UNIVERSIDADE FEDERAL DO AMAZONAS - UFAM

INSTITUTO DE COMPUTAÇÃO - ICOMP

BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

Uso de *Cluster* para Formação de Equipes Balanceadas na Prática do Futebol

Marcello Henrique Soares Cipriano

Manaus - AM

2026

Marcello Henrique Soares Cipriano

Uso de *Cluster* para Formação de Equipes Balanceadas na Prática do Futebol

Trabalho de Conclusão de Curso apresentado ao Instituto de Computação da Universidade Federal do Amazonas como requisito parcial para a obtenção do grau de Bacharel em Ciência da Computação.

Orientador

David Braga Fernandes de Oliveira, Prof. Dr.

Universidade Federal do Amazonas - UFAM

Instituto de Computação - IComp

Manaus - AM

2026

À minha esposa, Tarciana, e ao meu filho, Pedro. Agradeço imensamente pelo amor, pela paciência e por compreenderem minhas ausências ao longo desses cinco anos de curso. A vocês dedico a conclusão desta etapa, com todo o meu amor e a promessa de recuperar cada momento.

AGRADECIMENTOS

Um agradecimento especial a toda a equipe de professores do Instituto de Computação (IComp) da Universidade Federal do Amazonas. Mais do que transmitir o conhecimento técnico, vocês nos ensinaram a acreditar em nosso próprio potencial. O compromisso e a dedicação incansável de mestres que construíram suas carreiras na Universidade são a verdadeira engrenagem que transforma a vida de tantos jovens. Obrigado por nos passarem a confiança de que, com a base que recebemos aqui, estamos prontos para competir e brilhar nos mais altos níveis do mercado global de tecnologia.

“Playing football is very simple, but playing simple football is the hardest thing there is.”

J. Cruyff (frase atribuída em entrevistas e biografias).

Uso de *Cluster* para Formação de Equipes Balanceadas na Prática do Futebol

Autor: Marcello Henrique Soares Cipriano

Orientador: David Braga Fernandes de Oliveira, Prof. Dr.

Resumo

A utilização de dados para apoio à tomada de decisão no esporte tem se consolidado como uma importante frente de pesquisa e aplicação tecnológica. No futebol, o avanço da coleta de métricas de desempenho individual possibilitou o uso de métodos de aprendizado de máquina para análise de atletas e apoio à gestão esportiva. Em contextos informais e amadores, porém, a formação de equipes ainda ocorre de maneira subjetiva, frequentemente gerando partidas desequilibradas. Este trabalho investiga o uso de técnicas de clusterização, com ênfase na clusterização hierárquica com Ward, aplicadas a dados de desempenho de jogadores das cinco principais ligas europeias (temporada 2024/2025) como ferramenta para a formação de equipes balanceadas. A metodologia seguiu um *pipeline* clássico de ciência de dados: coleta e consolidação dos dados; pré-processamento; redução de dimensionalidade via PCA; clusterização; e aplicação dos agrupamentos para composição de equipes. Foram comparados dois algoritmos (K-Means e Hierárquico com Ward), avaliados por meio do coeficiente de Silhueta e do índice de Davies-Bouldin, além de critério interpretativo futebolístico e de aderência às formações medianas observadas em equipes reais. A solução hierárquica de Ward com sete *clusters* demonstrou o melhor equilíbrio entre refinamento funcional, interpretabilidade tática e aplicabilidade prática, identificando perfis como goleiros, zagueiros construtores, volantes organizadores, criadores ofensivos e finalizadores. O sistema proposto distribui jogadores entre equipes considerando tanto a diversidade funcional dos *clusters* quanto o equilíbrio técnico relativo dos atletas, superando estratégias simples de balanceamento por média global. Os resultados indicam que a clusterização constitui uma abordagem promissora e metodologicamente fundamentada para a formação de equipes equilibradas, com aplicabilidade em contextos recreativos, amadores e semiprofissionais.

Palavras-chave: aprendizado de máquina não supervisionado; clusterização hierárquica; *Ward's linkage*; futebol; *sports analytics*; PCA; formação de equipes balanceadas.

Uso de *Cluster* para Formação de Equipes Balanceadas na Prática do Futebol

Autor: Marcello Henrique Soares Cipriano

Orientador: David Braga Fernandes de Oliveira, Prof. Dr.

Abstract

The use of data to support decision-making in sports has established itself as an important research and technological application field. In football, advances in the collection of individual performance metrics have enabled the use of machine learning methods for athlete analysis and sports management support. In informal and amateur contexts, however, team formation still occurs subjectively, often resulting in unbalanced matches. Against this backdrop, this work investigates the use of clustering techniques — with emphasis on hierarchical clustering with the Ward method — applied to player performance data from the five main European leagues (2024/2025 season) as a tool for balanced team formation. The methodology followed a classical data science pipeline: data collection and consolidation; preprocessing; dimensionality reduction via Principal Component Analysis (PCA); clustering; and application of groupings for team composition. Two algorithms were compared — K-Means and Ward — evaluated through the Silhouette coefficient and the Davies-Bouldin index, as well as a football-based interpretive criterion and adherence to median formations observed in real teams. The Ward hierarchical solution with seven clusters demonstrated the best balance between functional refinement, tactical interpretability, and practical applicability, identifying profiles such as goalkeepers, ball-playing defenders, organizing midfielders, creative attackers, and finishers. The proposed system distributes players among teams considering both the functional diversity of clusters and the relative technical balance of athletes, outperforming simple average-based balancing strategies. The results indicate that clustering constitutes a promising and methodologically grounded approach for balanced team formation, with applicability in recreational, amateur, and semi-professional contexts.

Keywords: unsupervised machine learning; hierarchical clustering; Ward's linkage; football; sports analytics; PCA; balanced team formation.

LISTA DE ILUSTRAÇÕES

Figura 1 – Mapa de calor da matriz de correlação das variáveis da base de dados após o pré-processamento. 41
Figura 2 – Curva de Variância Acumulada. 42
Figura 3 – Critério de Kaiser. 43
Figura 4 – Curva de Variância Acumulada com Critério de Kaiser. 43
Figura 5 – Comparação dos índices de validação interna para o algoritmo K-Means, avaliados para k entre 5 e 11 *clusters*. 45
Figura 6 – Comparação dos índices de validação interna para o algoritmo Hierárquico com Ward, avaliados para k entre 5 e 11 *clusters*. 46
Figura 7 – Dendrograma da clusterização hierárquica com método de Ward. . . 48
Figura 8 – Dendrograma do método de *Ward* com projeção visual para o agrupamento com 7 *clusters*. 53
Figura 9 – Distribuição espacial 2D dos agrupamentos da solução Ward $k = 7$. . 54
Figura 10 – Distribuição espacial 2D dos agrupamentos da solução K-Means $k = 7$. 55
Figura 11 – Exemplos de equipes geradas pelo algoritmo de balanceamento proposto. 59

LISTA DE TABELAS

Tabela 1 – Evolução estrutural dos agrupamentos pelo método de ligação de Ward 48
Tabela 2 – Resumo interpretativo dos agrupamentos pelo algoritmo K-Means. . 50
Tabela 3 – Distribuição mediana de jogadores por *cluster* nas equipes reais. . . . 51
Tabela 4 – Nomeação funcional dos *clusters* da solução Ward para $k = 7$. 53
Tabela 5 – Correspondência entre *clusters* (Ward k=7) e posições nominais de jogadores. 54
Tabela 6 – Comparação entre os perfis identificados pelos métodos de Ward e K-Means. 56

SUMÁRIO

1 INTRODUÇÃO 11
1.1 Contextualização 11
1.2 Problematização 12
1.3 Justificativa 13
1.4 Objetivos 14
1.4.1 Objetivo Geral 14
1.4.2 Objetivos Específicos 15
1.5 Organização do Trabalho 16

2 REVISÃO BIBLIOGRÁFICA 17
2.1 Ciência de Dados e Análise no Futebol 17
2.2 Aprendizado de Máquina no Contexto Esportivo 18
2.3 Clusterização e Aprendizado Não Supervisionado 19
2.4 Clusterização de Jogadores no Futebol 22
2.5 Formação de Equipes e Equilíbrio Competitivo 25
2.6 Lacuna na Literatura e Contribuição do Trabalho 27

3 METODOLOGIA 28
3.1 Caracterização da Pesquisa 28
3.2 Coleta e Construção da Base de Dados 28
3.2.1 Fontes de Dados 28
3.2.2 Extração dos Dados 29
3.2.3 Estrutura dos Dados 30
3.3 Pré-processamento e Tratamento dos Dados 30
3.3.1 Tratamento de Valores Ausentes 30
3.3.2 Remoção de Registros Duplicados 31
3.3.3 Normalização por 90 Minutos 31
3.3.4 Seleção e Exclusão de Variáveis 32
3.3.5 Filtragem por Minutos Jogados 32
3.3.6 Padronização dos Dados 33

3.4 Redução de Dimensionalidade . . . 33
3.5 Clusterização dos Jogadores . . . 34
3.5.1 Algoritmos Utilizados . . . 34
3.5.2 Definição do Número de Clusters . . . 35
3.5.3 Métricas de Validação . . . 36
3.5.4 Análise Interpretativa dos Clusters . . . 37
3.6 Formação de Equipes Balanceadas . . . 37
3.6.1 Definição da Composição Estrutural das Equipes . . . 38
3.6.2 Algoritmo de Distribuição dos Jogadores . . . 38

4 RESULTADOS E DISCUSSÃO . . . 40
4.1 Construção e Caracterização da Base de Dados . . . 40
4.2 Redução de Dimensionalidade via PCA . . . 42
4.3 Avaliação dos Agrupamentos: Métricas de Validação . . . 44
4.4 Análise Interpretativa dos Agrupamentos . . . 47
4.4.1 Agrupamentos pelo Método Hierárquico com Ligação de Ward . 47
4.4.2 Agrupamentos pelo Algoritmo K-Means . . . 50
4.5 Análise das Formações Medianas e Definição do Modelo Final de Referência . . . 51
4.6 Perfis Funcionais do Modelo de Referência (Ward k=7) . . . 52
4.7 Comparação entre as soluções dos algoritmos Hierárquico com Ward e K-Means . . . 54
4.8 Justificativas da Escolha do Método de Referência e Considerações para Implementação . . . 57
4.9 Aplicação dos Agrupamentos na Formação de Equipes Balanceadas . . . 58

5 DESENVOLVIMENTO DA APLICAÇÃO . . . 60
5.1 Objetivos e Escopo da Aplicação . . . 60
5.2 Tecnologias e Arquitetura . . . 61
5.3 Módulos Funcionais . . . 62
5.4 Relação entre a Aplicação e a Pesquisa . . . 63
5.5 Disponibilização e Reprodutibilidade . . . 63

6 CONCLUSÃO . . . 64

Referências . . . 67

1

INTRODUÇÃO

1.1 Contextualização

A transformação digital tem impactado de maneira profunda os mais diversos setores da sociedade, e o esporte não constitui exceção. Nas últimas décadas, a crescente disponibilidade de dados e o avanço das técnicas de análise computacional abriram caminho para o surgimento de uma disciplina específica dedicada à aplicação de métodos quantitativos no contexto esportivo: a *sports analytics*.

Essa área, que combina ciência de dados, estatística e aprendizado de máquina, tem sido utilizada para apoiar decisões estratégicas, táticas e de gestão em diferentes modalidades, transformando a forma como atletas, equipes e competições são analisados e gerenciados.

No futebol, esporte mais popular do mundo e com um dos ecossistemas de dados mais ricos do esporte moderno, essa revolução analítica tem sido especialmente expressiva. A coleta sistemática de estatísticas individuais e coletivas durante partidas das principais ligas de futebol mundial possibilitou a criação de bases de dados ricas em atributos, que descrevem desde ações técnicas elementares, como passes e finalizações, até métricas avançadas derivadas de modelos probabilísticos, como o *expected goals* (xG) e o *expected assists* (xA).

Esses dados passaram a alimentar modelos preditivos, sistemas de recrutamento de talentos, análises táticas automatizadas e ferramentas de avaliação de desempenho, revolucionando práticas antes baseadas exclusivamente na observação empírica e no julgamento subjetivo de especialistas.

Um marco simbólico nessa transformação foi documentado por Lewis (2003) na obra *Moneyball*, que narra como a franquia Oakland Athletics, do beisebol norte-americano, utilizou análise estatística para identificar jogadores subestimados pelo mercado e construir uma equipe competitiva com orçamento reduzido.

O paradigma inaugurado pelo *Moneyball*, de que dados bem analisados podem revelar padrões não evidentes à observação e gerar vantagem competitiva, rapidamente extrapolou o beisebol e passou a influenciar o futebol e outras modalidades em todo o mundo, consolidando a *sports analytics* como campo de pesquisa e aplicação de grande relevância acadêmica e prática.

1.2 Problematização

Apesar dos avanços no contexto do futebol profissional, a maior parte das ferramentas e métodos desenvolvidos pela *sports analytics* destina-se a organizações com acesso a recursos tecnológicos e financeiros significativos, como: clubes profissionais, agências de *scouting*, plataformas especializadas. Em contextos informais e amadores, como jogos recreativos entre amigos, torneios universitários, ligas de "várzea" e campeonatos semiprofissionais, a realidade é substancialmente diferente.

Nesses ambientes, a formação de equipes costuma ocorrer de maneira inteiramente subjetiva, baseada na percepção empírica dos organizadores sobre a habilidade dos jogadores envolvidos. Métodos como sorteio aleatório, escolha alternada por capitães ou distribuição por posição nominal são os mais comuns.

Embora simples e de fácil aplicação, essas abordagens frequentemente resultam em equipes com grandes disparidades de nível e perfil funcional, conduzindo a partidas desequilibradas que comprometem a competitividade, a motivação dos participantes e a qualidade geral da experiência esportiva.

O problema do equilíbrio de equipes é, em sua essência, um problema de alocação: dado um conjunto de jogadores com características heterogêneas, como distribuí-los entre dois ou mais grupos de forma que as equipes resultantes sejam funcionalmente equivalentes? A complexidade desse problema reside no fato de que o equilíbrio não

depende apenas do nível técnico médio dos jogadores. É necessário também considerar a diversidade de perfis funcionais presentes em cada time. Cada equipe deve dispor de jogadores capazes de desempenhar as diferentes funções exigidas por uma partida de futebol: proteção defensiva, construção de jogo, criação ofensiva e finalização.

Nesse cenário, surge uma oportunidade relevante para a aplicação de técnicas computacionais de agrupamento. Se for possível identificar, a partir de dados de desempenho, grupos de jogadores com perfis funcionais similares, esses agrupamentos poderão ser utilizados como base para uma distribuição mais objetiva, equilibrada e fundamentada dos atletas entre as equipes. Essa é a motivação central do presente trabalho.

1.3 Justificativa

A relevância deste trabalho pode ser analisada sob duas perspectivas complementares: a acadêmica e a prática.

Do ponto de vista acadêmico, o trabalho propõe a integração entre três áreas do conhecimento: aprendizado de máquina não supervisionado, ciência de dados e análise esportiva, constituindo um problema interdisciplinar compatível com o perfil de formação do curso de Ciência da Computação. Em particular, o estudo explora a aplicação de algoritmos de clusterização em dados de alta dimensionalidade combinada com técnicas de redução dimensional em um domínio complexo, ruidoso e com grande variabilidade entre observações.

A investigação da capacidade de características individuais de desempenho representarem comportamentos funcionais coletivos em equipes de futebol representa uma contribuição original à literatura de *sports analytics*. Conforme documentado por Gu et al. (2025), essa área ainda carece de estudos que integrem o *pipeline* completo de clusterização à formação prática de equipes equilibradas.

Do ponto de vista prático, o trabalho endereça um problema cotidiano e amplamente vivenciado por milhões de praticantes de futebol amador e recreativo ao redor do mundo. A criação de um mecanismo objetivo e acessível de apoio à formação de equipes

equilibradas tem potencial de trazer melhorias à experiência competitiva em torneios universitários, ligas amadoras, peladas organizadas e outros contextos informais.

Além disso, a aplicação desenvolvida para apoiar a pesquisa apresenta, adicionalmente, potencial de continuidade e expansão. Uma versão mais completa poderia incluir funcionalidades de cadastro de jogadores, geração automática de tabelas de campeonato, monitoramento de desempenho ao longo do tempo e apoio à tomada de decisões táticas.

Um aspecto metodológico importante do trabalho refere-se à escolha da base de dados utilizada nos experimentos. Embora o objetivo final esteja voltado ao contexto do futebol amador e recreativo, optou-se por utilizar estatísticas históricas de jogadores das cinco principais ligas europeias devido à disponibilidade de uma base de dados pública, estruturada e suficientemente rica em atributos de desempenho. A construção de uma base equivalente para jogadores amadores exigiria a coleta sistemática de dados de partidas, seu registro, tratamento e validação, atividades incompatíveis com o tempo disponível para o desenvolvimento deste TCC.

Assim, a utilização de dados do futebol profissional não constitui o objetivo final da pesquisa, mas uma estratégia metodológica para avaliar a viabilidade da abordagem proposta em um ambiente controlado, utilizando dados de elevada qualidade e diversidade de perfis. Uma vez validada a metodologia de agrupamento, ela poderá ser posteriormente aplicada e ajustada a bases de dados provenientes do futebol amador.

1.4 Objetivos

1.4.1 Objetivo Geral

Desenvolver e avaliar um modelo de agrupamento para identificação de perfis funcionais de jogadores de futebol, utilizando estatísticas de desempenho de atletas profissionais como base experimental, e investigar como os resultados do agrupamento podem ser utilizados em um método automático de montagem de equipes técnica e funcionalmente balanceadas, com vistas a alcançar maior equilíbrio competitivo entre essas formações.

