INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA
DE MINAS GERAIS (IFMG – *CAMPUS* BAMBUÍ)
BACHARELADO EM ENGENHARIA DE COMPUTAÇÃO

Iuri José Rodrigues de Lima

APLICAÇÃO DE TÉCNICAS DE MINERAÇÃO DE DADOS PARA A IDENTIFICAÇÃO DE PADRÕES EM PARTIDAS DE FUTEBOL

2026

IURI JOSÉ RODRIGUES DE LIMA

APLICAÇÃO DE TÉCNICAS DE MINERAÇÃO DE DADOS PARA A IDENTIFICAÇÃO DE PADRÕES EM PARTIDAS DE FUTEBOL

Trabalho de conclusão de curso apresentado ao Curso de Bacharelado em Engenharia de Computação do Instituto Federal de Educação, Ciência e Tecnologia de Minas Gerais (IFMG – *Campus* Bambuí) – para obtenção do grau de em Engenharia de Computação.

Orientador: Prof. Dr. Mateus Clemente de Sousa

Coorientador: Prof. Dr. Marcos Roberto Ribeiro

2026

Catalogação na Fonte Biblioteca IFMG - *Campus* Bambuí

L732a Lima, Iuri José Rodrigues de.

Aplicação de técnicas de mineração de dados para a identificação de padrões em partidas de futebol [manuscrito] / Iuri José Rodrigues de Lima – 2026.

82 f. :il. ; color.

Orientador: Mateus Clemente de Sousa.

Coorientador: Marcos Roberto Ribeiro.

Trabalho de Conclusão de Curso (Bacharelado em Engenharia de Computação) – Instituto Federal de Minas Gerais. *Campus* Bambuí

1. Mineração de dados. 2. Agrupamento de dados. 3. Estatísticas de futebol. 4. Ciência de dados esportiva. I. Sousa, Mateus Clemente de. II. Ribeiro, Marcos Roberto. III. Instituto Federal de Minas Gerais – *Campus* Bambuí. IV. Título.

CDD 006.312

Catalogação: João Batista Rodrigues - CRB-6/2022

Iuri José Rodrigues de Lima

APLICAÇÃO DE TÉCNICAS DE MINERAÇÃO DE DADOS PARA A IDENTIFICAÇÃO DE PADRÕES EM PARTIDAS DE FUTEBOL

Trabalho de conclusão de curso apresentado ao Curso de Bacharelado em Engenharia de Computação do Instituto Federal de Educação, Ciência e Tecnologia de Minas Gerais (IFMG – *Campus* Bambuí) – para obtenção do grau de em Engenharia de Computação.

Aprovado em 28 de Janeiro de 2026 pela banca examinadora:

Prof. Dr. Mateus Clemente de Sousa – IFMG – *Campus* Bambuí – (Orientador)
Prof. Dr. Marcos Roberto Ribeiro – IFMG – *Campus* Bambuí – (Coorientador)
Prof. Me. Álvaro Antonio Fonseca de Souza – IFMG – *Campus* Bambuí
Prof. Esp. Carlos Renato Nolli – IFMG – *Campus* Bambuí

RESUMO

O crescimento na coleta e armazenamento de dados no futebol, impulsionado por sensores, softwares de monitoramento e plataformas analíticas, tem ampliado as possibilidades de aplicação da Ciência de Dados no esporte. Nesse contexto, a Mineração de Dados configura-se como uma abordagem promissora para explorar padrões e compreender relações entre atributos técnicos e o comportamento das equipes. Este trabalho teve como objetivo identificar padrões em partidas de futebol por meio da aplicação de algoritmos de agrupamento sobre uma base pública contendo estatísticas dos jogos. Inicialmente, foram comparadas diferentes bases de dados e selecionado um conjunto com informações de eventos de partida (gols, finalizações, posse de bola, escanteios, cruzamentos, faltas e cartões), sobre o qual se realizou um processo de extração, transformação e carga para normalização e validação da consistência dos registros. Em seguida, foram construídos vetores de desempenho em janelas temporais de partidas, combinando históricos de cinco jogos com estatísticas agregadas (médias e desvios-padrão em janelas de 3, 4 e 5 partidas), que serviram de entrada para algoritmos de agrupamento não supervisionado. Os resultados indicam que os *clusters* formados capturam sobretudo diferenças de estilo e intensidade entre ligas, com padrões de desempenho recente, mas exibem baixa correspondência direta com o desfecho das partidas (vitória, empate ou derrota). Assim, os achados reforçam o uso das técnicas de Mineração de Dados como ferramenta exploratória no contexto do futebol, oferecendo subsídios para estudos futuros na Ciência de Dados Esportiva.

