Marinha Guerra	176	64	7	57	7	7	39	5,571	72	2	4	8	9	22	50	0	0	5	6	2	3	3	17	7	5	4	5	7	1	0
MONTEIRO Jelsia Marlia Miguel	Left wing	16.02.2001	17	Clube Desportivo 1° De Agosto	175	79	7		7	7	11	1,571	25	9	19	0	0	1	2	0	0	0	0	1	4	1	4	3	1	0	5	0	0	0
MASSEU Beatriz Diana	Center back	04.04.2000	18	Clube Desportivo 1° De Agosto	173	58	7	15	7	7	50	7,142	99	0	0	9	12	23	60	1	4	12	14	5	9	4	20	4	5	5	0	9	1	0
CASTRO Thayany Cacaica B	Pivot	16.11.2002	15	Nucleo Do Clube 1° De Agosto Benguela	175	89			7	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
DA SILVA Paulina Bazolau	Goalkeeper	14.05.2000	18	Clube Desportivo 1° De Agosto	182	55	7		7	7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0
TCHITONGUA Isabel Marcelino	Center back	27.08.2001	16	Nucleo Do Clube 1° De Agosto Benguela	169	60			7	6	9	1,285	17	0	0	0	0	7	13	0	0	1	1	1	3	2	10	3	0	0	2	1	0	0
COIO Ana Jandira A	Right back	12.10.2001	16	Clube Desportivo Marinha Guerra	177	73			7	6	5	0,714	7	0	0	0	0	0	1	4	4	0	0	1	2	1	2	2	1	2	4	1	2	0
GABRIEL Chelsia Cristina	Center back	24.09.2000	17	Clube Desportivo 1° De Agosto	170	56	7	13	7	7	12	1,714	24	0	0	1	1	6	15	1	1	3	5	1	2	7	17	1	1	8	1	7	0	0
SEBASTIAO Zuraida Agnela M.	Pivot	01.04.2001	17	Clube Desportivo 1° De Agosto	177	79			7	7	4	0,571	5	0	0	1	1	1	1	0	1	0	0	2	2	1	4	1	0	2	5	0	1	0
FERREIRA Ligia Alexandra J	Goalkeeper	16.04.2002	16	Nucleo Do Clube 1° De Agosto Benguela	178	77	7		7	7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ROSARIO Dolores Palmira L	Left wing	01.04.2003	15	Clube Desportivo 1° De Agosto	174	69			7	7	9	1,285	26	7	20	0	1	0	0	0	0	0	0	2	5	0	2	5	2	0	1	0	1	0
MATEUS Roseana Sebastino	Left back	11.09.2000	17	Clube Desportivo Marinha Guerra	183	69	7		7	7	7	1	20	0	0	0	0	6	19	0	0	0	0	1	1	1	11	1	4	0	6	0	2	0

Apêndice B

Tabela das jogadoras do grupo dois pelo método Ward

A seguir veremos algumas informações das jogadoras que, pelo método Ward, compoem o segundo grupo que foi formado.

Tabela B.1: Atletas do grupo dois, pelo método Ward.