1.4.2 Objetivos Específicos

- Realizar revisão bibliográfica focada em ciência de dados e técnicas de aprendizado de máquina aplicados ao esporte, com ênfase em algoritmos de clusterização e sua aplicação ao futebol;
- Obter, integrar e tratar uma base de dados histórica contendo estatísticas individuais de jogadores das cinco principais ligas do futebol europeu (temporada 2024/2025), preparando-a para os experimentos de clusterização;
- Aplicar técnicas de pré-processamento, incluindo normalização por tempo de jogo, filtragem por minutagem mínima e padronização das variáveis, para preparar os dados para a etapa de modelagem;
- Aplicar a técnica de Análise de Componentes Principais (PCA) para redução de dimensionalidade da base, eliminando redundâncias e preparando os dados para os algoritmos de agrupamento;
- Executar e comparar algoritmos de clusterização (K-Means e Hierárquico com método de ligação de Ward), avaliando a qualidade dos agrupamentos por meio de índices de validação interna e de critério interpretativo futebolístico;
- Comparar os *clusters* gerados com as posições nominais dos atletas na base de dados original, como forma de validação externa da capacidade discriminativa dos agrupamentos;
- Definir uma estratégia algorítmica para distribuição equilibrada dos jogadores entre equipes com base nos agrupamentos obtidos, considerando simultaneamente a diversidade funcional dos perfis e o equilíbrio técnico relativo dos atletas;
- Desenvolver uma *dashboard* interativa em Python que permita a aplicação prática do sistema de geração automática de equipes balanceadas.

1.5 Organização do Trabalho

O presente trabalho está organizado em seis capítulos, iniciando pela presente Introdução, além dos elementos pré-textuais e das referências bibliográficas.

O Capítulo 2 apresenta a revisão bibliográfica, cobrindo os fundamentos teóricos e o estado da arte nas áreas relevantes para o trabalho: ciência de dados e *sports analytics* no futebol; aprendizado de máquina no contexto esportivo; técnicas de clusterização e aprendizado não supervisionado; aplicações de clusterização a jogadores de futebol; e a problemática de formação de equipes equilibradas.

O Capítulo 3 descreve a metodologia adotada, detalhando cada etapa do *pipeline* de ciência de dados utilizado. Para cada etapa, são apresentadas as escolhas metodológicas realizadas e seus fundamentos na literatura.

O Capítulo 4 apresenta e discute os resultados obtidos em cada etapa do *pipeline*, incluindo a caracterização da base de dados após o pré-processamento, os resultados da redução dimensional, a análise qualitativa e quantitativa dos agrupamentos produzidos pelos dois algoritmos avaliados, a definição do modelo final e sua aplicação na formação de equipes balanceadas.

O Capítulo 5 descreve o desenvolvimento da aplicação interativa *Football Cluster Generator*, apresentando seus objetivos, arquitetura, módulos funcionais e a sua relação com os resultados da pesquisa.

O Capítulo 6 sintetiza as conclusões do estudo, discute as contribuições acadêmicas e práticas do trabalho, aponta suas limitações e propõe direções para investigações futuras.

2

REVISÃO BIBLIOGRÁFICA

2.1 Ciência de Dados e Análise no Futebol

A aplicação de métodos quantitativos no esporte tem crescido significativamente nas últimas décadas, impulsionada pelo aumento da disponibilidade de dados e pelo avanço das técnicas de análise computacional. No futebol, em particular, a coleta sistemática de estatísticas de jogadores e partidas possibilitou o surgimento da área conhecida como *sports analytics*, campo dedicado a extrair conhecimento a partir de dados para apoiar decisões esportivas estratégicas, táticas e de gestão de elenco.

Essa transformação tem raízes em uma mudança de paradigma mais ampla no esporte profissional. Como ilustrado por Lewis (2003) em *Moneyball*, a análise estatística pode revelar padrões não evidentes à observação empírica e gerar vantagem competitiva mesmo com recursos financeiros limitados. O paradigma inaugurado no beisebol norte-americano rapidamente influenciou outras modalidades, incluindo o futebol.

No futebol, por sua vez, já tem sido observado a aplicação da abordagem analítica em diversos contextos, como, por exemplo: na avaliação de desempenho individual e coletivo das equipes; no recrutamento de talentos; na definição de estratégias de treino e de jogo; e até mesmo na tentativa de previsão de resultados.

Ressalte-se que as bases de dados modernas do futebol apresentam grande riqueza de atributos, frequentemente reunindo dezenas ou centenas de variáveis relativas a ações ofensivas, defensivas, de movimentação e de contexto nos jogos. Esse volume de informações, embora amplie o potencial analítico, introduz desafios metodológicos relevantes.

Em particular, essas variáveis costumam apresentar elevado grau de correlação entre si, dificultando a análise direta e impactando negativamente as técnicas de aprendizado de máquina, fenômeno denominado *maldição da dimensionalidade* (VEN, 2018). Diante disso, torna-se necessária a aplicação de técnicas de pré-processamento e redução de dimensionalidade antes da modelagem, etapas que preservam a estrutura essencial dos dados enquanto eliminam redundâncias e ruídos (GU et al., 2025).

2.2 Aprendizado de Máquina no Contexto Esportivo

O aprendizado de máquina constitui uma das principais ferramentas da ciência de dados aplicada ao esporte. Essa área envolve o desenvolvimento de modelos computacionais capazes de aprender padrões a partir de dados históricos e realizar tarefas como classificação, regressão e agrupamento, sem que regras explícitas sejam definidas pelo programador. Sua aplicação ao contexto esportivo tem crescido de forma acelerada à medida que o volume e a qualidade dos dados disponíveis aumentam.

No futebol, modelos supervisionados têm sido amplamente utilizados para previsão de resultados de partidas, detecção de padrões táticos e estimação do valor de mercado de jogadores. No entanto, o aprendizado não supervisionado também desempenha papel relevante e crescente, especialmente na identificação de padrões ocultos em dados sem rótulos predefinidos, como é o caso da segmentação de perfis de jogadores (D'URSO; GIOVANNI; VITALE, 2023; GARCíA-ALIAGA et al., 2021).

Técnicas de aprendizado de máquina permitem a exploração eficiente de grandes volumes de dados, sendo amplamente implementadas por meio de bibliotecas como o Scikit-learn (PEDREGOSA et al., 2011), que oferece modelos robustos e bem documentados de algoritmos de classificação, regressão e clusterização. A popularização dessas ferramentas tem democratizado o acesso a métodos sofisticados de análise em contextos esportivos.

Em uma revisão sistemática sobre o tema, Gu et al. (2025) documentam que o uso combinado de redução de dimensionalidade e aprendizado não supervisionado representa um dos fluxos metodológicos mais recorrentes na literatura de *sports analy-*

tics, com destaque para aplicações no futebol, que respondem por mais de 80% dos estudos em comparação com outros esportes. Os autores identificam aplicações tanto na caracterização de estilos de jogo de equipes quanto na definição de perfis de jogadores.

A escolha das técnicas de aprendizado de máquina depende das características do problema e dos dados disponíveis. Em cenários exploratórios, nos quais não existem categorias previamente definidas que possam guiar o aprendizado, o aprendizado não supervisionado, e em especial a clusterização, apresenta-se como a abordagem mais adequada. Conforme enfatizado por Hennig (2015a), a escolha do método de agrupamento deve ser orientada pelo objetivo da análise e pelo "conceito de *cluster*" que melhor representa o interesse da aplicação específica.

2.3 Clusterização e Aprendizado Não Supervisionado

A clusterização é uma técnica de aprendizado não supervisionado cujo objetivo é agrupar elementos semelhantes em um mesmo conjunto, de modo que os elementos dentro de um grupo apresentem maior similaridade entre si do que em relação aos elementos de outros grupos. Por não exigir exemplos rotulados, ela é especialmente adequada para explorar estruturas latentes em dados complexos e de alta dimensionalidade (PEDREGOSA et al., 2011).

Entre os algoritmos mais utilizados, destaca-se o K-Means, que busca particionar os dados em um número pré-definido de *clusters* minimizando a variância intra-grupo em relação ao centroide de cada agrupamento. Conforme descrito por Hartigan e Wong (1979) (apud Haan et al. (2025)), todas as observações são atribuídas ao centroide mais próximo, e grandes distâncias dentro dos *clusters* são fortemente penalizadas, comportamento que favorece a formação de grupos compactos e aproximadamente esféricos.

A definição do número ideal de *clusters* é um aspecto crítico do método, sendo frequentemente realizada por meio de técnicas auxiliares como o método do "cotovelo", baseado na variação da inércia total, e o coeficiente de silhueta, que mensura a coesão intra-*cluster* e a separação entre *clusters* (MAUX, 2024; AZZAMI et al., 2025).

Além do K-Means, os algoritmos de clusterização hierárquica têm recebido considerável atenção na literatura. Conforme demonstrado por Ran et al. (2023), esses métodos geram múltiplas partições consistentes dos dados em diferentes níveis de granularidade para o mesmo conjunto de dados, sem a necessidade de executar o agrupamento novamente. Essa característica permite analisar de forma mais rica e flexível a estrutura dos dados.

Os métodos hierárquicos constroem uma estrutura em árvore de agrupamentos, representada graficamente por um dendrograma, na qual cada fusão sucessiva de grupos pode ser avaliada visualmente. Isso permite que o analista identifique o número mais adequado de *clusters* com base tanto em critérios matemáticos quanto em considerações práticas e interpretativas. Gere (2023) destaca que o dendrograma é visualmente atraente e sugere *clusters* bem distribuídos, mas recomenda que sua análise seja complementada por índices de validade quantitativos.

No contexto de dados com múltiplos atributos correlacionados, como ocorre frequentemente nas estatísticas de futebol, a aplicação de redução de dimensionalidade torna-se fundamental antes da clusterização. Uma das abordagens mais utilizadas é a Análise de Componentes Principais (PCA, do inglês *Principal Component Analysis*), que transforma o conjunto original de variáveis em um novo conjunto de componentes ortogonais, e, portanto, não correlacionados, preservando a maior parte da variância dos dados e eliminando redundâncias (GARCíA-ALIAGA et al., 2021; AZZAMI et al., 2025).

Em uma revisão sistemática sobre o uso do PCA em ciências do esporte coletivo, Rojas-Valverde et al. (2020) apontam que uma retenção de variância acumulada a partir de 70% é bastante aceitável na média dos trabalhos analisados. De forma convergente, estudos de PCA em esportes coletivos normalmente retêm componentes que explicam entre 70% e 85% da variância, com média próxima de 80% (PINO-ORTEGA et al., 2021). Essa faixa representa um equilíbrio adequado entre simplificação do problema e preservação da estrutura informacional dos dados.

A avaliação da qualidade dos agrupamentos gerados é etapa crucial em qualquer análise de clusterização. Hennig (2015a) lista múltiplas abordagens complementares

para essa finalidade: uso de informações externas; índices de validação interna; avaliações de estabilidade; exploração visual; e comparação de diversas soluções de agrupamento. Entre os índices internos mais empregados em *sports analytics*, destacam-se:

- **Inércia (método do "cotovelo")**: mede a soma das distâncias quadráticas de cada ponto ao centroide de seu *cluster*. Quanto menor a inércia, mais coesos são os grupos. Ao plotar a inércia em função do número de *clusters*, busca-se o ponto em que a redução marginal passa a ser pequena (o "cotovelo" da curva), indicando o número de *clusters* a partir do qual ganhos adicionais de coesão não justificam a maior complexidade do modelo (MAUX, 2024).

- **Coeficiente de Silhueta**: para cada ponto, calcula a diferença entre a distância média às demais observações de seu próprio *cluster* e a distância média ao *cluster* vizinho mais próximo, normalizando esse valor pelo maior dos dois. O resultado varia de -1 a $+1$: valores próximos de $+1$ indicam que o ponto está bem alocado em seu *cluster* e claramente separado dos demais; valores próximos de 0 indicam sobreposição entre *clusters*; e valores negativos sugerem má alocação. O *Average Silhouette Width* (**ASW**) é a média desse índice para todos os pontos do conjunto de dados, sendo amplamente utilizado para comparar soluções de agrupamento com diferentes números de *clusters* (WIJNGAARD, 2020).

- **Índice de *Dunn***: razão entre a menor distância entre pontos de *clusters* diferentes e o maior diâmetro intra-*cluster*. Valores mais altos indicam *clusters* bem separados e internamente compactos. Por considerar simultaneamente separação e coesão, é considerado um índice equilibrado, embora sensível a ruído e *outliers* (AKHANLI; HENNIG, 2023).

- **Índice de *Calinski-Harabász* (CH)**: calcula a razão entre a dispersão inter-*cluster* (quão distantes estão os centróides dos *clusters* entre si) e a dispersão intra-*cluster* (quão compactos são os *clusters* internamente), ponderada pelo número de *clusters* e de observações. Valores mais elevados indicam *clusters* bem definidos e separados. Esse índice é computacionalmente eficiente e favorece a separação entre grupos, podendo subestimar soluções com maior número de *clusters* (VEN, 2018).

- **Índice de *Davies-Bouldin* (DBI)**: para cada *cluster*, calcula a razão média entre a dispersão interna daquele *cluster* e a distância ao *cluster* mais próximo e toma a média dessas razões sobre todos os *clusters*. Ao contrário dos índices anteriores, valores menores indicam melhor qualidade de agrupamento, pois refletem *clusters* mais compactos e mais bem separados. (VEN, 2018) observa que o *DBI* tende a favorecer soluções com mais *clusters* em relação a índices como *ASW* e *CH*, sendo mais sensível à homogeneidade interna do que à separação global, o que o torna útil quando esse aspecto é prioritário (MAUX, 2024; VEN, 2018).

Um aspecto conceitual relevante, discutido por Hennig (2015b), é que não existe um agrupamento "verdadeiro" de forma objetiva e independente do contexto. O que constitui um agrupamento "bom" depende do objetivo da análise: diferentes agrupamentos podem ser ótimos em aspectos distintos dos mesmos dados, e os dados, por si sós, não determinam o agrupamento ideal. Por isso, a seleção do método e do número de *clusters* deve ser guiada, não só pelos índices de validade, como os enumerados acima, mas também pelas necessidades práticas da aplicação avaliadas pelo pesquisador.

2.4 Clusterização de Jogadores no Futebol

No futebol, a clusterização tem sido empregada para identificar padrões de comportamento e perfis de jogadores com base em atributos técnicos, táticos e físicos. Os estudos nessa área demonstram que os agrupamentos frequentemente transcendem as categorias posicionais tradicionais, revelando estruturas funcionais mais ricas e informativas (DAğ; YüKSEL; ATMACA, 2023; HAAN et al., 2025).

D'Urso, Giovanni e Vitale (2023) sistematizam as principais aplicações identificadas na literatura: orientar a formação e composição de equipes; apoiar a seleção e o recrutamento de jogadores; determinar o valor de mercado de atletas; classificar e monitorar o desempenho; direcionar o planejamento de treinamentos; e explicar fatores associados ao sucesso das equipes. Essa diversidade de aplicações evidencia o potencial transformador da técnica para diferentes agentes do ecossistema futebolístico.

Entre os trabalhos de referência na área, Akhanli e Hennig (2023) propõem a

utilização de medidas de dissimilaridade combinadas com índices agregados de validação para segmentar jogadores das principais ligas europeias em grupos homogêneos. Uma contribuição central do estudo é justamente a avaliação comparativa de diferentes soluções de agrupamento, provenientes de diferentes métodos e diferentes números de *clusters*, selecionando aquela que melhor atende ao objetivo específico da análise.

Os resultados de Akhanli e Hennig (2023) apontaram o método aglomerativo de Ward (JR., 1963) com 5 *clusters* como melhor solução, identificando os seguintes perfis (excluindo goleiros): meio-campistas; laterais; zagueiros; meio-campistas ofensivos; e atacantes. Os autores destacam que esse modelo de agrupamento fornece uma representação interpretável da estrutura dos dados, útil para analisar a composição das equipes em termos de perfis funcionais e relacioná-los ao sucesso esportivo, perspectiva diretamente alinhada ao objetivo do presente trabalho.

Do ponto de vista do pré-processamento, Akhanli e Hennig (2023) adotam a exclusão de jogadores com menor tempo de participação e a normalização dos dados de contagem por 90 minutos, além da padronização (*standardization*) das variáveis, procedimento que controla o peso relativo das variáveis entre si quando agregadas. Esses cuidados visam evitar distorções causadas por jogadores com poucos minutos disputados e garantir comparabilidade entre diferentes escalas de medição. García-Aliaga et al. (2021) adotam critério similar, excluindo jogadores com menos de 360 minutos disputados para evitar estatísticas irreais.

Wijngaard (2020) investiga a aplicação do K-Means em jogadores da *La Liga* espanhola, utilizando o coeficiente de silhueta como critério de validação. Os resultados indicam a existência de aproximadamente oito perfis distintos de jogadores, número que transcende as posições clássicas em campo. Esse achado é corroborado por Behravan et al. (2019), que, utilizando um algoritmo baseado em otimização por enxame de partículas (PSO), com seleção automática do número de *clusters*, também identificaram cerca de oito posições funcionais predominantes entre os jogadores de linha.

A comparação entre o agrupamento obtido por algoritmos não supervisionados e as posições nominais dos jogadores é um procedimento validatório recorrente na literatura. Dağ, Yüksel e Atmaca (2023) obtiveram 81% de precisão na recuperação

de quatro categorias posicionais tradicionais. Haan et al. (2025), por outro prisma, demonstram que agrupamentos baseados em demandas físicas de partidas podem ser mais coerentes do que as posições nominais, encontrando cinco *clusters* ideais (atacantes; meias ofensivos; meias defensivos; zagueiros centrais; e laterais), validados por gráficos de cotovelo e de silhueta.

García-Aliaga et al. (2021) acrescentam contribuição metodológica relevante ao aplicar técnicas de visualização agrupada de jogadores representados por suas características em campo. Os resultados mostram que jogadores de funções semelhantes tendem a ocupar regiões próximas no espaço de características, validando empiricamente a premissa de que o agrupamento por desempenho captura estruturas funcionais reais. O estudo também compara os resultados do agrupamento com as posições registradas na base de dados original, abordagem que se pretende replicar neste trabalho.

Amhimmid et al. (2025), em estudo comparativo utilizando diferentes tipos de dados sintéticos (*clusters* esféricos, sobrepostos, com densidades variadas, alongados e com ruído), demonstra as vantagens do método de *Ward* na redução da variância interna dos *clusters*. Gere (2023) corrobora essa conclusão, indicando que o método de *Ward* é o preferido quando se deseja grupos bem definidos, compactos e homogêneos, características especialmente desejáveis quando se trabalha com atributos numéricos normalizados, como as métricas de desempenho de futebol.