Palavras-chave: Mineração de Dados. Agrupamento de Dados. Estatísticas de Futebol. Ciência de Dados Esportiva.

ABSTRACT

The growth in the collection and storage of football data, driven by sensors, tracking software, and analytical platforms, has expanded the possibilities for applying Data Science in sports. In this context, Data Mining stands out as a promising approach to explore patterns and understand relationships between technical attributes and team behavior. This study aimed to identify patterns in football matches by applying clustering algorithms to a public dataset containing match statistics. First, different datasets were compared, and one was selected with information on match events (goals, shots, possession, corners, crosses, fouls, and cards). An extract, transform and load process was then carried out to normalize the data and validate record consistency. Next, performance vectors were constructed using temporal match windows, combining five-match histories with aggregated statistics (means and standard deviations over 3, 4, and 5 match windows), which served as input to unsupervised clustering algorithms. The results indicate that the resulting clusters primarily capture differences in style and intensity across leagues, reflecting recent performance patterns, but show low direct correspondence with match outcomes (win, draw, or loss). Therefore, these findings support the use of Data Mining techniques as an exploratory tool in football analytics, providing a basis for future studies in Sports Data Science.

Keywords: Data Mining. Clustering. Football Statistics. Sports Data Science.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

Nas últimas décadas, a expansão da capacidade de armazenamento e a evolução de algoritmos de aprendizado de máquina impulsionaram o uso de técnicas de análise de dados no setor de esportes (PROVOST; FAWCETT, 2013; HAN; PEI; TONG, 2022; FERRAZ *et al.*, 2023). Nesse cenário, destaca-se a Mineração de Dados como uma área fundamental da Ciência da Computação, dedicada à identificação de padrões e à extração de conhecimento em grandes volumes de dados. O processo está formalmente inserido no ciclo conhecido como *Knowledge Discovery in Databases* (KDD), que significa Descoberta de Conhecimento em Bancos de Dados, entendido como um processo que abrange as etapas de seleção, pré-processamento, transformação, mineração de dados e interpretação dos resultados (HOCHKAMP; SCHEIDLER; RABE, 2025).

No campo esportivo, o avanço de tecnologias como *Global Positioning System* (GPS), sensores inerciais, câmeras multivista e plataformas de análise estatística permitiu a coleta e integração de dados técnicos, físicos e táticos (GUDMUNDSSON; HORTON, 2017; PERIN *et al.*, 2018). Esse volume crescente de dados favoreceu o surgimento da Ciência de Dados Esportiva. Essa abordagem interdisciplinar aplica métodos estatísticos, computacionais e matemáticos para apoiar decisões técnicas e administrativas no esporte. Segundo Han, Pei e Tong (2022), a análise automatizada de dados supera limitações humanas ao reconhecer padrões que não são perceptíveis por observação empírica, oferecendo suporte à tomada de decisão.