Name	País	Position	Goals	Índice 1	Índice 2	Índice 3	Índice 4	Índice 5
AOUIJ Fadwa	Tunísia	Center back	8.429	10.83	2.69	0.68	1.48	-2.83
ZHOU Mengxue	China	Center back	7.571	9.47	0.47	-1.25	2.1	-0.04
MASSEU Beatriz Diana	Angola	Center back	7.143	8.44	1.14	0.07	-0.56	1.15
TREPÁCOVÁ Zuzana	Eslováquia	Right back	6.143	7.34	0.33	0.08	-1.81	1.17
BONO Carolina	Argentina	Center back	6	7.15	-0.96	0.29	0.99	2.46
ANDERSSON Isabelle	Suécia	Left back	6.111	6.82	0.83	2.56	-2.25	-0.96
OKADA Ayame	Japão	Left back	5.714	6.1	3.23	1.67	-3.01	1.11
KRULLAARS Nyala	Países Baixos	Left back	6	6.06	1.62	0.28	0.88	-1.92
STEPANOVA Kristina	Cazaquistão	Right back	6.286	5.78	2.58	0.44	-0.49	-0.49
MÉSZÁROS Réka	Eslováquia	Center back	5.286	5.69	-0.27	-0.51	2.36	-0.58
ZHAPAROVA Aida	Cazaquistão	Right wing	5.286	5.48	-0.38	0.53	0.55	-2.29
KIALA Luzia Santana	Angola	Right back	5.571	5.45	0.44	-0.5	0.56	-0.81
ZALFANI Nada	Tunísia	Right back	4.714	5.44	0.54	2.55	-2.18	3.02
WOO Bitna	Coreia	Left back	6.333	5.38	0.36	-1.75	6.49	0.18
OH Yedam	Coreia	Center back	4.556	5.27	1.1	3.12	-4.16	-0.34
ABE Miyuki	Japão	Right wing	4.571	5.26	0.26	0.34	-2.1	0.06
STEFFENSEN Emilie Bodholdt	Dinamarca	Left back	5.125	5.07	-0.61	-0.56	2.52	1.43
MALEC Ana	Croacia	Center back	4	5.03	-0.39	1.13	-4.4	-0.66
SOBREPERA CASOL Janna	Espanha	Left back	4.889	4.95	0.71	-1.13	2.31	-0.72
SEPULVEDA Valentina	Chile	Pivot	3.857	4.95	-1.01	-0.71	-1.89	1.39
PÁL Tamara	Hungria	Left back	5.333	4.94	0.35	-1.23	2.25	0.34
MIKHAYLICHENKO Elena	Russia	Center back	4.889	4.88	0.15	0.37	-2.76	1.62
POPA Andreea Cristina	Romênia	Center back	3.143	4.73	-1.28	-0.16	-0.64	1.24
MOHAMED RADY AHMED Rana	Egito	Left back	4.286	4.68	-1.24	-1.05	1.74	-1.94
SOBECKA Lucyna	Polônia	Right back	4.143	3.58	-0.44	-0.17	-2.27	0.52
BALACEANU Ioana Beatrice	Romênia	Left back	3.857	3.57	-1.09	-0.63	1.14	2.37
TARSOAGA Andreea Rebeca	Romênia	Left back	3.286	3.49	-1.69	0	-3.1	3.27
DANO Nina	Suécia	Right back	3.889	3.1	-0.34	-0.64	2.81	-1.12
NEIDHART Nina	Austria	Left wing	7.143	3.06	6.66	-1.11	5.07	1.39
WIERZBA Anna Berger	Dinamarca	Center back	2.889	2.84	-0.8	-0.52	-3.19	2.14
PANDZA Katarina	Austria	Back	3.286	2.7	-0.87	-1.13	1.36	1.27
PARK Soyoun	Coreia	Right back	3.778	2.57	0.04	1.8	-2.41	-0.96
VON PEREIRA Aimée	Alemanha	Left back	4.333	2.52	-0.37	-0.85	3.03	-0.13
VÁMOS Petra	Hungria	Center back	3	2.4	-1.64	-1.02	-2.3	2.02
MAMDOU MOHAMED MAHMOUD Nada	Egito	Center back	3	2.36	-1	-0.84	2.18	0.58
PETIKA Tena	Croacia	Left back	2.857	2.09	-1.13	-0.7	1.44	0.11
SVELE Mia	Noruega	Center back	3.5	1.92	-0.03	-0.06	2.01	-1.08
BOZOVIC Marija	Montenegro	Left wing	3.286	1.9	-1.05	-0.4	1.93	1.16
CYGAN Katarzyna	Polônia	Left back	3	1.56	-1.46	-0.6	3.18	-2.31
CHOI Gyeongbin	Coreia	Left back	3.167	1.33	-0.73	-2.01	3.74	0.6
PEILLON Melina	França	Right wing	3.667	1.14	1.06	-0.32	4.39	0.23
CAVO Guadalupe	Argentina	Left wing	3.286	0.97	2.48	-0.08	-2.01	1.05

Referências Bibliográficas

Barroso, L. P. e Artes, R. (2003). Análise multivariada. *Lavras: UFLA*, page 151.

Commission, J. R. C.-E. *et al.* (2008). *Handbook on constructing composite indicators: methodology and user guide*. OECD publishing.

Donatelli, A. C. (2017). *Agrupamento dos Atacantes do Campeonato Inglês Utilizando Indicadores de Performance*. Trabalho de Conclusão de Curso (Graduação) - Departamento de Estatística, Universidade Federal de São Carlos, São Carlos, 2017.

Greco, P. J. e Romero, J. J. F. (2011). *Manual de handebol: da iniciação ao alto nível*. Phorte Editora LTDA.

Hair, J. F., Black, W. C., Babin, B. J., Anderson, R. E. e Tatham, R. L. (2009). *Análise multivariada de dados*. Bookman editora.

Härdle, W. K. e Simar, L. (2015). *Applied multivariate statistical analysis*. Springer-Verlag Berlin Heidelberg.

IHF (2016). Regulations documents. Disponível em: < `https://www.ihf.info/regulations-documents/361?selected=Rules\%20of\%20the\%20Game` >. Acessado em: 4 abr. 2021.

IHF (2018). 2018 women's youth (u18) world championship. Disponível em: < `https://archive.ihf.info/en-us/ihfcompetitions/worldchampionships/womensyouthworldchampionships/2018womensyouth(u18)worldchampionship/teaminfo.aspx` >. Acessado em: 3 feb. 2021.

Johnson, R. A., Wichern, D. W. *et al.* (2002). *Applied multivariate statistical analysis*, volume 5. Prentice hall Upper Saddle River, NJ.

Lattin, J., Carroll, J. D. e Green, P. E. (2011). Análise de dados multivariados. *São Paulo: Cengage Learning*, **475**.

MacQueen, J. *et al.* (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA.

Menezes, R. P. (2011). *Modelo de análise técnico-tática do jogo de handebol: necessidades, perspectivas e implicações de um modelo de interpretação das situações de jogo em tempo real*. Dissertação (Doutorado em Educação Física) - Faculdade de Educação Física, Universidade Estatual de Campinas, Campinas, 2011.

R Core Team (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Xu, R. e Wunsch, D. (2005). Survey of clustering algorithms. *IEEE Transactions on neural networks*, **16**(3), 645–678.