O método de *Ward* busca minimizar a variância dentro dos *clusters* a cada fusão, resultando em grupos mais compactos e aproximadamente esféricos. Essa propriedade é especialmente relevante no contexto do futebol, onde a similaridade dos jogadores no mesmo *cluster* é uma condição elementar para a interpretabilidade dos agrupamentos. Akhanli e Hennig (2023) observam que índices voltados à homogeneidade intra-*cluster* são especialmente úteis nesse contexto, uma vez que a coesão interna é um requisito essencial para que os *clusters* de jogadores sejam funcionalmente significativos.

Além da homogeneidade, Ran et al. (2023) destacam uma vantagem estrutural dos métodos hierárquicos frente ao K-Means: a possibilidade de gerar múltiplas partições consistentes dos dados em diferentes níveis de granularidade, sem a necessidade de reexecutar o algoritmo. Akhanli e Hennig (2020) ressaltam ainda a tensão

inerente entre homogeneidade intra-*cluster* (que melhora com o aumento do número de *clusters*) e separação entre *clusters* (que melhora com a sua redução). No presente trabalho, privilegia-se a homogeneidade intra-*cluster*, buscando o maior número de *clusters* possível sem ultrapassar o número de posições funcionais existentes em campo.

Experiências em outros esportes reforçam a validade dessa abordagem. Dalton-Barron et al. (2022) aplicaram com sucesso a clusterização hierárquica com *Ward* na *rugby league profissional* para avaliar similaridades entre posições de jogo. Wing et al. (2025) empregaram *K-Means* na organização de grupos de treinamento no futebol australiano. E Penner (2025) utilizou agrupamentos para definir arquétipos funcionais no basquete. Em conjunto, esses precedentes constroem uma base sólida para a aplicação proposta neste trabalho.

2.5 Formação de Equipes e Equilíbrio Competitivo

A formação de equipes equilibradas é um problema relevante tanto no esporte profissional quanto em competições amadoras. O equilíbrio competitivo pode ser definido como a redução das disparidades de desempenho entre equipes, promovendo maior imprevisibilidade e competitividade nas partidas. Do ponto de vista da teoria esportiva, esse equilíbrio é reconhecido como fator determinante para o engajamento de participantes e espectadores.

Do ponto de vista computacional, o problema pode ser enquadrado como uma questão de otimização e alocação, na qual se busca distribuir elementos entre grupos de forma a minimizar diferenças entre eles segundo algum critério de interesse. Métodos simples, como distribuição aleatória ou balanceamento por médias globais de desempenho, frequentemente não garantem equilíbrio real, especialmente quando há alta variabilidade de habilidades e perfis entre os jogadores. Isso ocorre porque a média global não captura a diversidade de funções que cada jogador desempenha em campo.

É nesse contexto que a clusterização emerge como abordagem promissora. Ao agrupar jogadores com base em perfis de desempenho similares, vai além de uma simples medida de "nível geral" e considera a diversidade funcional dos atletas. O

equilíbrio competitivo passa a ser entendido não apenas como igualdade de potencial médio, mas como equivalência na distribuição de tipos funcionais entre as equipes, garantindo a todos uma composição plural e complementar.

Esse princípio encontra respaldo empírico em Ven (2018), que utiliza K-Means e PCA para diferenciar tipos de jogadores e investigar como diferentes combinações de perfis influenciam o resultado de partidas. O estudo, apontado como pioneiro no uso de características técnicas individuais para estimar resultados no futebol, demonstra que a composição equilibrada de perfis distintos pode ser fator determinante para o desempenho coletivo da equipe.

Em trabalho análogo no basquete, Penner (2025) demonstra que o agrupamento de jogadores com base em suas performances em jogo, para definir arquétipos funcionais, pode ser empregado diretamente na composição de equipes completas e equilibradas. O autor também adota a normalização dos dados individuais por tempo de jogo, assegurando a comparabilidade entre jogadores com diferentes cargas de participação. Embora relativo a outro esporte, esse trabalho oferece suporte metodológico direto ao objetivo central do presente trabalho.

No futebol, Prasetyo, Priyatama e Setyatama (2023) utilizam clusterização para determinar perfis de preço de jogadores no mercado a partir de dados de desempenho separados por posição dominante, evidenciando que os *clusters* obtidos são suficientemente informativos para apoiar decisões práticas de gestão de elenco. García-Aliaga et al. (2021) reforçam essa perspectiva ao demonstrar que a distribuição equilibrada de exemplares de *clusters* distintos entre equipes tende a produzir composições mais completas e funcionalmente diversas.

Em síntese, a literatura evidencia que a clusterização oferece uma base objetiva e estruturada para a formação de equipes equilibradas. Ao representar cada jogador como pertencente a um perfil funcional determinado por seus atributos de desempenho, torna-se possível compor equipes não apenas igualmente fortes em termos absolutos, mas igualmente diversas em termos de tipos de jogadores, condição que, conforme demonstrado por Ven (2018) e Penner (2025), está associada ao desempenho coletivo superior e ao equilíbrio competitivo.

2.6 Lacuna na Literatura e Contribuição do Trabalho

A revisão apresentada nas seções anteriores evidencia um campo de pesquisa ativo e crescente, no qual a clusterização de jogadores com base em dados de desempenho está bem estabelecida. Contudo, ao examinar os objetivos e resultados dos trabalhos disponíveis, identifica-se uma lacuna relevante: a aplicação direta e sistemática dessas técnicas para a formação de equipes equilibradas permanece pouco explorada.

A maior parte dos trabalhos concentra-se na identificação de perfis, na validação metodológica dos agrupamentos ou na comparação com categorias posicionais tradicionais, sem explorar de forma aprofundada a utilização dos *clusters* como base para estratégias de composição de equipes. Os trabalhos de Ven (2018) e Penner (2025) constituem exceções parciais, ao investigar a relação entre combinações de perfis e resultados coletivos, mas sem propor uma metodologia sistemática e replicável para formação de equipes equilibradas a partir de *clusters*.

Adicionalmente, poucos estudos exploram de forma integrada o *pipeline* completo: pré-processamento e normalização dos dados, redução de dimensionalidade via PCA, clusterização hierárquica com validação por múltiplos índices e aplicação dos resultados para equilíbrio de equipes. O presente trabalho propõe investigar exatamente essa lacuna, utilizando dados de desempenho de jogadores das principais ligas europeias e o método hierárquico de *Ward* como ferramenta para formação de equipes balanceadas, mediante a distribuição equilibrada de exemplares dos *clusters* entre as equipes.

Do ponto de vista metodológico, o trabalho se funda em contribuições consolidadas: o uso de PCA para redução de dimensionalidade (GARCíA-ALIAGA et al., 2021; ROJAS-VALVERDE et al., 2020; GU et al., 2025); a clusterização hierárquica com método *Ward* para grupos compactos e homogêneos (AKHANLI; HENNIG, 2023; AMHIMMID et al., 2025; GERE, 2023); a validação por múltiplos índices internos (AKHANLI; HENNIG, 2020; VEN, 2018); e a comparação com posições nominais como um dos processos de validação externa (WIJNGAARD, 2020; DAğ; YüKSEL; ATMACA, 2023; HAAN et al., 2025).

3

METODOLOGIA

3.1 Caracterização da Pesquisa

O presente trabalho caracteriza-se como uma pesquisa aplicada, de abordagem quantitativa e natureza experimental, inserida no contexto da Ciência de Dados aplicada ao esporte. O estudo possui caráter exploratório e descritivo: exploratório porque busca identificar estruturas latentes em dados de desempenho de jogadores sem a imposição de categorias predefinidas; e descritivo porque visa caracterizar os perfis funcionais encontrados e descrever o comportamento do modelo de balanceamento proposto.

A metodologia adotada segue um *pipeline* clássico de trabalhos em ciência de dados aplicada ao esporte (MAUX, 2024), estruturado em cinco etapas sequenciais: (i) coleta e consolidação dos dados; (ii) pré-processamento e tratamento da base; (iii) redução de dimensionalidade; (iv) clusterização dos jogadores e validação dos agrupamentos; e (v) aplicação dos agrupamentos para formação de equipes balanceadas. Do ponto de vista computacional, a pesquisa foi desenvolvida em Python com as bibliotecas Pandas, NumPy, Scikit-learn (PEDREGOSA et al., 2011), SciPy, Matplotlib e Seaborn.

3.2 Coleta e Construção da Base de Dados

3.2.1 Fontes de Dados

Os dados utilizados neste trabalho foram obtidos a partir de duas plataformas amplamente reconhecidas na área de *sports analytics*: Sofascore e FBref.

A plataforma Sofascore (Sofascore, 2026) disponibiliza estatísticas avançadas relacionadas ao desempenho individual dos jogadores durante partidas, incluindo estatísticas ofensivas, defensivas, de construção de jogo, posse de bola e ações específicas por posição. Já a plataforma FBref (FBref, 2026) foi utilizada para complementar a base com informações contextuais, como a posição nominal dos atletas, idade e dados de desempenho das equipes nas respectivas competições.

Foram coletados dados referentes à temporada 2024/2025 das cinco principais ligas do futebol europeu: Premier League (Inglaterra), La Liga (Espanha), Serie A (Itália), Bundesliga (Alemanha) e Ligue 1 (França). A escolha por essas competições segue o mesmo escopo adotado por Akhanli e Hennig (2023) e D'Urso, Giovanni e Vitale (2023), garantindo representatividade geográfica, estilística e de nível competitivo ao conjunto de dados. Os dados extraídos das diferentes ligas foram posteriormente integrados em um único *dataframe*, constituindo a base consolidada utilizada nos experimentos.

A escolha de uma base composta por jogadores profissionais não altera o propósito final da pesquisa, voltado à formação de equipes balanceadas no futebol amador. Essa decisão foi motivada pela necessidade de avaliar a metodologia proposta utilizando um conjunto de dados amplo, consistente e previamente estruturado, permitindo concentrar os esforços da pesquisa na avaliação dos métodos de agrupamento e da estratégia de balanceamento das equipes.

3.2.2 Extração dos Dados

A extração dos dados da plataforma Sofascore foi realizada por meio da biblioteca ScraperFC (NORRIS, 2024), disponível para a linguagem Python, que permitiu a coleta automatizada das estatísticas dos jogadores. No caso da plataforma FBref, mecanismos de proteção da página limitaram a utilização de métodos tradicionais de *web scraping*. Em razão disso, a obtenção das tabelas complementares foi realizada manualmente, por meio da exportação dos conteúdos renderizados em HTML para arquivos no formato `.xlsx`, posteriormente integrados à base principal via Pandas.

3.2.3 Estrutura dos Dados

A base construída contém mais de uma centena de atributos relacionados ao desempenho dos jogadores durante a temporada analisada. As variáveis contemplam cinco grandes dimensões do jogo: evidências ofensivas (gols, assistências, finalizações, criação de grandes chances, xG e xA); atributos defensivos (desarmes, interceptações, bloqueios e recuperações de bola); de construção de jogo (passes completados, passes progressivos, cruzamentos e bolas longas).

Incluem-se ainda atributos de posse e movimentação (dribles, conduções e pressão sobre adversários) e estatísticas específicas de goleiros (defesas, saídas e distribuição). Essa diversidade de atributos permite representar os jogadores de forma multidimensional, possibilitando a identificação de padrões de similaridade baseados em desempenho efetivo em campo, abordagem alinhada à perspectiva de García-Aliaga et al. (2021) e Akhanli e Hennig (2023).

3.3 Pré-processamento e Tratamento dos Dados

A etapa de pré-processamento teve como objetivo garantir a qualidade, consistência e comparabilidade dos dados antes da modelagem. Cada procedimento adotado está fundamentado em práticas recomendadas na literatura e aplicadas de forma sequencial.

3.3.1 Tratamento de Valores Ausentes

Foram identificadas colunas com elevado percentual de valores ausentes. As variáveis `outfielderBlocks` e `goalsPrevented` foram removidas por apresentarem mais de 90% de valores nulos, percentual que inviabiliza qualquer estratégia de imputação sem comprometer a integridade dos dados. As colunas `attemptPenaltyMiss` e `attemptPenaltyPost` foram excluídas por apresentarem exclusivamente valores iguais a zero em todos os registros, não contribuindo com informação discriminativa para o processo de agrupamento.

Para variáveis com pequena quantidade de valores ausentes, foram adotadas

estratégias de imputação localmente justificadas. A coluna categórica `Pos` recebeu a categoria "UD" (undefined) para os registros sem posição definida. Variáveis numéricas como `expectedGoals` e `expectedAssists` foram preenchidas com a mediana dos jogadores da mesma posição nominal, preservando a estrutura posicional dos dados. A variável `Age` recebeu a mediana global da base.

3.3.2 Remoção de Registros Duplicados

Foi identificada a ocorrência de jogadores que atuaram por mais de uma equipe durante a temporada, resultando em registros duplicados na base integrada. Nesses casos, manteve-se apenas o registro correspondente ao maior número de minutos jogados (`minutesPlayed`). Essa decisão parte do pressuposto de que o registro com maior minutagem representa de forma mais fiel e estável o padrão de desempenho predominante do atleta ao longo da temporada.

3.3.3 Normalização por 90 Minutos

Diversas estatísticas acumulativas foram normalizadas por 90 minutos jogados, convertendo-as em taxas de ocorrência por unidade de tempo padronizada. Esse procedimento, também adotado por Akhanli e Hennig (2023) e Penner (2025), é fundamental para garantir a comparabilidade entre jogadores com diferentes cargas de participação na temporada.

Sem essa normalização, jogadores com mais minutos disputados tenderiam a apresentar estatísticas acumuladas mais elevadas simplesmente em função do maior tempo de jogo, distorcendo as distâncias calculadas pelos algoritmos de agrupamento. A normalização foi aplicada a todas as métricas de natureza acumulativa nas dimensões ofensiva, defensiva, de construção, de posse e específicas de goleiros.

3.3.4 Seleção e Exclusão de Variáveis

Após as etapas de limpeza e normalização, realizou-se uma filtragem criteriosa das variáveis a serem utilizadas no processo de clusterização. Foram removidas variáveis categóricas e identificatórias, como nome do atleta, equipe de origem e posição nominal, preservadas separadamente para uso posterior na interpretação e validação dos agrupamentos. Essa remoção é essencial para garantir que o algoritmo opere exclusivamente sobre atributos de desempenho, sem ser influenciado por rótulos que poderiam induzir artificialmente a formação de grupos.

Também foram excluídas variáveis de natureza subjetiva, como atributos de avaliação geral (*ratings*) e médias de desempenho compostas, para evitar que avaliações previamente sintetizadas influenciassem artificialmente a estrutura dos agrupamentos. Por fim, foram removidas variáveis redundantes ou fortemente dependentes do contexto coletivo, como `minutesPlayed`, `appearances`, `scoringFrequency`, `matchesStarted` e `cleanSheet`, cujos efeitos já haviam sido parcialmente capturados pela normalização por 90 minutos ou que refletem predominantemente fatores coletivos da equipe em vez de características individuais do atleta.

3.3.5 Filtragem por Minutos Jogados

Jogadores com baixa minutagem acumulada na temporada tendem a apresentar estatísticas normalizadas artificialmente infladas. Por exemplo, um jogador que disputou apenas dois jogos e marcou um gol apresentará, após a normalização, uma taxa de gols por 90 minutos consideravelmente superior à de um artilheiro regular, distorção que pode comprometer gravemente a qualidade dos agrupamentos. Esse risco é amplamente documentado na literatura: Akhanli e Hennig (2023) adotaram um corte correspondente a 37% do tempo máximo disputado; García-Aliaga et al. (2021) excluíram jogadores com menos de 360 minutos; e D'Urso, Giovanni e Vitale (2023) utilizaram o limiar de 200 minutos.

No presente trabalho, foram removidos os atletas com menos de 450 minutos disputados na temporada, equivalente a cinco partidas completas. A definição desse limiar

foi realizada empiricamente, por meio da análise da estabilização dos valores máximos observados em variáveis-chave à medida que o ponto de corte era progressivamente elevado em intervalos de 90 minutos. Observou-se que, a partir dos 450 minutos, os valores extremos deixaram de apresentar reduções significativas, indicando diminuição substancial da influência de registros estatisticamente instáveis na base.

3.3.6 Padronização dos Dados

Após as etapas de limpeza, seleção de variáveis e normalização por tempo de jogo, os dados foram padronizados utilizando o método `StandardScaler` da biblioteca Scikit-learn, que transforma cada variável para que apresente média zero e desvio padrão unitário. Essa padronização é essencial para o bom funcionamento de algoritmos de clusterização baseados em distância euclidiana, pois garante que variáveis com escalas numéricas mais elevadas não exerçam influência desproporcional na formação dos agrupamentos. Conforme observado por Akhanli e Hennig (2023), esse procedimento controla o peso relativo das variáveis entre si quando são agregadas na medida de dissimilaridade.

3.4 Redução de Dimensionalidade

Mesmo após as etapas de seleção e exclusão de variáveis, a base resultante ainda continha quase uma centena de atributos numéricos. Em bases com essa característica, é comum que muitas variáveis apresentem alto grau de correlação entre si, fenômeno que pode prejudicar os algoritmos de clusterização, uma vez que variáveis fortemente correlacionadas carregam informação redundante e distorcem as distâncias calculadas entre os jogadores.

Para lidar com esse problema, foi aplicada a técnica de Análise de Componentes Principais (PCA), que transforma o conjunto original de variáveis em um novo espaço de componentes ortogonais entre si, e, portanto, não correlacionados, preservando a maior parte da variância dos dados originais (GARCíA-ALIAGA et al., 2021; AZZAMI

et al., 2025).

A aplicação do PCA antes da clusterização é amplamente recomendada e documentada na literatura de *sports analytics*, especialmente em contextos de alta dimensionalidade (GU et al., 2025; ROJAS-VALVERDE et al., 2020).

A definição do número adequado de componentes principais a reter foi realizada com base em dois critérios complementares, em consonância com as recomendações de Rojas-Valverde et al. (2020) e Pino-Ortega et al. (2021).