Nos últimos anos, estudos aplicaram técnicas de aprendizado de máquina e Mineração de Dados ao futebol, explorando desde previsões de resultados (BUNKER; THABTAH, 2019), até ranqueamento de desempenho de atletas (PAPPALARDO *et al.*, 2019), e identificação de padrões táticos e espaciais (GYARMATI; STANOJEVIĆ, 2016; GUDMUNDSSON; HORTON, 2017). A detecção automática de eventos em partidas de futebol, como passes, finalizações e movimentações táticas, tem sido investigada na literatura. Abordagens baseadas em dados de rastreamento, vídeo e estatísticas têm sido aplicadas com êxito (VIDAL-CODINA *et al.*, 2022; FERNÁNDEZ; BORNN; CERVONE, 2021; RANA, 2023; NARAYANAN *et al.*, 2023).

Trabalhos como o de Cao (2012) demonstram que a previsão de resultados e o entendimento de padrões táticos têm sido aprimorados por meio de Mineração de Dados, evidenciando a influência dessas abordagens e seu potencial no contexto esportivo.

Paralelamente, tecnologias como inteligência artificial, holografia e realidade aumentada estão sendo integradas aos treinamentos e análises no futebol. Um exemplo recente foi a adoção de goleiros holográficos em sessões de treino, o que resultou em melhorias nas taxas de conversão de pênaltis e no tempo de reação dos

atletas (REUTERS, 2024). Apesar desse progresso, desafios permanecem: a padronização dos dados, sua qualidade e a integração eficiente entre áreas técnicas e científicas ainda são gargalos importantes para a plena aplicação da Ciência de Dados no esporte (DELLO IACONO *et al.*, 2025).

Diante desse cenário, o presente trabalho utilizou técnicas de Mineração de Dados para explorar indicadores de desempenho de equipes em partidas de futebol. A partir de conjuntos de dados públicos, foram construídos vetores que agregaram informações como mando de campo, posse de bola, finalizações, escanteios, faltas e cartões ao longo de janelas de partidas, sobre os quais são aplicados algoritmos de agrupamento. A análise buscou compreender como esses padrões de desempenho se distribuem entre diferentes ligas e em que medida se relacionam com os desfechos dos jogos, contribuindo para o aprofundamento da interseção entre Ciência de Dados e futebol, com foco em uma análise exploratória de estilos de jogo.

1.1 Objetivos

O presente trabalho teve como objetivo explorar padrões em indicadores de desempenho de equipes em partidas de futebol por meio da aplicação de técnicas de Mineração de Dados, em particular métodos de agrupamento, sobre um *dataset* público. Busca-se analisar métricas como posse de bola, finalizações, escanteios, faltas e cartões se organizam em grupos de comportamento, bem como de que forma esses padrões se distribuem entre diferentes ligas e se relacionam, de maneira exploratória, com os desfechos das partidas.

Os seguintes objetivos específicos foram definidos:

- pesquisar e selecionar bases de dados adequadas ao escopo do estudo;
- realizar a preparação, normalização e validação de consistência da base de dados escolhida;
- construir vetores de atributos baseados em janelas de partidas, agregando estatísticas de desempenho das equipes;
- aplicar algoritmos de Agrupamento de Dados (*clustering*) sobre diferentes representações de atributos;
- analisar e interpretar os agrupamentos obtidos, apresentando os achados por meio de métricas e visualizações gráficas.

1.2 Justificativa

A previsão esportiva tem ganhado importância com o avanço da tecnologia e da Inteligência Artificial, estimulando assim a busca por métodos refinados de coleta de dados e reconhecimento de padrões (GOMES, 2024). Essa tendência acompanha

o movimento de transformação digital no esporte, em que a análise baseada em dados torna-se uma ferramenta para clubes, analistas e pesquisadores.

Estudos recentes demonstram o potencial dessas técnicas no contexto esportivo. Linhares (2024) investigou a influência da distância percorrida por times da Premier League nos resultados das partidas, utilizando regressão logística para concluir que esse fator não teve impacto significativo nos desfechos dos jogos.

Outros estudos apontam que indicadores técnicos e táticos podem revelar perfis de atuação recorrentes e diferenças entre competições. Por exemplo, Plakias *et al.* (2023) analisaram ligas europeias por meio de técnicas multivariadas e de agrupamento, evidenciando que combinações de estatísticas de jogo permitem caracterizar estilos distintos e agrupar ligas com padrões semelhantes.

Além disso, levantamentos indicam um crescimento na adoção de departamentos de análise de dados por clubes profissionais. A série de relatórios *MLS Analytics Survey*, realizada entre 2023 e 2025, aponta um crescimento na adoção de equipes especializadas em análise de dados na *Major League Soccer* (MLS), liga de futebol profissional dos Estados Unidos e Canadá. O número de clubes com analistas dedicados passou de 21, em 2023, para 25 dos 30 clubes da liga em 2025, evidenciando a consolidação da análise de dados como uma prática institucional no futebol profissional (AMERICAN SOCCER ANALYSIS, 2025). Esse panorama reforça a relevância científica e aplicada de estudos voltados à mineração e modelagem de informações esportivas.

Maimone e Yasseri (2021) apontam que, ao longo dos anos, os jogos nas principais ligas europeias tornaram-se mais previsíveis, sugerindo uma crescente desigualdade entre as equipes. Apesar desses avanços, existem lacunas especialmente no que tange à identificação e análise de padrões estatísticos recorrentes em partidas de futebol. Diante desse cenário, torna-se necessário aprofundar os conceitos e métodos que embasam a análise de dados no contexto esportivo.

1.3 Organização do trabalho

Este trabalho está organizado em capítulos, além desta introdução. O Capítulo 2 apresenta a fundamentação teórica. O Capítulo 3 descreve a metodologia, incluindo a base de dados, o pré-processamento e os algoritmos empregados. O Capítulo 4 reúne e discute os resultados experimentais. Por fim, o Capítulo 5 apresenta as conclusões e indica direções para trabalhos futuros.

2 FUNDAMENTAÇÃO TEÓRICA

Este capítulo apresenta os fundamentos teóricos que embasam o desenvolvimento deste trabalho, reunindo os conceitos e técnicas de Mineração de Dados aplicados ao contexto do futebol.

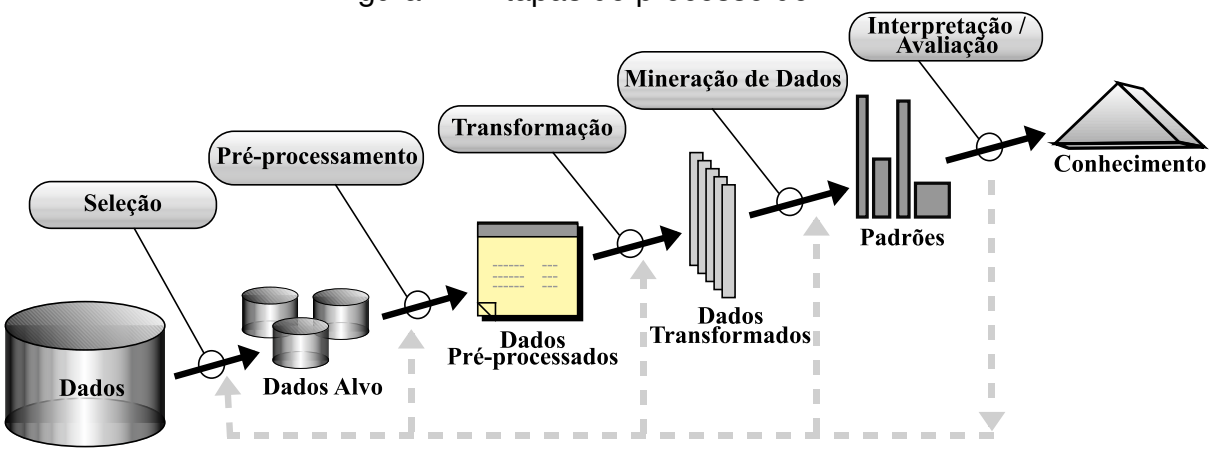
Na Seção 2.1, são discutidos os conceitos de Mineração de Dados e sua inserção no processo de descoberta de conhecimento. Em seguida, a Seção 2.2 descreve as técnicas supervisionadas e não supervisionadas. A Seção 2.3 apresenta os métodos de agrupamento de dados adotados como base para a análise. Posteriormente, a Seção 2.4 contextualiza o uso de estatísticas de futebol na Ciência de Dados Esportiva, e a Seção 2.5 discute as bases de dados públicas empregadas em estudos da área. Por fim, a Seção 2.6 sintetiza o estado da arte sobre aplicações de agrupamento ao futebol, evidenciando contribuições e lacunas de pesquisa.