O primeiro critério é a análise da curva de variância acumulada (*scree plot*), que permite identificar visualmente o ponto de inflexão a partir do qual os ganhos adicionais de variância explicada tornam-se marginais. O segundo é o critério de Kaiser, que estabelece que apenas os componentes com autovalor superior a 1 devem ser retidos, por explicarem uma quantidade de variância superior à de uma única variável padronizada original. A utilização conjunta dos dois critérios buscou equilibrar redução de dimensionalidade e preservação da estrutura informacional da base.

3.5 Clusterização dos Jogadores

3.5.1 Algoritmos Utilizados

A seleção dos algoritmos de clusterização levou em conta tanto a adequação ao problema quanto o alinhamento com as práticas documentadas na literatura de *sports analytics* (AKHANLI; HENNIG, 2023; MAUX, 2024). Foram avaliados dois métodos com características complementares.

O **K-Means** é o algoritmo de particionamento mais amplamente utilizado em aplicações de clusterização. Ele funciona iterativamente: inicializa k centróides no espaço de atributos, atribui cada observação ao centroide mais próximo, recalcula os centróides como médias dos pontos atribuídos e repete até a convergência (HARTIGAN; WONG, 1979). Suas principais vantagens são a eficiência computacional, a escalabilidade e a facilidade de interpretação dos centróides. Como limitações, exige a especificação prévia de k e é sensível à inicialização dos centróides e à presença de *outliers*.

A **clusterização hierárquica aglomerativa com método de ligação de Ward** constrói progressivamente uma árvore de agrupamentos (dendrograma), partindo de cada observação como seu próprio *cluster* e fundindo iterativamente os dois grupos cuja fusão resulte no menor incremento da variância intra-*cluster* total, critério que define o método de ligação de Ward (*Ward's linkage*) (JR., 1963).

Essa abordagem tem duas vantagens estruturais sobre o K-Means: não exige a especificação prévia de k, pois o dendrograma permite explorar todas as partições possíveis a partir de uma única execução do algoritmo (RAN et al., 2023); e tende a produzir *clusters* mais compactos e homogêneos em espaço euclidiano normalizado (AMHIMMID et al., 2025; GERE, 2023). Como desvantagem, apresenta maior custo computacional e não permite a reatribuição dinâmica de novos pontos sem reprocessamento completo da estrutura hierárquica.

A análise conjunta dos dois métodos visa não apenas identificar a solução de agrupamento mais adequada, mas também verificar a robustez dos perfis encontrados por meio da convergência entre abordagens com lógicas distintas.

3.5.2 Definição do Número de Clusters

O intervalo de valores de k avaliado, entre 5 e 11 *clusters*, foi definido com base em dois critérios conceituais. O limite inferior ($k = 5$) foi estabelecido considerando que o número de *clusters* deveria superar a divisão tradicional em quatro grandes posições (goleiros, defensores, meio-campistas e atacantes), capturando especializações funcionais mais refinadas. O limite superior ($k = 11$) corresponde ao número total de jogadores em campo por equipe, representando o nível máximo de granularidade funcionalmente justificável para fins de composição de times.

Esse intervalo é coerente com a lógica apontada por Akhanli e Hennig (2020), que ressaltam a tensão inerente entre dois objetivos concorrentes: a homogeneidade intra-*cluster*, que melhora com o aumento de k; e a separação entre *clusters*, que melhora com a redução de k. No presente estudo, privilegia-se a homogeneidade intra-*cluster*, buscando o maior número de partições possível que ainda preserve a interpretabilidade

futebolística e a viabilidade prática para composição de equipes.

3.5.3 Métricas de Validação

Dado que a clusterização é uma técnica não supervisionada, sem rótulos de referência para comparação direta, a avaliação da qualidade dos agrupamentos requer o uso de índices de validação interna, que medem características desejáveis dos *clusters* com base apenas na estrutura dos próprios dados (HENNIG, 2015a). No presente trabalho, foram utilizados dois índices complementares.

O **coeficiente de Silhueta** avalia, para cada observação, a razão entre sua distância média aos demais pontos do mesmo *cluster* (coesão interna) e sua distância média ao *cluster* vizinho mais próximo (separação externa), normalizada pelo maior dos dois valores. Varia de -1 a $+1$: valores próximos de $+1$ indicam que a observação está bem alocada e claramente separada dos demais *clusters*; valores próximos de 0 indicam sobreposição; e valores negativos sugerem possível má alocação (WIJNGAARD, 2020; VEN, 2018).

O **índice de Davies-Bouldin (DBI)** calcula, para cada *cluster*, a razão média entre sua dispersão interna e a distância ao *cluster* mais próximo, tomando a média dessas razões sobre todos os *clusters*. Valores menores indicam melhor qualidade, refletindo *clusters* internamente compactos e mutuamente bem separados. Ven (2018) observa que o DBI é particularmente sensível à homogeneidade interna, tornando-o especialmente adequado ao objetivo deste trabalho.

Ambos os índices foram utilizados como mecanismo de filtragem preliminar das soluções com desempenho claramente insatisfatório, e não como critério absoluto e suficiente para a escolha do modelo final. Conforme enfatizado por Hennig (2015a) e Akhanli e Hennig (2023), o agrupamento ideal não pode ser determinado exclusivamente pelos dados, sendo indispensável a análise interpretativa à luz do domínio de aplicação.

3.5.4 Análise Interpretativa dos Clusters

Após a filtragem quantitativa pelas métricas de validação, os agrupamentos selecionados foram submetidos a uma análise interpretativa baseada na coerência futebolística dos perfis identificados. Essa etapa consistiu em três procedimentos complementares.

O primeiro consistiu na análise das variáveis com maior poder discriminativo em cada *cluster*, aquelas cujos valores médios no grupo apresentam maior diferença em relação à média global da base, buscando identificar o padrão funcional predominante em cada agrupamento: características predominantemente defensivas, ofensivas, de construção ou de pressão.

O segundo procedimento foi a construção de projeções visuais dos agrupamentos nos espaços dos dois e três primeiros componentes principais do PCA, permitindo verificar a separação espacial dos *clusters* e identificar eventuais sobreposições.

Por fim, foram elaboradas tabelas de correspondência entre os *clusters* obtidos e as posições nominais presentes na base de dados original, abordagem adotada por García-Aliaga et al. (2021), Dağ, Yüksel e Atmaca (2023) e Haan et al. (2025) como forma de validação externa. Espera-se que jogadores de posições nominais semelhantes se concentrem predominantemente em *clusters* específicos, confirmando a capacidade do modelo de capturar estruturas funcionais reais.

A definição final do agrupamento mais adequado combinou critérios quantitativos (métricas de validação), interpretativos (coerência futebolística) e estruturais (aderência às composições medianas observadas nas equipes reais), seguindo o critério híbrido recomendado por Akhanli e Hennig (2023).

3.6 Formação de Equipes Balanceadas

A etapa final da metodologia consistiu na definição de uma estratégia algorítmica para composição de equipes balanceadas a partir dos agrupamentos obtidos. A proposta parte do pressuposto, sustentado por Ven (2018) e Penner (2025), de que equipes equilibradas não dependem apenas de níveis médios semelhantes de desempenho,

mas também da distribuição funcional equilibrada dos diferentes perfis de jogadores, representados aqui pelos *clusters* identificados.

3.6.1 Definição da Composição Estrutural das Equipes

O ponto de partida para a composição das equipes balanceadas foi a identificação de uma estrutura funcional de referência, isto é, quantos jogadores de cada perfil (*cluster*) uma equipe típica das ligas analisadas costuma ter em seu elenco principal.

Para isso, foram considerados os jogadores com maior frequência de participação em partidas por cada equipe real presente na base de dados, buscando identificar padrões médios de distribuição dos perfis funcionais entre os times das principais ligas europeias. Para cada *cluster* identificado, calculou-se a quantidade mediana de representantes presentes nas equipes reais analisadas. A mediana foi preferida à média por ser mais robusta à influência de distribuições extremas, como equipes com perfis táticos muito atípicos em relação ao conjunto.

Quando a simples soma das medianas por *cluster* não resultava exatamente em 11 jogadores, adotou-se um critério complementar: identificaram-se as equipes reais cuja distribuição de *clusters* apresentava maior aderência às medianas calculadas e, a partir desse subconjunto, determinou-se a moda das formações observadas como padrão final de composição.

3.6.2 Algoritmo de Distribuição dos Jogadores

Com a estrutura funcional de referência definida, foi desenvolvida uma estratégia algorítmica para seleção e distribuição equilibrada dos jogadores entre as equipes. O algoritmo opera em duas camadas de balanceamento: a da diversidade funcional, garantida pela exigência de representantes de todos os *clusters* previstos nessa estrutura; e a do equilíbrio técnico relativo, garantido pela restrição sobre faixas de desempenho.

Para implementar o balanceamento técnico, os atletas de cada *cluster* foram inicialmente ordenados de forma decrescente pelo *rating* médio disponibilizado pelas

plataformas de dados utilizadas, atributo de desempenho geral adotado como critério de ordenação relativa dentro de cada perfil funcional. Em seguida, os jogadores de cada *cluster* foram divididos em 11 faixas de desempenho aproximadamente equivalentes, variando das faixas superiores (jogadores mais bem avaliados) até as inferiores.

A opção por onze faixas de desempenho está relacionada com o quantitativo de atletas que compõem uma equipe de futebol completa. A lógica central consiste em garantir que cada equipe formada receba jogadores provenientes de diferentes níveis relativos de desempenho dentro de cada perfil funcional. Durante a seleção dos jogadores de determinado *cluster* para uma equipe, apenas faixas ainda não utilizadas naquela composição podem ser acionadas para os demais *clusters*.

Essa restrição faz com que os onze integrantes do time sejam provenientes de onze estratos distintos de desempenho relativo, reduzindo a possibilidade de concentração de jogadores mais fortes ou mais fracos em uma mesma equipe e promovendo um balanceamento técnico mais uniforme, sem comprometer a diversidade funcional assegurada pelos *clusters*.

O procedimento adotado está formalizado no Algoritmo 1.

Algoritmo 1: Formação de equipes balanceadas orientadas por *clusters*

Entrada: Conjunto de jogadores agrupados em *clusters*; distribuição padrão de jogadores por *cluster*; *rating* de cada jogador; número de equipes N

Saída: N equipes balanceadas

para cada *cluster* c *identificado* **faça**
  Ordenar os jogadores de c em ordem decrescente de *rating*
  Dividir os jogadores de c em 11 faixas de desempenho $F_1, F_2, \ldots, F_{11}$ aproximadamente equivalentes
para $i \leftarrow 1$ N **faça**
  $E_i \leftarrow \emptyset$
  FaixasUsadas $\leftarrow \emptyset$
  para cada *cluster* c *previsto na composição estrutural da equipe* **faça**
    Selecionar faixa $f \notin$ FaixasUsadas
    Sortear aleatoriamente jogador j pertencente a $F_f^{(c)}$ ainda disponível
    $E_i \leftarrow E_i \cup \{j\}$
    Marcar j como indisponível
    FaixasUsadas $\leftarrow$ FaixasUsadas $\cup \{f\}$
retorna $\{E_1, E_2, \ldots, E_N\}$

4

RESULTADOS E DISCUSSÃO

4.1 Construção e Caracterização da Base de Dados

O processo de integração dos *datasets* contendo estatísticas de desempenho de jogadores das cinco principais ligas europeias na temporada 2024/2025, seguido da remoção de registros duplicados conforme os critérios descritos na metodologia, resultou inicialmente em uma base de dados composta por aproximadamente 2.705 jogadores e mais de uma centena de variáveis explicativas.

Em seguida, foi aplicado o critério de corte por minutagem mínima (450 minutos), com o objetivo de excluir atletas com baixa participação competitiva e reduzir a presença de ruídos estatísticos capazes de comprometer a qualidade da clusterização. Também foram removidas variáveis não relevantes para o agrupamento, como as relacionadas à identificação individual dos atletas ou a aspectos subjetivos e coletivos do jogo.

Após esse tratamento, a base final utilizada no processo de clusterização passou a conter 1.956 registros de jogadores e 98 variáveis numéricas, correspondendo a uma redução de aproximadamente 28% em relação à base original.

Embora essa etapa tenha reduzido significativamente o volume de dados, a base resultante ainda apresentava elevada dimensionalidade, com um número expressivo de atributos potencialmente redundantes. Em bases de estatísticas esportivas, é comum que diferentes atributos descrevam aspectos semelhantes do desempenho dos atletas, produzindo fortes relações de dependência entre as variáveis.

A Figura 1 apresenta o mapa de calor da matriz de correlação das variáveis utilizadas na etapa de modelagem.

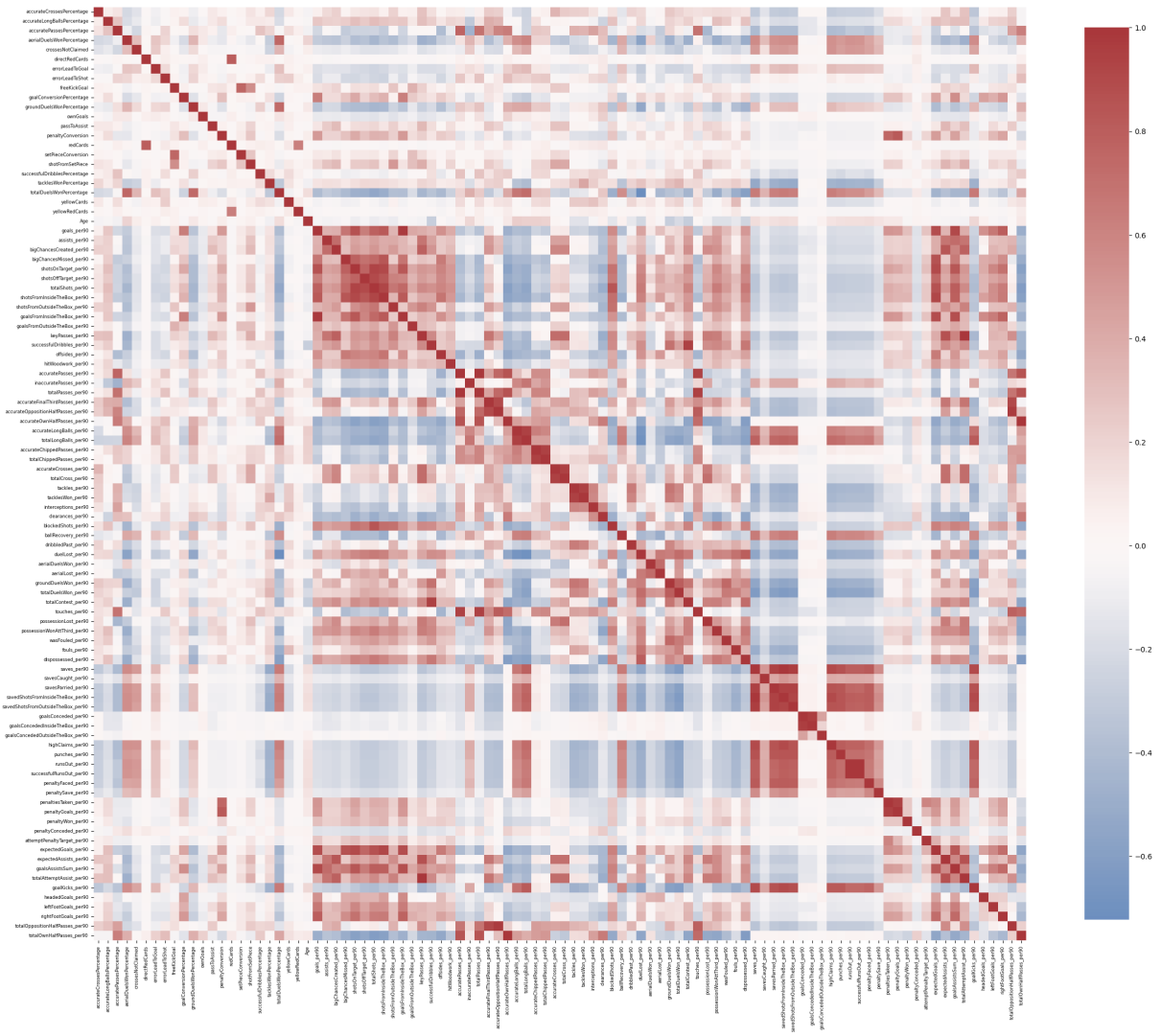

Figura 1 – Mapa de calor da matriz de correlação das variáveis da base de dados após o pré-processamento.

A figura mostra extensas regiões com elevada correlação positiva (tons em vermelho) e negativa (tons em azul), evidenciando a presença de grupos de atributos que carregam informações semelhantes ou proporcionais. Esse comportamento caracteriza um cenário típico de alta multicolinearidade, no qual a utilização direta de todas as variáveis pode introduzir redundâncias, aumentar o custo computacional e dificultar a identificação dos padrões mais relevantes para os algoritmos de clusterização.

Diante desse cenário, tornou-se necessária a aplicação de uma técnica de redução de dimensionalidade capaz de eliminar redundâncias preservando a maior parte da variabilidade presente nos dados. Para esse fim, foi empregada a Análise de Componentes Principais (PCA), cujos resultados são apresentados na seção seguinte.

4.2 Redução de Dimensionalidade via PCA

Sobre os dados padronizados, procedeu-se à aplicação do PCA. Em um primeiro momento, foi construído o gráfico de variância acumulada em função do número de componentes principais (*scree plot*), buscando identificar visualmente um ponto de inflexão da curva pelo método do cotovelo.

Entretanto, a curva apresentou comportamento bastante suave, sem um ponto de inflexão claramente identificável, comportamento esperado em bases com muitas variáveis correlacionadas (PINO-ORTEGA et al., 2021) e que tornaria a determinação do número de componentes por inspeção visual excessivamente arbitrária. A Figura 2 apresenta o *scree plot* resultante.

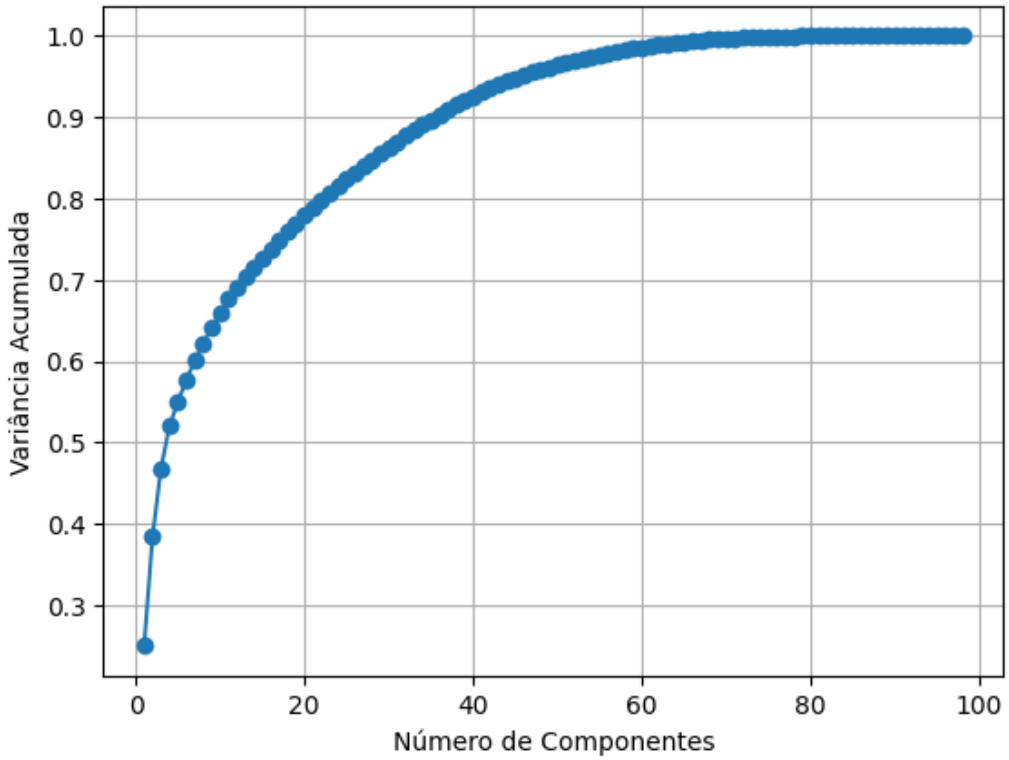

Figura 2 – Curva de Variância Acumulada.

Diante da ausência de cotovelo visual claro, adotou-se o critério de Kaiser como método complementar, conforme previsto na metodologia (Capítulo 3). O critério de Kaiser estabelece que devem ser mantidos apenas os componentes cujos autovalores sejam superiores a 1, uma vez que tais componentes explicam uma quantidade de variância superior àquela originalmente atribuída a uma variável padronizada individual,

ou seja, componentes com autovalores inferiores a 1 contribuiriam menos informação do que uma variável original isolada, não justificando sua manutenção no modelo reduzido.

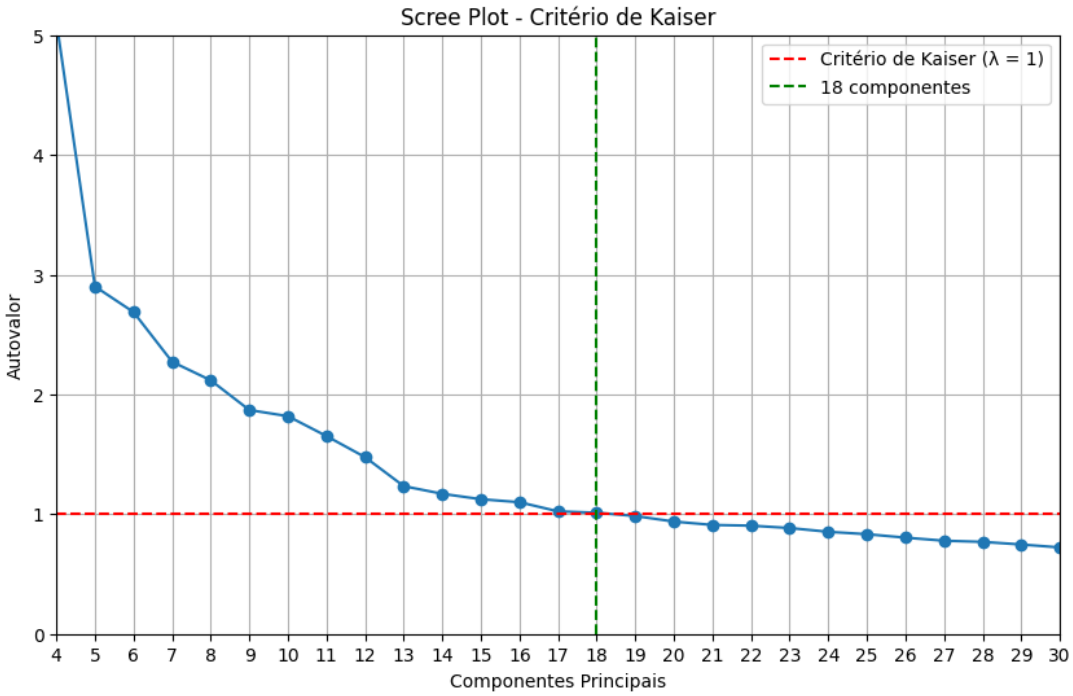

Figura 3 – Critério de Kaiser.

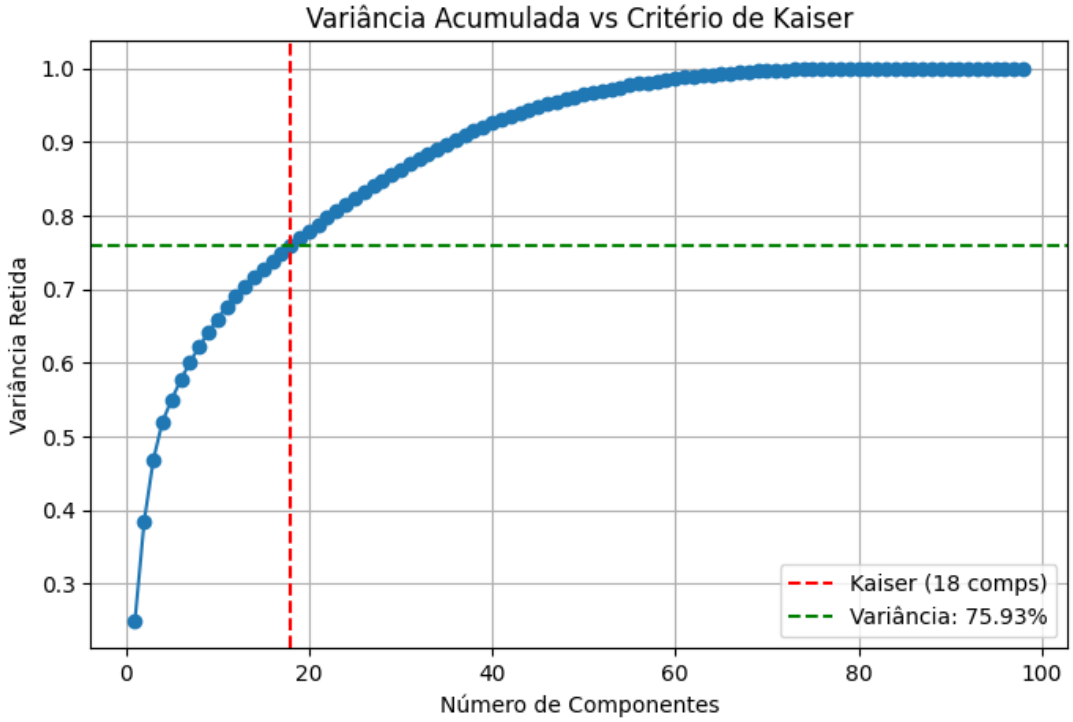

Figura 4 – Curva de Variância Acumulada com Critério de Kaiser.

Conforme demonstrado nas Figuras 3 e 4, a aplicação do critério de Kaiser indicou a retenção de 18 componentes principais, correspondentes ao último componente com autovalor superior a 1, responsáveis por preservar aproximadamente 75,93% da variância total dos dados originais. Esse percentual situa-se dentro da faixa de 70% a 85% frequentemente reportada como adequada em estudos de *sports analytics* (ROJAS-VALVERDE et al., 2020; PINO-ORTEGA et al., 2021). A transformação representou uma redução de aproximadamente 82% na dimensionalidade original, eliminando redundâncias e contribuindo para a estabilidade dos agrupamentos.

4.3 Avaliação dos Agrupamentos: Métricas de Validação

Após a projeção dos dados no espaço de 18 componentes principais, procedeu-se à avaliação da qualidade interna dos agrupamentos produzidos pelos algoritmos K-Means e Hierárquico com Ward para soluções variando entre 5 e 11 *clusters*, utilizando o coeficiente de Silhueta e o índice de Davies-Bouldin como critérios de filtragem preliminar.

Lembrando que o índice de Davies-Bouldin mede simultaneamente a compactação interna dos grupos e a separação entre *clusters* distintos: valores menores indicam agrupamentos mais eficientes, caracterizados por *clusters* internamente homogêneos e externamente bem separados. Já o coeficiente de Silhueta avalia o grau de similaridade entre os elementos de um mesmo *cluster* em comparação com os elementos dos demais grupos: valores mais elevados, na escala de -1 a $+1$, indicam maior coesão interna e melhor separação entre agrupamentos.

A seleção dessas duas métricas é diretamente orientada pelo objetivo do trabalho. O índice de Davies-Bouldin (DBI) foi preferido por sua maior sensibilidade à homogeneidade interna dos agrupamentos, propriedade essencial para a identificação de perfis funcionais coesos. O coeficiente de Silhueta, amplamente adotado na literatura de clusterização aplicada ao futebol, complementa essa avaliação ao considerar de maneira mais equilibrada a coesão interna com a separação entre *clusters*. A combinação dos índices busca então cobrir perspectivas complementares da qualidade dos agrupamentos.

Os gráficos comparativos dos dois índices para os dois algoritmos são apresentados nas Figuras 5 e 6.

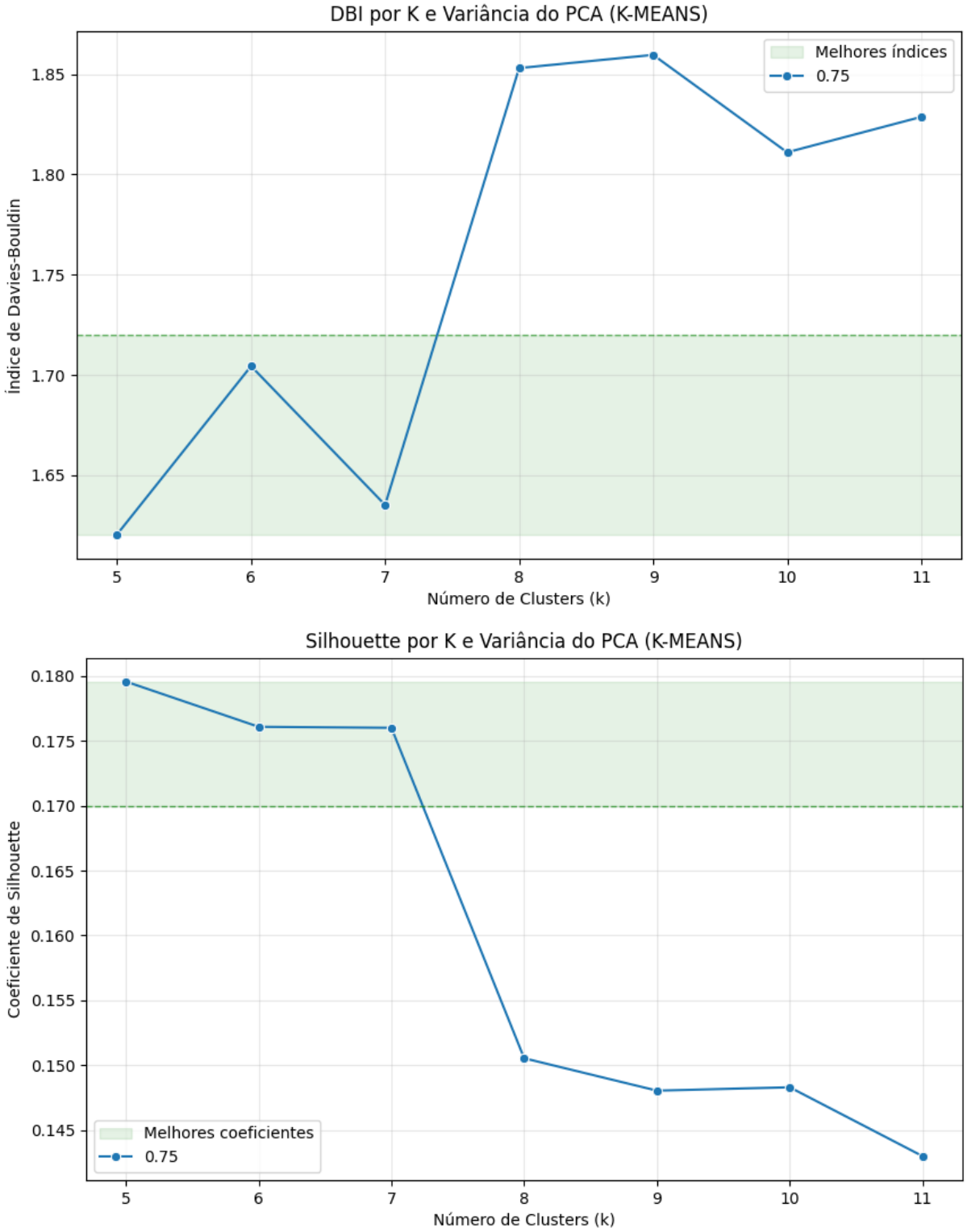

Figura 5 – Comparação dos índices de validação interna para o algoritmo K-Means, avaliados para k entre 5 e 11 *clusters*.

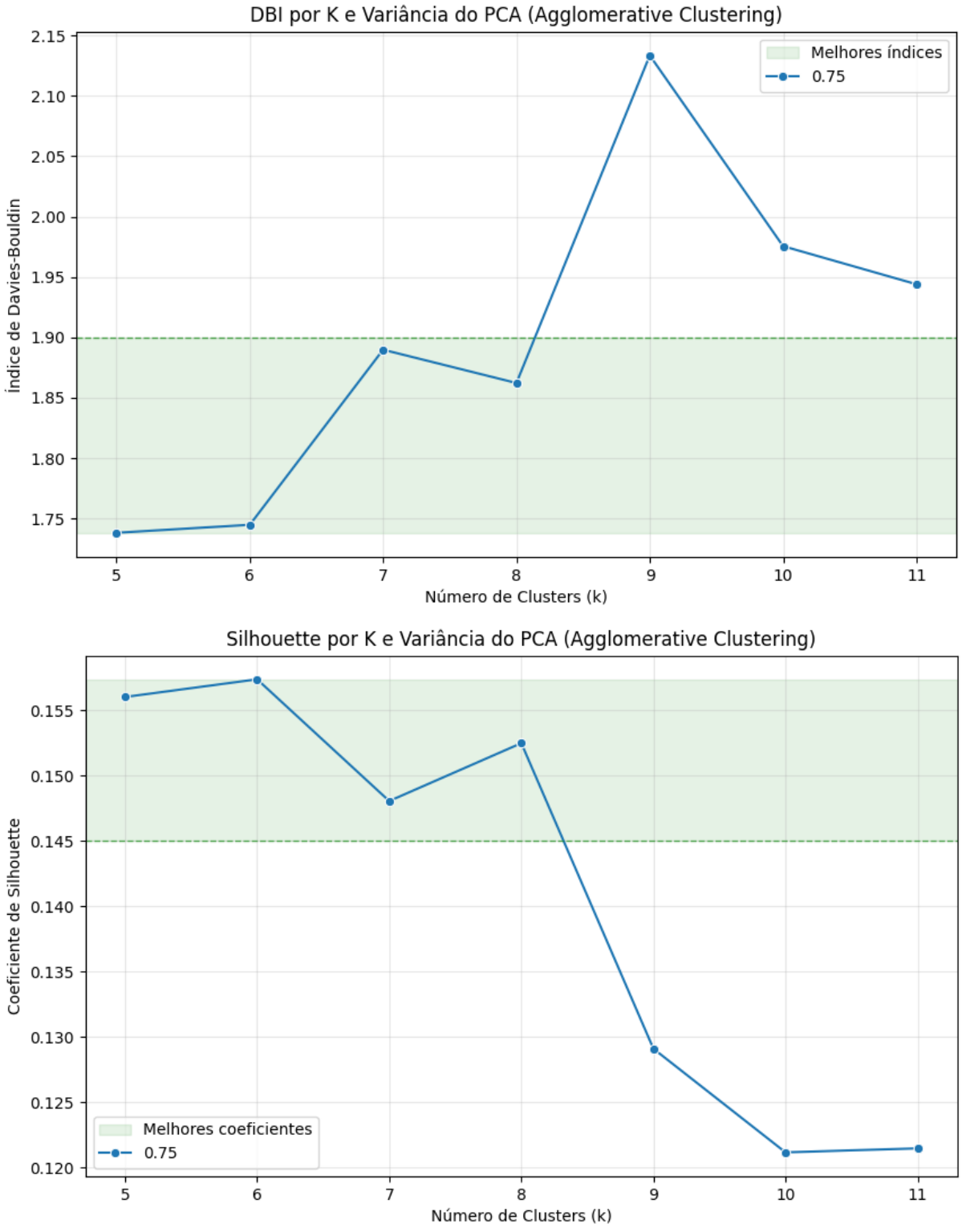

Figura 6 – Comparação dos índices de validação interna para o algoritmo Hierárquico com Ward, avaliados para k entre 5 e 11 *clusters*.

Conforme evidenciado pelos gráficos, os melhores resultados para o algoritmo K-Means concentraram-se nos agrupamentos com 5, 6 e 7 *clusters*, enquanto no método de ligação de Ward as soluções com 5, 6, 7 e 8 *clusters* destacaram-se positivamente

em ambos os índices. Com base nesses resultados, as análises subsequentes foram concentradas exatamente nessas soluções pré-selecionadas: $k \in \{5, 6, 7\}$ para o K-Means e $k \in \{5, 6, 7, 8\}$ para o método de ligação de Ward.