2.1 Mineração de Dados

A Mineração de Dados constitui uma etapa no processo de KDD, sendo responsável pela identificação automática ou semiautomática de padrões relevantes, estruturas e relações não triviais em grandes conjuntos de dados. Trata-se de uma área interdisciplinar que integra conceitos da estatística, Aprendizado de Máquina e banco de dados, com o objetivo de extrair informações a partir de dados brutos (HAN; PEI; TONG, 2022; LESKOVEC; RAJARAMAN; ULLMAN, 2020; RIBEIRO, 2022).

O processo de KDD foi formalizado por Fayyad, Piatetsky-Shapiro e Smyth (1996) como um ciclo composto por cinco etapas sequenciais: seleção, pré-processamento, transformação, mineração e interpretação ou avaliação dos dados. Esse ciclo permite a conversão de dados brutos em conhecimento útil, estruturado e interpretável, como ilustrado na Figura 1.

Figura 1 – Etapas do processo de KDD



Fonte: Adaptado de FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996.

A seleção consiste na identificação e extração de dados provenientes de fontes distintas, como bancos relacionais, planilhas, arquivos estruturados ou sistemas transacionais. A relevância e a representatividade dos dados selecionados impactam no resultado analítico, sendo fundamental que estejam alinhados com o objetivo da investigação (HAN; PEI; TONG, 2022; PROVOST; FAWCETT, 2013).

O pré-processamento envolve a limpeza e preparação dos dados. São aplicadas técnicas para correção de inconsistências, tratamento de valores ausentes, remoção de ruídos e *outliers*, e padronização de formatos. Essa etapa visa garantir a qualidade e a integridade do conjunto de dados, fundamentais para resultados analíticos (RIBEIRO, 2022; HAN; PEI; TONG, 2022; JAMES *et al.*, 2021).

No processo de transformação, os dados são transformados em formatos adequados aos algoritmos de mineração. Isso inclui a normalização de atributos, discretização de variáveis contínuas, redução de dimensionalidade e codificação de variáveis categóricas. Tais transformações são cruciais para garantir que os algoritmos operem de forma eficiente e que os padrões extraídos sejam representativos (HAN; PEI; TONG, 2022; JAMES *et al.*, 2021).

A mineração de dados é a etapa na qual algoritmos são aplicados para extrair padrões, regras, estruturas e agrupamentos a partir dos dados processados. As técnicas podem incluir modelos de classificação, regressão, associação ou agrupamento, conforme os objetivos da análise e a natureza dos dados disponíveis (HAN; PEI; TONG, 2022; JAMES *et al.*, 2021; LESKOVEC; RAJARAMAN; ULLMAN, 2020).

Após a mineração, os padrões descobertos devem ser avaliados quanto à sua validade, utilidade e interpretabilidade. Métricas como acurácia, precisão, índice de silhueta ou medidas estatísticas são utilizadas para validar os resultados. A apresentação dos achados é essencial para viabilizar sua aplicação prática ou teórica (JAMES *et al.*, 2021; RIBEIRO, 2022).