Ressalta-se, entretanto, que os índices foram utilizados apenas como mecanismo preliminar de exclusão de soluções claramente inadequadas, e não como critério absoluto para a escolha do modelo final, em consonância com a perspectiva de Hennig (2015a), segundo a qual o agrupamento ideal não pode ser determinado exclusivamente pelos dados, mas exigindo que a escolha da melhor partição incorpore a experiência e a avaliação crítica do pesquisador sobre o fenômeno analisado.

Nessa esteira, a definição do agrupamento mais adequado fundamentar-se-á em um critério híbrido, combinando: qualidade estatística interna dos agrupamentos; coerência futebolística dos perfis encontrados; capacidade de diferenciação tática; e aplicabilidade prática para composição automática de equipes balanceadas.

4.4 Análise Interpretativa dos Agrupamentos

4.4.1 Agrupamentos pelo Método Hierárquico com Ligação de Ward

As soluções produzidas pelo método de ligação de Ward evidenciaram um processo progressivo e coerente de refinamento dos perfis táticos à medida que o número de *clusters* aumentava, comportamento característico dos métodos hierárquicos e uma de suas principais vantagens para fins exploratórios (RAN et al., 2023).

O dendrograma resultante é apresentado na Figura 7, com indicação das linhas de corte para os valores de k avaliados. Observa-se que os cortes selecionados recaem em regiões de maior altura no dendrograma, isto é, onde os saltos de variância entre fusões sucessivas são mais expressivos, o que indica que os grupos formados nessas partições são bem distintos entre si. Essa convergência entre a evidência visual do dendrograma e os resultados dos índices de validação reforça a consistência das soluções pré-selecionadas.

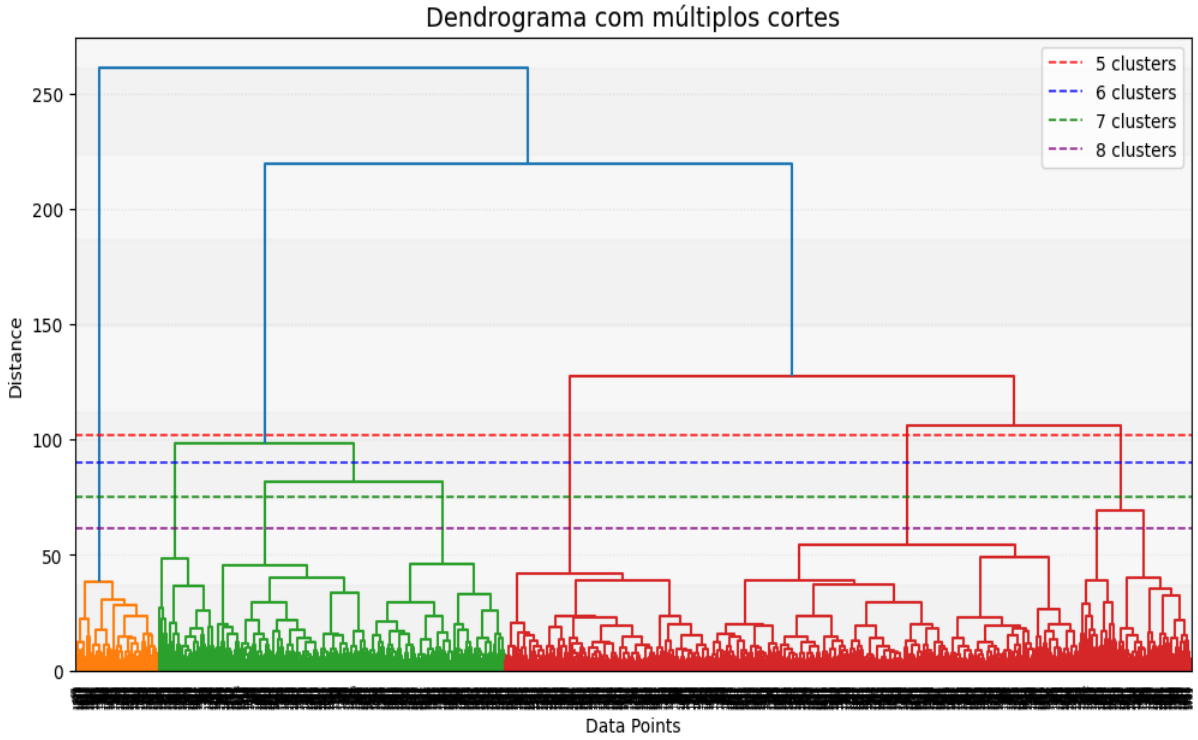

Figura 7 – Dendrograma da clusterização hierárquica com método de Ward.

A interpretação de cada solução foi conduzida com base na análise das variáveis de maior relevância estatística em cada *cluster*, aquelas cujos valores médios no grupo apresentam maior afastamento em relação à média global da base, complementada por projeções visuais bidimensionais e tridimensionais dos agrupamentos no espaço do PCA e por tabelas de correspondência com as posições nominais da base de dados.

A tabela 1 sintetiza a evolução estrutural dos agrupamentos à medida que o número de partições (k) aumenta.

Tabela 1 – Evolução estrutural dos agrupamentos pelo método de ligação de Ward

Transição	**Refinamento identificado**
$5 \rightarrow 6$ clusters	Separação entre construção ofensiva e finalização
$6 \rightarrow 7$ clusters	Separação entre atacantes associativos e atacantes verticais
$7 \rightarrow 8$ clusters	Separação entre meias ofensivos e armadores de elite

Na solução com cinco *clusters*, o algoritmo identificou grandes macrofunções do futebol moderno, separando: goleiros; defensores construtores; jogadores de meio-campo equilibrados; criadores pelos corredores laterais; e atacantes móveis. Essa partição representa uma estrutura funcional ampla, mas internamente coesa, na qual as

diferenças entre os perfis são claramente perceptíveis nas variáveis de construção, pressão, finalização e posse de bola.

Com seis *clusters*, observou-se um refinamento importante da estrutura ofensiva. O algoritmo passou a distinguir jogadores ofensivos voltados predominantemente à finalização daqueles com maior participação associativa e construtiva, separação evidenciada pela divergência nos valores médios de atributos como finalizações por 90 minutos, *xG* e criação de grandes chances. Tal divisão revelou uma separação coerente entre atletas responsáveis pela conclusão das jogadas e jogadores especializados na progressão e criação de oportunidades.

A solução com 7 *clusters* aprofundou ainda mais essa especialização, permitindo distinguir atacantes associativos e criativos de atacantes mais verticais e físicos. Essa configuração apresentou alta coerência tática e maior granularidade funcional sem comprometer a estabilidade estrutural dos agrupamentos, o que, combinado aos resultados das métricas de validação, tornou tal solução a mais equilibrada dentre as avaliadas.

Por fim, a solução com oito *clusters* produziu o maior nível de refinamento observado no estudo, separando inclusive jogadores ofensivos extremamente agressivos e verticais de atletas técnicos voltados à circulação e sustentação da posse ofensiva.

Embora esse refinamento tenha gerado agrupamentos altamente especializados e com elevada coerência interpretativa, observou-se também o surgimento de um *cluster* excessivamente específico e de dimensão reduzida, com frequência mediana nula nas equipes reais analisadas, comprometendo sua viabilidade prática para composição equilibrada de equipes.

De maneira geral, o método de ligação de Ward demonstrou elevada capacidade de captar dimensões estruturais clássicas do futebol contemporâneo, produzindo agrupamentos compatíveis com funções reais observadas em equipes profissionais. Os agrupamentos defensivos apresentaram separações claras entre goleiros, zagueiros físicos e defensores construtores.

Os agrupamentos intermediários permitiram distinguir jogadores de circulação, organizadores e atletas híbridos de corredor. Os agrupamentos ofensivos revelaram diferentes modos de contribuição, incluindo jogadores criativos, progressivos, associativos

e finalizadores verticais. Essa capacidade de refinamento progressivo representou uma das principais vantagens do método hierárquico em relação ao K-Means, sobretudo por permitir acompanhar o surgimento gradual de subperfis táticos conforme o aumento do número de *clusters*.

4.4.2 Agrupamentos pelo Algoritmo K-Means

As soluções produzidas pelo K-Means apresentaram comportamento estrutural semelhante ao obtido pelo método de ligação de Ward, embora com menor refinamento hierárquico dos perfis e maior tendência a separar os jogadores conforme padrões globais de comportamento estatístico, em vez de dividir progressivamente macrofunções.

Na solução com cinco *clusters*, o algoritmo separou os jogadores em grandes macrofunções, distinguindo: atacantes finalizadores; defensores construtores; goleiros; criadores ofensivos; e jogadores híbridos de corredor.

Com seis agrupamentos, o modelo passou a representar uma estrutura funcional bastante próxima da organização real das equipes contemporâneas, identificando, além dos goleiros: defensores de contenção; laterais de apoio; criadores ofensivos; construtores equilibrados; e atacantes móveis e verticais.

Já na solução com sete *clusters*, observou-se maior especialização dos perfis ofensivos e de circulação, embora com menor clareza hierárquica em comparação ao método de ligação de Ward.

A tabela 2 sintetiza os principais perfis encontrados nas três soluções K-means pré-selecionadas pelos índices de validação.

Tabela 2 – Resumo interpretativo dos agrupamentos pelo algoritmo K-Means.

Agrupamento	**Principais perfis encontrados**	**Característica predominante**
K-Means $k = 5$	Finalizadores; defensores construtores; goleiros; criadores; híbridos	Macrofunções equilibradas
K-Means $k = 6$	Defensores de contenção; laterais; criadores; organizadores; atacantes móveis; goleiros	Estrutura funcional moderna
K-Means $k = 7$	Maior especialização nos perfis de função ofensiva e de circulação	Refinamento intermediário

4.5 Análise das Formações Medianas e Definição do Modelo Final de Referência

Além das métricas estatísticas e da interpretação futebolística, este trabalho incorporou um terceiro critério de avaliação: a aderência estrutural dos *clusters* às formações medianas observadas nas equipes reais. Para os casos em que a busca pela formação mediana não apontasse para uma equipe completa de 11 jogadores, estabeleceu-se o critério subsidiário de buscar a formação mais comum (moda) entre as formações reais mais próximas da mediana.

O objetivo dessa etapa consistiu em verificar se os agrupamentos encontrados eram compatíveis com composições funcionais plausíveis dentro de uma equipe de futebol, condição necessária para sua utilização prática no sistema proposto.

Ao adotar essa sistemática, os resultados confirmaram a possibilidade de se estabelecer, para a base de dados estudada, a formação completa de uma equipe (de 11 jogadores) a partir do quantitativo mediano de jogadores por *cluster* para todas as quatro partições selecionadas da clusterização hierárquica (método de Ward), assim como para o agrupamento de 5 *clusters* do algoritmo K-Means. Para as partições com 6 e 7 *clusters* do K-Means, foi necessário recorrer ao critério subsidiário da moda das formações próximas da mediana para se atingir a composição de referência das equipes.

A tabela 3 apresenta a distribuição mediana (ou mediana com moda) de jogadores por *cluster* para cada solução avaliada.

Tabela 3 – Distribuição mediana de jogadores por *cluster* nas equipes reais.

Modelo	**Distribuição mediana (ou mediana com moda) dos *clusters***
Ward $k = 5$	C1=1; C2=3; C3=2; C4=4; C5=1
Ward $k = 6$	C1=1; C2=1; C3=2; C4=2; C5=4; C6=1
Ward $k = 7$	C1=1; C2=1; C3=1; C4=1; C5=2; C6=4; C7=1
Ward $k = 8$	C1=1; C2=1; C3=1; C4=1; C5=2; C6=4; C7=0; C8=1
K-Means $k = 5$	C0=1; C1=3; C2=1; C3=4; C4=2
K-Means $k = 6$	C0=3; C1=1; C2=1; C3=1; C4=4; C5=1
K-Means $k = 7$	C0=1; C1=1; C2=1; C3=2; C4=1; C5=2; C6=3

A análise dessas distribuições revelou um aspecto metodológico fundamental: embora o aumento do número de *clusters* favoreça maior especialização e refinamento

dos perfis táticos, soluções excessivamente detalhadas podem produzir agrupamentos pouco representativos dentro das estruturas reais de composição das equipes.

Esse fenômeno tornou-se particularmente evidente na solução Ward com oito *clusters*: o Cluster 7, associado aos jogadores ofensivos mais agressivos e verticais, não apresentou representatividade dentro das formações medianas identificadas, resultando em frequência nula na composição típica das equipes analisadas. Tal comportamento sugere que esse agrupamento, embora relevante para aplicações específicas de *scouting* e recrutamento, mostrou-se excessivamente especializado para utilização como categoria estrutural obrigatória num sistema de montagem equilibrada de equipes.

Por outro lado, soluções com cinco *clusters* apresentaram segmentações excessivamente amplas, reduzindo a capacidade de diferenciação funcional dos jogadores. As soluções intermediárias, particularmente Ward com seis e sete *clusters*, apresentaram melhor equilíbrio entre granularidade tática, estabilidade estrutural, interpretabilidade e aderência às composições reais das equipes.

Entre essas, o agrupamento hierárquico de Ward com sete *clusters* apresentou o melhor compromisso geral entre refinamento funcional e aplicabilidade prática, permitindo representar adequadamente diferentes funções táticas sem gerar *clusters* excessivamente específicos ou estruturalmente inviáveis. Dessa forma, a solução Ward para $k = 7$ foi adotada como modelo de referência para análise e discussão dos resultados.

4.6 Perfis Funcionais do Modelo de Referência (Ward k=7)

A partir da interpretação estatística e futebolística dos agrupamentos da solução Ward para $k = 7$, foi possível atribuir denominações funcionais representativas para cada *cluster*. Essas denominações foram estabelecidas com base na análise das variáveis mais discriminativas de cada grupo, aquelas que mais se afastam da média global da base, e também com base na correspondência com as posições nominais predominantes entre os jogadores alocados em cada agrupamento.

A tabela 4 apresenta os perfis identificados e a Figura 8 permite visualizar com clareza tanto o grau de proximidade quanto de especialização desse perfis.

Tabela 4 – Nomeação funcional dos *clusters* da solução Ward para $k = 7$.

Cluster	**Perfil / Função representativa**
Cluster 1	Goleiros
Cluster 2	Atacantes finalizadores
Cluster 3	Criadores ofensivos / Pontas
Cluster 4	Atacantes móveis / Híbridos ofensivos
Cluster 5	Zagueiros / Defensores de contenção
Cluster 6	Meias e Defensores construtores
Cluster 7	Meias organizadores e de circulação

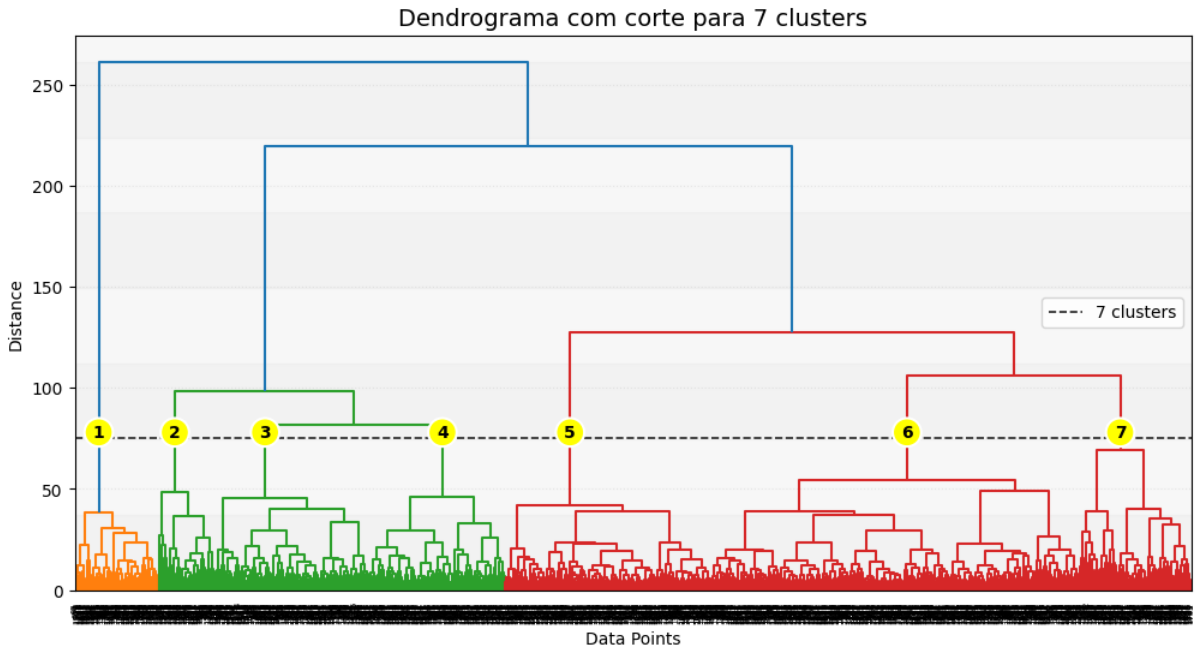

Figura 8 – Dendrograma do método de *Ward* com projeção visual para o agrupamento com 7 *clusters*.

Os sete perfis identificados representam de forma relativamente completa as principais macrofunções táticas do futebol contemporâneo: proteção defensiva; construção desde a defesa; circulação e sustentação da posse; progressão pelos corredores; criação ofensiva; atacantes de ruptura; e finalização vertical. Essa estrutura reforça a coerência futebolística da solução escolhida e sua adequação ao objetivo central do trabalho.

A tabela 5 apresenta a correspondência entre os *clusters* e as posições nominais dos jogadores na base de dados original, permitindo avaliar o grau de alinhamento entre os agrupamentos obtidos e os papéis posicionais tradicionais, sem que o algoritmo tenha tido acesso a essa informação durante o processo de agrupamento.

Tabela 5 – Correspondência entre *clusters* (Ward k=7) e posições nominais de jogadores.