2.2 Técnicas de Mineração de Dados

A aplicação de técnicas de Mineração de Dados se diferencia da estatística clássica, pois está voltada à descoberta de padrões desconhecidos, sem a necessidade de hipóteses prévias. A relação entre Mineração de Dados e Aprendizado de Máquina é objeto de diferentes interpretações. Enquanto alguns autores tratam o Aprendizado de Máquina como um subconjunto da Mineração de Dados, aplicável principalmente às tarefas de modelagem supervisionada (HAN; PEI; TONG, 2022; TAN *et al.*, 2018), outros destacam distinções conceituais, especialmente no que se refere aos objetivos de previsão *versus* interpretação (PROVOST; FAWCETT, 2013; LAROSE, 2014).

As técnicas de mineração podem ser divididas em dois grupos principais:

supervisionadas e não supervisionadas. Essa distinção se baseia na presença ou ausência de variáveis-alvo conhecidas durante o treinamento dos modelos (HAN; PEI; TONG, 2022; RIBEIRO, 2022; JAMES *et al.*, 2021).

2.2.1 Técnicas supervisionadas

As técnicas supervisionadas baseiam-se na existência de um conjunto de dados previamente rotulado, no qual cada instância de entrada está associada a uma saída conhecida. Esse conjunto serve de base para o processo de treinamento, no qual o algoritmo aprende uma função que mapeia as variáveis independentes (também chamadas de preditoras ou atributos) para a variável dependente (ou rótulo). O objetivo é que essa função tenha capacidade de generalização, ou seja, consiga prever corretamente os rótulos de novos dados não vistos, com base nos padrões aprendidos (MITCHELL, 1997; HAN; PEI; TONG, 2022).

Entre as abordagens supervisionadas, destacam-se os métodos de classificação e regressão. A classificação visa prever rótulos discretos, ou seja, categorias ou classes finitas. Um exemplo é o algoritmo *k-Nearest Neighbors* (*k*-NN), que significa *k*-Vizinhos Mais Próximos, que atribui o rótulo de uma nova instância com base nos rótulos das instâncias mais próximas no espaço métrico. Sua simplicidade, aliada à eficácia em problemas com baixo ruído e dados bem distribuídos, faz do *k*-NN uma técnica útil em tarefas como reconhecimento de padrões (JAMES *et al.*, 2021). Outro modelo utilizado é a regressão logística, que, apesar do nome, é um método de classificação binária. Seu diferencial está na modelagem da probabilidade de pertencimento a uma classe a partir de uma função logística, sendo empregada em contextos de modelagem estatística e análise preditiva (HAN; PEI; TONG, 2022).

Já a regressão busca estimar valores contínuos a partir de variáveis explicativas. Um exemplo é a regressão linear, que assume uma relação linear entre os atributos de entrada e a variável de saída. A partir de métodos de mínimos quadrados, o modelo ajusta uma função que minimiza o erro entre os valores previstos e os observados (JAMES *et al.*, 2021). Outra variação é a regressão com árvores de decisão, que divide o espaço de atributos em regiões homogêneas por meio de partições sucessivas, facilitando a interpretação dos modelos mesmo em contextos não lineares (HAN; PEI; TONG, 2022).

2.2.2 Técnicas não supervisionadas

Ao contrário do aprendizado supervisionado, as técnicas não supervisionadas operam sobre dados não rotulados, isto é, sem uma variável-alvo explícita. O foco está na exploração dos dados em busca de estruturas ocultas, padrões recorren-

tes, agrupamentos naturais ou correlações latentes. Essas técnicas são fundamentais em contextos onde o conhecimento prévio é limitado ou onde se deseja realizar uma análise exploratória preliminar (LAROSE, 2014; HAN; PEI; TONG, 2022).

Uma das abordagens mais relevantes é o agrupamento (*clustering*), cujo objetivo é particionar os dados em subconjuntos (*clusters*) de elementos similares entre si e distintos dos demais. O algoritmo *k*-Means, por exemplo, parte da definição prévia de *k* centróides e busca minimizar a distância intra-*cluster*, ajustando iterativamente os agrupamentos até a convergência (JAMES *et al.*, 2021). Já o algoritmo *Density-Based Spatial Clustering of Applications with Noise* (DBSCAN), que significa Agrupamento Espacial Baseado em Densidade com Aplicações em Ruído, define *clusters* com base na densidade local de pontos, sendo capaz de identificar agrupamentos e lidar com ruídos e *outliers* sem a necessidade de definir previamente o número de *clusters*. Isso o torna eficaz em conjuntos de dados com distribuições irregulares (HAN; PEI; TONG, 2022).