Posição	**C1**	**C2**	**C3**	**C4**	**C5**	**C6**	**C7**
GK (goleiro)	100%	–	–	–	–	–	–
DF (defensor)	–	–	0,4%	–	93,2%	26%	10,4%
DF,MF (meia/defensor)	–	–	2,4%	2,3%	5,3%	23,9%	19,2%
MF (meio-campo)	–	6,3%	39,9%	17,5%	1,5%	47,8%	58,8%
MF,FW (meia/atacante)	–	23,5%	54%	35,5%	–	2,3%	11%
FW (atacante)	–	70,2%	3,3%	44,7%	–	–	0,6%

4.7 Comparação entre as soluções dos algoritmos Hierárquico com Ward e K-Means

A análise comparativa entre os métodos demonstrou elevada convergência estrutural na identificação dos perfis funcionais, em especial considerando os respectivos agrupamentos gerados para a partição de referência adotada (de 7 clusters).

A distribuição espacial dos clusters e a identificação dos respectivos pontos médios ou centróides, conforme as projeções apresentadas nas Figuras 9 e 10, permitem conferir essa grande proximidade entre os resultados.

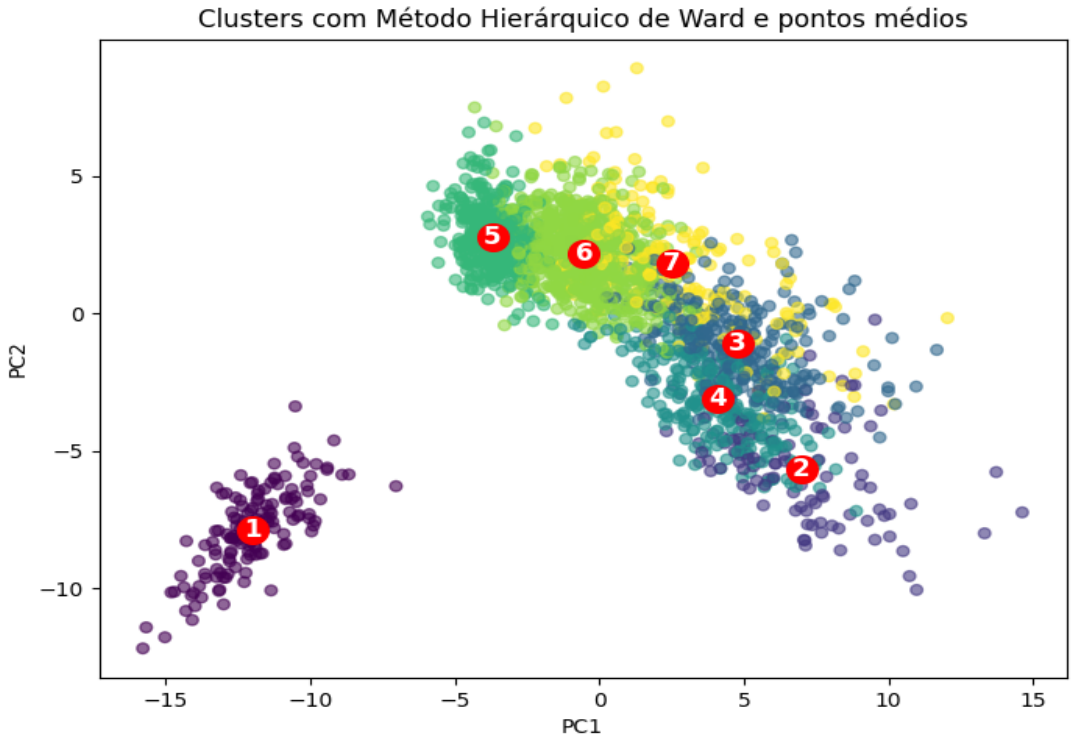

Figura 9 – Distribuição espacial 2D dos agrupamentos da solução Ward $k = 7$.

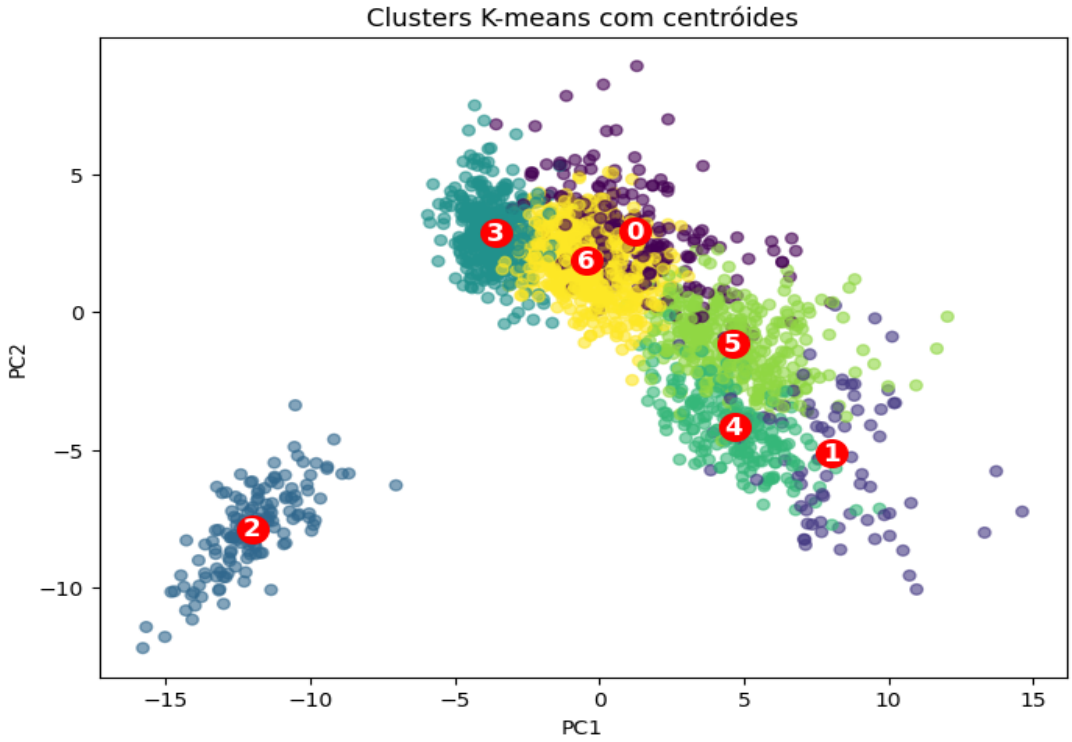

Figura 10 – Distribuição espacial 2D dos agrupamentos da solução K-Means $k = 7$.

Apesar das diferenças metodológicas entre os algoritmos, ambos apresentaram capacidade consistente de separar grandes grupos táticos relacionados às funções desempenhadas em campo, como goleiros, defensores estruturais, jogadores de sustentação e circulação, criadores de jogadas, jogadores de ruptura e finalizadores.

Inclusive, é visualmente fácil identificar e consistentemente possível mapear a correspondência aproximada entre os grupos formados por cada um dos algoritmos. O mapeamento entre as soluções, obtido por inspeção dos gráficos de dispersão, pode ser assim determinado:

- Cluster 1 Ward (goleiros) → Cluster 2 K-Means;
- Cluster 5 Ward (zagueiros e defensores de contenção) → Cluster 3 K-Means;
- Cluster 6 Ward (meias e defensores construtores) → Cluster 6 K-Means;
- Cluster 7 Ward (meias organizadores e de circulação) → Cluster 0 K-Means;
- Cluster 3 Ward (criadores ofensivos e pontas) → Cluster 5 K-Means;
- Cluster 4 Ward (atacantes móveis e híbridos ofensivos) → Cluster 4 K-Means;
- Cluster 2 Ward (atacantes finalizadores) → Cluster 2 K-Means.

Esse alinhamento entre as soluções evidencia que ambos os algoritmos convergiram para estruturas funcionais equivalentes, ainda que com numerações distintas para os *clusters*. A tabela 6 apresenta a comparação detalhada entre os métodos e respectivos perfis identificados na solução com sete *clusters*.

Tabela 6 – Comparação entre os perfis identificados pelos métodos de Ward e K-Means.

Ward	**N**	**K-Means**	**N**	**Principais diferenças**
Goleiros	147	Goleiros	147	Separação estável em ambos
Zagueiros estruturais	362	Defensores estruturais	395	K-Means incorporou defensores híbridos e laterais mais físicos
Volantes organizadores	645	Conectores de sustentação	576	K-Means subdividiu em dois subgrupos distintos
Executores ofensivos	101	Executores de elite	79	K-Means isolou mais agressivamente os de maior produção
Desequilibradores criativos	268	Geradores ofensivos	353	K-Means enfatizou mais o volume de criação
Atacantes de ruptura	238	Atacantes de ruptura	200	Elevada convergência entre os métodos
Construtores dinâmicos	195	Orquestradores de posse	206	Ward mais hierárquico; K-Means distribuiu entre subgrupos

As maiores diferenças entre os métodos ocorreram principalmente na forma de organizar os jogadores responsáveis pela construção ofensiva e pela circulação de bola. Enquanto o método de ligação de Ward apresentou uma estrutura mais hierárquica e interpretativa, agrupando inicialmente grandes macrofunções que progressivamente se subdividem, o K-Means mostrou maior tendência a separar os jogadores conforme padrões globais de comportamento estatístico, aproximando jogadores de funções híbridas e ecossistemas ofensivos semelhantes.

Essa diferença ficou especialmente evidente nos grupos relacionados à criação e progressão ofensiva. O método de ligação de Ward preservou relações conceituais mais claras entre os perfis, permitindo compreender a evolução funcional dos agrupamentos ao longo da árvore hierárquica.

O K-Means, por sua vez, apresentou agrupamentos mais orientados ao comportamento prático em campo. Ainda assim, a forte convergência estrutural observada entre os dois métodos reforça a robustez dos perfis encontrados e sugere que as variáveis utilizadas foram capazes de capturar dimensões relevantes do comportamento tático dos jogadores.

4.8 Justificativas da Escolha do Método de Referência e Considerações para Implementação

Com base no conjunto de evidências apresentadas, métricas de validação interna, interpretação futebolística, aderência às formações medianas reais e análise comparativa entre os métodos, o presente trabalho adota a clusterização hierárquica com método de ligação de Ward (*Ward's linkage*) e sete *clusters* (Ward $k = 7$) como modelo de referência para análise e discussão dos resultados.

Essa escolha fundamenta-se em três razões. Primeiramente, o método de ligação de Ward demonstrou superior capacidade interpretativa: sua estrutura hierárquica permite compreender como os perfis funcionais emergem progressivamente a partir de grandes macrofunções do jogo, produzindo uma espécie de taxonomia tática dos jogadores (HENNIG, 2015a; AKHANLI; HENNIG, 2023).

Em segundo lugar, o dendrograma (vide Figura 7) documenta visualmente a lógica das fusões sucessivas, facilitando a compreensão e a justificativa das escolhas metodológicas (GERE, 2023; RAN et al., 2023). Em terceiro lugar, a solução *Ward* $k = 7$ demonstrou aderência satisfatória às composições funcionais das equipes reais, com todos os 7 *clusters* apresentando representatividade nas formações medianas analisadas.

Não obstante essas vantagens analíticas, reconhece-se que, em um contexto de implementação em produção, o algoritmo K-Means apresenta vantagens práticas relevantes. O algoritmo possui menor complexidade computacional e maior escalabilidade, tornando-o mais adequado para aplicações interativas e em tempo real.

Além disso, o K-Means permite a atribuição dinâmica de novos jogadores a *clusters* já estabelecidos sem necessidade de reprocessar toda a estrutura hierárquica, o que é particularmente relevante em sistemas que precisam incorporar novos atletas de forma incremental. Outro ponto é que, nas métricas de validação, o K-Means apresentou resultados ligeiramente superiores, e os dois métodos convergiram para estruturas funcionais muito semelhantes para sete *clusters*, como bem evidencia a Tabela 6, indicando robustez dos padrões identificados.

Conclui-se que os dois métodos desempenham papéis complementares: o método de ligação de Ward é mais eficiente para a interpretação conceitual e para a

construção de uma taxonomia funcional dos jogadores com fins acadêmicos; enquanto o K-Means demonstra maior adequação para a operacionalização prática de um sistema automatizado de balanceamento de equipes. Recomenda-se que implementações futuras voltadas à produção utilizem o K-Means com sete *clusters* como mecanismo principal, utilizando os resultados do método de ligação de Ward como suporte interpretativo e explicativo dos perfis.

4.9 Aplicação dos Agrupamentos na Formação de Equipes Balanceadas

A partir dos resultados do método de clusterização definido como referência (Hierárquico com Ward para k=7), aplicou-se o Algoritmo 1 descrito no Capítulo 3 para a geração de equipes balanceadas, usando como base para a composição integral de cada equipe a distribuição mediana dos jogadores por *cluster* nas equipes reais constantes da base de dados, conforme indicado na Tabela 3, com destaque abaixo:

- Cluster 1 (goleiro) = 1 componente;
- Cluster 2 (atacante finalizador) = 1 componente;
- Cluster 3 (criador/ponta) = 1 componente;
- Cluster 4 (atacante híbrido) = 1 componente;
- Cluster 5 (zagueiros) = 2 componentes;
- Cluster 6 (meias/volantes) = 4 componentes;
- Cluster 7 (meia organizador) = 1 componente;

Essa composição é consistente com formações táticas modernas como o 4-3-3 e suas variantes, frequentemente observadas nas ligas de futebol avaliadas. A lógica de distribuição por faixas de desempenho, combinada com a restrição de não repetição de faixas ordinais dentro de uma mesma equipe, garantiu simultaneamente: (i) diversidade funcional, pela presença de representantes de todos os *clusters* na composição de cada time; e (ii) equilíbrio técnico global, pela distribuição de jogadores de diferentes estratos de desempenho relativo entre as equipes.

A estratégia buscou assegurar que nenhum time concentrasse simultaneamente jogadores dos estratos superiores de desempenho em múltiplos perfis funcionais, promovendo uma distribuição mais homogênea dos níveis técnicos globais.

A Figura 11 apresenta exemplos de equipes geradas pelo sistema, com indicação do *cluster* funcional e da faixa de desempenho (*'stratum'*) de cada jogador selecionado dentro do respectivo *cluster*.

Equipe Balanceada 1

player	team	Age	Pos	rating	cluster	stratum
Diego Conde	Villarreal	25	GK	6.91	1	8
André Silva	SV Werder Bremen	28	FW	6.75	2	10
Matías Fernández	Lille	24	UD	6.86	3	7
Esteban Lepaul	Angers	24	FW	6.88	4	4
Mickael Nade	Saint-Étienne	25	DF	6.91	5	6
José Luis Palomino	Cagliari	34	DF	6.78	5	9
Ben White	Arsenal	26	DF	6.96	6	3
Jack Taylor	Ipswich Town	26	MF	6.63	6	11
Frédéric Guilbert	Lecce	29	DF	6.86	6	5
Frank Anguissa	Napoli	25	UD	7.07	6	1
Declan Rice	Arsenal	25	MF	7.36	7	2

Equipe Balanceada 2

player	team	Age	Pos	rating	cluster	stratum
Nediljko Labrović	FC Augsburg	24	GK	6.87	1	9
Kylian Mbappé	Real Madrid	25	FW	7.7	2	1
Álex Sola	Getafe	25	MF	6.53	3	11
Mostafa Mohamed	Nantes	26	FW	6.8	4	5
Cédric Kipré	Stade de Reims	27	DF	6.98	5	4
Patrick Mainka	1. FC Heidenheim	29	DF	7.09	5	2
Niklas Dorsch	1. FC Heidenheim	26	MF	6.83	6	6
Amir Murillo	Olympique de Marseille	24	UD	6.97	6	3
Maxim Leitsch	1. FSV Mainz 05	26	DF	6.74	6	8
Amadou Haidara	RB Leipzig	26	MF	6.79	6	7
Jordan Teze	AS Monaco	24	DF,MF	6.83	7	10

Equipe Balanceada 3

player	team	Age	Pos	rating	cluster	stratum
Manuel Neuer	FC Bayern München	38	GK	6.97	1	7
Mohamed Salah	Liverpool	32	MF	7.78	2	1
Jon Guridi	Deportivo Alavés	29	MF	6.73	3	9
Bobby Decordova-Reid	Leicester City	25	UD	6.55	4	11
Leandro Cabrera	Espanyol	33	DF	7	5	4
Cameron Burgess	Ipswich Town	28	DF	6.89	5	6
Javi Galán	Atlético Madrid	29	DF	6.85	6	5
Jack Harrison	Everton	27	FW,MF	6.68	6	10
Alex Král	Espanyol	26	MF	6.74	6	8
Yangel Herrera	Girona FC	26	MF	7	6	2
Alexis Mac Allister	Liverpool	25	MF	7.24	7	3

Equipe Balanceada 4

player	team	Age	Pos	rating	cluster	stratum
Emiliano Martínez	Aston Villa	31	GK	7.07	1	4
Romelu Lukaku	Napoli	31	FW	6.99	2	6
Savinho	Manchester City	25	UD	7.31	3	1
Sirlord Conteh	1. FC Heidenheim	28	FW,MF	6.67	4	9
Diego Carlos	Aston Villa	31	DF	6.64	5	11
Julien Le Cardinal	Stade Brestois	26	DF	7.04	5	3
Stefano Sabelli	Genoa	31	DF,MF	6.75	6	8
Patrick Ciurria	Monza	29	MF	6.79	6	7
Adrià Pedrosa	Sevilla	26	DF	6.85	6	5
Ander Guevara	Deportivo Alavés	27	MF	6.68	6	10
Eberechi Eze	Crystal Palace	26	MF	7.28	7	2

Figura 11 – Exemplos de equipes geradas pelo algoritmo de balanceamento proposto.

Notadamente, a sistemática proposta tende a superar estratégias simples de distribuição por média global ou por posição nominal, pois busca contemplar as duas dimensões do equilíbrio competitivo identificadas na literatura: a igualdade de potencial médio, derivado da escolha por estratos de qualidade técnica; e a equivalência na distribuição de tipos funcionais de jogadores entre as diversas equipes (VEN, 2018; PENNER, 2025).