Outra técnica é a descoberta de regras de associação, cujo propósito é identificar padrões de associação entre itens ou eventos dentro de grandes bases de dados. O algoritmo Apriori, por exemplo, utiliza uma abordagem iterativa para gerar conjuntos frequentes de itens com base no princípio do suporte mínimo, construindo regras do tipo “se A, então B” (HAN; PEI; TONG, 2022). O FP-Growth, por sua vez, propõe uma estrutura chamada FP-Tree, que permite extrair regras sem a geração explícita de candidatos, otimizando o desempenho em bases extensas e densas (HAN; PEI; TONG, 2022). A escolha da técnica de Mineração de Dados mais adequada depende da natureza do problema, do tipo e estrutura dos dados disponíveis, da existência ou não de variáveis-alvo, bem como dos objetivos específicos da análise (PROVOST; FAWCETT, 2013).

Com o advento do *Big Data*, da Computação em Nuvem e da Inteligência Artificial, a Mineração de Dados assumiu um papel central na geração de conhecimento aplicado. Em setores como saúde, finanças, *marketing* e educação, ela tem sido empregada para apoiar decisões baseadas em evidências, otimizar recursos e antecipar comportamentos (HAN; PEI; TONG, 2022). No contexto esportivo, dados gerados por sensores, câmeras, softwares de rastreamento e sistemas de análise de desempenho tornaram viável a aplicação de técnicas de mineração. Seus resultados colaboram tanto para a análise de atletas quanto para o entendimento de dinâmicas coletivas de jogo (GUDMUNDSSON; HORTON, 2017; PERIN *et al.*, 2018).

O uso da Mineração de Dados em domínios específicos, como o esportivo, ainda enfrentam desafios metodológicos e técnicos. Entre esses desafios destacam-se a ausência de padronização nas estruturas das bases de dados, a necessidade de uma curadoria criteriosa das variáveis e a definição de procedimentos para a validação dos resultados obtidos. Como ressalta Cao (2012), a eficácia dos modelos preditivos

no contexto esportivo está ligada à qualidade e à relevância dos dados disponíveis, além da escolha adequada das métricas de avaliação, que devem refletir o objetivo analítico do problema em estudo. Esses fatores se tornam relevantes em estudos voltados para a interpretação de padrões, onde a confiabilidade das conclusões depende diretamente da robustez metodológica e da consistência do processo analítico (PIETRASZEWSKI *et al.*, 2025).

2.3 Agrupamento de dados

A inclusão deste capítulo justifica-se pelo fato de que a presente pesquisa faz uso de técnicas de Agrupamento de Dados como abordagem para a análise e identificação de padrões em indicadores de desempenho de equipes em partidas de futebol. Nesse contexto, é essencial compreender os fundamentos teóricos dessa técnica, de modo a embasar metodologicamente as escolhas realizadas.

O Agrupamento de Dados tem como objetivo identificar estruturas naturais ou padrões latentes em um conjunto de dados, agrupando instâncias semelhantes entre si com base em medidas de proximidade, como distâncias ou similaridades. Cada subconjunto formado, denominado *cluster*, deve apresentar coesão interna (alta similaridade entre seus elementos) e separação externa (baixa similaridade com elementos de outros grupos) (HAN; PEI; TONG, 2022; LESKOVEC; RAJARAMAN; ULLMAN, 2020).