5

DESENVOLVIMENTO DA APLICAÇÃO

Com base nos perfis funcionais e na lógica de balanceamento definidos ao longo desta pesquisa, foi desenvolvida uma aplicação interativa de apoio à demonstração dos resultados obtidos. O sistema, denominado *Football Cluster Generator*, foi implementado em linguagem Python com o *framework* Streamlit e disponibilizado em repositório público no GitHub, garantindo transparência e reprodutibilidade ao trabalho.

É importante delimitar o papel desta aplicação no contexto do trabalho. Ela não constitui um sistema de produção para uso contínuo em contextos esportivos reais, mas uma ferramenta demonstrativa que permite explorar interativamente os dados, os processos analíticos e os resultados da pesquisa sem necessidade de acesso direto ao código-fonte.

Para um sistema de produção seriam necessários desenvolvimentos adicionais, como integração com APIs de dados esportivos e persistência em banco de dados, além da adoção do K-Means como mecanismo central de clusterização, por suas vantagens de escalabilidade e de suporte à atribuição dinâmica de novos jogadores sem reprocessamento completo da estrutura de agrupamento, conforme visto no Capítulo 4.

5.1 Objetivos e Escopo da Aplicação

A aplicação cumpre dois objetivos complementares: (i) permitir a exploração interativa da metodologia empregada na pesquisa, tornando configuráveis e visíveis etapas que exigiriam execução direta de código; e (ii) demonstrar, de forma tangível e acessível,

como os agrupamentos obtidos podem ser utilizados para a geração automática de equipes balanceadas.

O escopo abrange os dados da temporada 2024/2025 das cinco principais ligas de futebol masculino europeias, com as configurações de agrupamento restritas às opções validadas pela pesquisa: os algoritmos K-Means e *Ward's linkage*, com número de partições entre cinco e oito *clusters*.

5.2 Tecnologias e Arquitetura

Python foi adotado como linguagem principal do sistema por ter se consolidado como uma das linguagens mais populares para computação científica e análise exploratória de dados, apoiada por um ecossistema maduro de bibliotecas especializadas que cobre desde a manipulação de dados até o aprendizado de máquina e a visualização interativa (HARRIS et al., 2020; VIRTANEN et al., 2020).

A interface web interativa foi construída com **Streamlit**, *framework* que permite transformar scripts Python em aplicações web sem necessidade de conhecimento em desenvolvimento front-end, tornando a exposição dos resultados da pesquisa acessível a usuários não técnicos.

Para manipulação e tratamento dos dados, foram utilizadas as bibliotecas **pandas** (MCKINNEY, 2010) e **NumPy** (HARRIS et al., 2020). O pandas fornece estruturas de dados e ferramentas para trabalhar com conjuntos de dados tabulares, servindo como camada fundamental para computação estatística em Python (MCKINNEY, 2010). O NumPy, por sua vez, é a principal biblioteca de programação vetorial em Python e fundamento do ecossistema científico, amplamente utilizada em *pipelines* de análise em áreas como física, engenharia e economia (HARRIS et al., 2020).

O **Scikit-learn** (PEDREGOSA et al., 2011) foi utilizado para a implementação dos algoritmos de PCA, K-Means e clusterização hierárquica. A biblioteca integra uma ampla gama de algoritmos de aprendizado de máquina supervisionados e não supervisionados em uma API unificada, com ênfase em facilidade de uso e desempenho, sendo amplamente adotada tanto na academia quanto na indústria (PEDREGOSA et al., 2011).

Para a geração dos dendrogramas e demais rotinas científicas de suporte, foi utilizado o **SciPy** (VIRTANEN et al., 2020), biblioteca padrão de algoritmos científicos em Python, que inclui rotinas de clusterização hierárquica, estatística, otimização e álgebra linear. As visualizações interativas foram produzidas com **Plotly** (Plotly Technologies Inc., 2015), complementado por **Matplotlib** e **Seaborn** para representações estáticas.

A combinação NumPy / pandas / SciPy / Scikit-learn é apontada na literatura como a base típica de *pipelines* exploratórios e de prototipagem rápida em pesquisa científica (HARRIS et al., 2020; VIRTANEN et al., 2020; PEDREGOSA et al., 2011; MCKINNEY, 2010), reforçando a adequação desse conjunto de ferramentas ao trabalho.

O projeto é organizado em diretórios com separação clara de responsabilidades: `data/` (datasets, tipos e configurações), `logic/` (módulos de análise, clusterização e balanceamento), `visuals/` (geradores de gráficos e tabelas), `utils/` (funções auxiliares) e `pages/` (módulos funcionais da interface).

5.3 Módulos Funcionais

A interface é composta por três módulos acessíveis por menu lateral, cada um correspondendo a uma dimensão distinta do trabalho.

Módulo 1 — Análise dos Algoritmos de Agrupamento: permite percorrer interativamente as etapas metodológicas da pesquisa. O usuário pode configurar diferentes limiares de minutagem mínima e observar seu impacto na base de dados; visualizar o *scree plot* de variância acumulada e o gráfico de autovalores com destaque para o critério de Kaiser; comparar os índices de Silhueta e Davies-Bouldin para os dois algoritmos em função do número de *clusters* e identificar as melhores partições para aquele conjunto de dados; além de explorar dendrogramas quando aplicado o método hierárquico.

Módulo 2 — Gerador de Agrupamentos: oferece interface para aplicação configurável dos algoritmos de clusterização sobre a base de dados. O usuário seleciona o algoritmo e o número de partições entre as opções validadas pelo estudo e obtém a distribuição dos jogadores nos *clusters*, a visualização espacial dos agrupamentos no espaço do PCA e os perfis funcionais atribuídos a cada *cluster*.

Módulo 3 — Montador de Equipes Balanceadas: implementa o algoritmo de formação de equipes descrito na Seção 3. O usuário define o número de equipes a formar e a composição estrutural por *clusters*, e o sistema aplica a lógica de distribuição por faixas de desempenho, gerando times funcionalmente diversos e tecnicamente equilibrados com filtros configuráveis por liga de origem e por faixas de idade e *rating* dos jogadores.

5.4 Relação entre a Aplicação e a Pesquisa

As opções de configuração disponíveis na aplicação refletem diretamente as escolhas metodológicas validadas ao longo do estudo. Os algoritmos disponíveis, os intervalos de número de *clusters* e a lógica de composição das equipes são os mesmos utilizados e justificados na pesquisa, garantindo que a aplicação opere dentro de um espaço analiticamente fundamentado. Ela não substitui a análise científica realizada, cujos detalhes estão descritos nos Capítulos 3 e 4, mas a complementa, tornando-a acessível, reprodutível e explorável de forma interativa.

5.5 Disponibilização e Reprodutibilidade

O código-fonte completo, os *datasets* utilizados e as instruções de instalação estão disponíveis em repositório público: <https://github.com/MHCipriano/football-cluster-generator>. A reprodutibilidade do ambiente é garantida pelo arquivo `requirements.txt`, que lista todas as dependências com suas versões. Para executar a aplicação localmente, basta clonar o repositório e sequir as orientações de instalação e aplicação descritas no arquivo `readme.md`.

6

Conclusão

O trabalho investigou o uso de técnicas de clusterização em dados de desempenho de jogadores das cinco principais ligas de futebol europeias como ferramenta para a identificação de perfis funcionais e, a partir deles, para a formação de equipes balanceadas. A proposta partiu da hipótese de que o agrupamento de jogadores por similaridade de perfil, ao invés de simples balanceamento por médias globais, permitiria compor equipes mais equilibradas tanto em termos técnicos quanto funcionais.

O *pipeline* metodológico adotado, estruturado em coleta, pré-processamento, redução de dimensionalidade via PCA, clusterização e aplicação dos agrupamentos, demonstrou-se adequado e replicável para o domínio da análise esportiva. A aplicação do critério de Kaiser na definição dos componentes principais do PCA resultou na retenção de 18 componentes, responsáveis por preservar 75,93% da variância original, percentual consistente com as recomendações da literatura especializada (ROJAS-VALVERDE et al., 2020; PINO-ORTEGA et al., 2021).

A comparação entre os algoritmos K-Means e Hierárquico com ligação de *Ward*, avaliada por meio do coeficiente de Silhueta e do índice de Davies-Bouldin, complementada por análise interpretativa futebolística e por critério de aderência às formações medianas observadas em equipes reais, revelou que a solução hierárquica de *Ward's linkage* com sete *clusters* oferece o melhor equilíbrio entre refinamento funcional, interpretabilidade tática e viabilidade prática.

Os sete perfis identificados (goleiros; atacantes finalizadores; criadores ofensivos; atacantes móveis; zagueiros; meias e defensores construtores; e meias organizadores)

cobrem de forma representativa as principais macrofunções táticas do futebol moderno, mostrando elevada correspondência com as posições nominais dos jogadores na base de dados original.

A estratégia algorítmica de distribuição de jogadores entre equipes, baseada na combinação de diversidade funcional (garantida pelos *clusters*) e equilíbrio técnico relativo (garantido pelas faixas de desempenho), demonstrou-se capaz de produzir composições mais equilibradas do que métodos simples de distribuição aleatória ou por média global. Esse resultado sustenta a hipótese central do trabalho e contribui para preencher a lacuna identificada na literatura, na qual poucos estudos exploram de forma integrada o *pipeline* completo desde a clusterização até a aplicação prática na formação de equipes.

Do ponto de vista das contribuições acadêmicas, o trabalho propõe e valida uma metodologia integrada de clusterização aplicada ao futebol, com documentação detalhada das escolhas metodológicas e de seus fundamentos na literatura. A análise comparativa entre os métodos Hierárquico com Ward e K-Means, com discussão explícita de seus papéis complementares, constitui uma contribuição relevante para a área de *sports analytics*.

Sob a perspectiva prática, a *dashboard* interativa desenvolvida em *Python* garante uma base estruturada e configurável para aplicação dos resultados e métodos discutidos a contextos de futebol recreativo, amador ou semiprofissional, com possibilidade de extensão a sistemas mais completos.

Limitações

Como limitações do trabalho, destacam-se: (i) a dependência de dados de desempenho de ligas profissionais, que podem não capturar adequadamente os perfis de jogadores amadores, contexto-alvo da aplicação proposta; (ii) a utilização do *rating* médio como critério de ordenação relativa dentro dos *clusters*, métrica de natureza sintética que pode incorporar avaliações subjetivas; e (iii) a ausência de validação experimental do equilíbrio das equipes geradas por meio de partidas reais ou simuladas.

Trabalhos Futuros

Para trabalhos futuros, sugere-se: a validação experimental do sistema em contextos amadores reais, com coleta de dados sobre equilíbrio percebido e resultados das partidas; a investigação de métricas alternativas de desempenho relativo dentro dos *clusters*; a extensão do sistema para suporte à organização de campeonatos completos; e a exploração de técnicas de clusterização baseadas em modelos probabilísticos, como o Gaussian Mixture Model (GMM), como alternativa ao K-Means para implementação em produção.

Em síntese, o trabalho demonstra que a clusterização constitui uma abordagem promissora, objetiva e metodologicamente fundamentada para a formação de equipes equilibradas no futebol. A integração entre ciência de dados, aprendizado de máquina não supervisionado e análise esportiva abre perspectivas relevantes tanto para a pesquisa acadêmica em *sports analytics* quanto para o desenvolvimento de ferramentas práticas de apoio ao esporte amador e semiprofissional.

REFERÊNCIAS

AKHANLI, S. E.; HENNIG, C. Comparing clusterings and numbers of clusters by aggregation of calibrated clustering validity indexes. *Statistics and Computing*, v. 30, n. 5, p. 1523–1544, 2020. 24, 27, 35

AKHANLI, S. E.; HENNIG, C. Clustering of football players based on performance data and aggregated clustering validity indexes. *Journal of Quantitative Analysis in Sports*, v. 19, n. 2, p. 103–123, 2023. 21, 22, 23, 24, 27, 29, 30, 31, 32, 33, 34, 36, 37, 57

AMHIMMID, H. A. et al. Comparative study of four methods in hierarchical cluster analysis. *Libyan Journal of Medical and Applied Sciences*, p. 11–16, 2025. 24, 27, 35

AZZAMI, S. Y. A. et al. Clustering and profiling analysis for FIFA football player using K-Means. *Jurnal Informatika: Jurnal Pengembangan IT*, v. 10, n. 1, p. 178–189, 2025. 19, 20, 33, 34

BEHRAVAN, I. et al. Finding roles of players in football using automatic PSO-clustering algorithm. *Big Data*, v. 7, n. 1, p. 35–56, 2019. 23

DALTON-BARRON, N. et al. Clustering of match running and performance indicators to assess between- and within-playing position similarity in professional rugby league. *Journal of Sports Sciences*, v. 40, n. 15, p. 1712–1721, 2022. 25

DAğ, O.; YüKSEL, A. S.; ATMACA, Examination of player positions by cluster analysis. *Bilge International Journal of Science and Technology Research*, v. 7, n. 1, p. 43–48, 2023. 22, 23, 27, 37

D'URSO, P.; GIOVANNI, L. D.; VITALE, V. A robust method for clustering football players with mixed attributes. *Annals of Operations Research*, v. 325, n. 1, p. 9–36, 2023. 18, 22, 29, 32

FBref. *Football Statistics and History*. 2026. Disponível em: <https://fbref.com>. Acesso em: abr. 2026. 29

GARCíA-ALIAGA, A. et al. In-game behaviour analysis of football players using machine learning techniques based on player statistics. *International Journal of Sports Science & Coaching*, v. 16, n. 1, p. 148–157, 2021. 18, 20, 23, 24, 26, 27, 30, 32, 33, 34, 37

GERE, A. Recommendations for validating hierarchical clustering in consumer sensory projects. *Current Research in Food Science*, v. 6, p. 100522, 2023. 20, 24, 27, 35, 57

GU, L. et al. Playing style identification in team sports: a systematic review from statistical dimensionality reduction to unsupervised machine learning. *International Journal of Sports Science & Coaching*, v. 20, n. 6, p. 2742–2761, 2025. 13, 18, 27, 34

HAAN, M. de et al. Beyond playing positions: categorizing soccer players based on match-specific running performance using machine learning. *Journal of Sports Science & Medicine*, v. 24, n. 3, p. 565, 2025. 19, 22, 24, 27, 37

HARRIS, C. R. et al. Array programming with NumPy. *Nature*, v. 585, n. 7825, p. 357–362, 2020. 61, 62

HARTIGAN, J. A.; WONG, M. A. Algorithm AS 136: a K-means clustering algorithm. In: *Applied Statistics*. [S.l.: s.n.], 1979. v. 28, n. 1, p. 100–108. 19, 34

HENNIG, C. Clustering strategy and method selection. In: HENNIG, C. et al. (Ed.). *Handbook of Cluster Analysis*. Boca Raton: Chapman & Hall/CRC, 2015. p. 703–730. 19, 20, 36, 47, 57

HENNIG, C. What are the true clusters? *Pattern Recognition Letters*, v. 64, p. 53–62, 2015. 22

JR., J. H. W. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, v. 58, n. 301, p. 236–244, 1963. 23, 35

LEWIS, M. *Moneyball: The Art of Winning an Unfair Game*. New York: W. W. Norton & Company, 2003. 12, 17

MAUX, A. L. *Clustering Analysis in Football*. Tese (Doutorado) — Dublin Business School, Dublin, 2024. 19, 21, 22, 28, 34

MCKINNEY, W. Data structures for statistical computing in Python. In: WALT, S. van der; MILLMAN, J. (Ed.). *Proceedings of the 9th Python in Science Conference (SciPy 2010)*. Austin, TX: [s.n.], 2010. p. 56–61. 61, 62

NORRIS, O. *ScraperFC: a Python package for scraping soccer data*. 2024. Disponível em: <https://github.com/oseymour/ScraperFC>. Acesso em: abr. 2026. 29

PEDREGOSA, F. et al. Scikit-learn: machine learning in Python. *Journal of Machine Learning Research*, v. 12, p. 2825–2830, 2011. 18, 19, 28, 61, 62

PENNER, L. S. J. Player archetypes within basketball: optimizing roster composition to create a championship team. *Frontiers in Sports and Active Living*, v. 7, p. 1639431, 2025. 25, 26, 27, 31, 37, 59

PINO-ORTEGA, J. et al. Training design, performance analysis, and talent identification—a systematic review about the most relevant variables through the principal component analysis in soccer, basketball, and rugby. *International Journal of Environmental Research and Public Health*, v. 18, n. 5, p. 2642, 2021. 20, 34, 42, 44, 64

Plotly Technologies Inc. Collaborative data science. *Plotly Technologies Inc.*, Montreal, QC, 2015. Disponível em: <https://plot.ly>. 62

PRASETYO, E.; PRIYATAMA, A. S.; SETYATAMA, F. Constellation of football players determination based on cost and performance history using the K-Means clustering. *Digital Zone: Jurnal Teknologi Informasi dan Komunikasi*, v. 14, n. 2, p. 194–205, 2023. 26

RAN, X. et al. Comprehensive survey on hierarchical clustering algorithms and the recent developments. *Artificial Intelligence Review*, v. 56, n. 8, p. 8219–8264, 2023. 20, 24, 35, 47, 57

ROJAS-VALVERDE, D. et al. A systematic review of methods and criteria standard proposal for the use of principal component analysis in team's sports science. *International Journal of Environmental Research and Public Health*, v. 17, n. 23, p. 8712, 2020. 20, 27, 34, 44, 64

Sofascore. *Sports Statistics and Live Scores*. 2026. Disponível em: <https://www.sofascore.com>. Acesso em: abr. 2026. 29

VEN, E. van de. *Clustering soccer players to find the drivers of soccer team performance*. Dissertação (Mestrado) — Erasmus School of Economics, Erasmus University Rotterdam, Rotterdam, 2018. 18, 21, 22, 26, 27, 36, 37, 59

VIRTANEN, P. et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods*, v. 17, p. 261–272, 2020. 61, 62

WIJNGAARD, G. W. A. *Clustering soccer players: investigating unsupervised learning on player positions*. Utrecht: [s.n.], 2020. 21, 23, 27, 36

WING, C. et al. The use of K-means clustering for the organisation of training groups in Australian football. *International Journal of Sports Science & Coaching*, v. 20, n. 4, p. 1642–1650, 2025. 25