Diferentemente das técnicas supervisionadas, que dependem de variáveis-alvo rotuladas para aprender uma função preditiva, os métodos de agrupamento operam de forma exploratória, buscando inferir padrões diretamente dos dados brutos, sem conhecimento prévio de categorias. Conforme argumenta Rai e Singh (2010), o agrupamento permite reduzir a complexidade de bases volumosas ao representar milhares de instâncias por meio de poucos grupos representativos, servindo como base para tarefas de análise, modelagem e tomada de decisão.

Han, Pei e Tong (2022) afirma que os algoritmos de agrupamento podem ser organizados em quatro paradigmas principais, sendo métodos particionais, hierárquicos, baseados em densidade e baseados em modelos. Cada um desses adota diferentes premissas sobre a distribuição dos dados e usa estratégias específicas para formação dos agrupamentos. A seguir, são detalhados os fundamentos, vantagens e limitações de cada abordagem.

2.3.1 Métodos particionais

Os métodos particionais buscam dividir um conjunto de n instâncias em k grupos não sobrepostos, de modo que a variabilidade dentro dos grupos seja minimi-

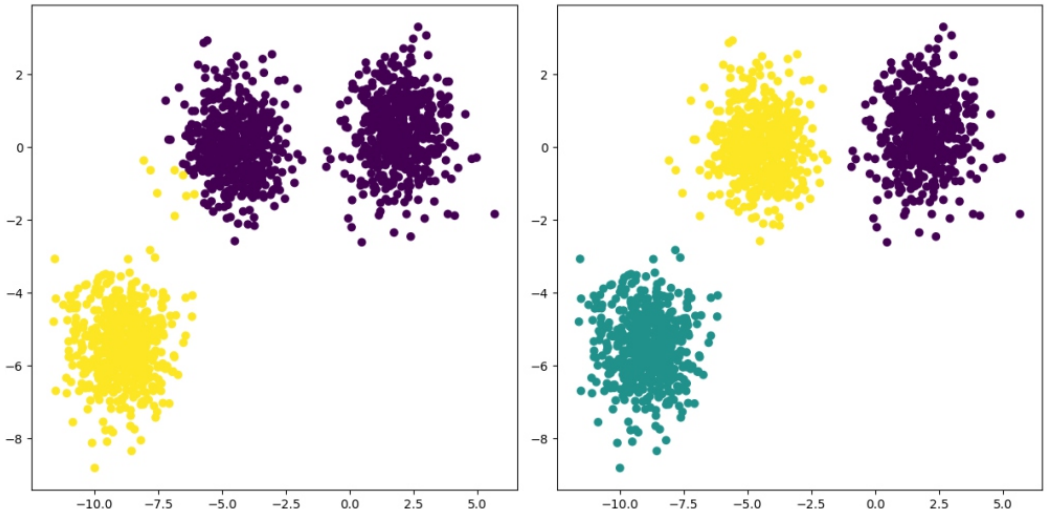
zada. O algoritmo K-Means é conhecido por operar no meio de um processo iterativo de otimização. A Equação (2.1) mostra a função objetivo do algoritmo que busca minimizar a soma das distâncias quadráticas entre os pontos e os centróides dos *clusters* aos quais pertencem.

$$J = \sum_{i=1}^{k} \sum_{x \in C_i} |x - \mu_i|^2 \tag{2.1}$$

em que x representa um ponto de dados (vetor de características) pertencente ao *cluster* C_i e μ_i é o centróide do cluster C_i. O algoritmo inicia com a seleção aleatória de *k* centróides, atribui cada instância ao centróide mais próximo, recalcula os centróides com base nas médias das instâncias atribuídas e repete esse processo até a convergência (JAMES *et al.*, 2021; HAN; PEI; TONG, 2022).

A Figura 2 apresenta a representação gráfica a partir de resultados da execução do algoritmo K-Means para o valor de *k* igual a 2 e 3, respectivamente, onde cada cor representa um grupo identificado. Observa-se que quando se considera apenas dois *clusters* (gráfico à esquerda), a distinção entre os grupos fica prejudicada, pois uma parte dos pontos do *cluster* amarelo encontra-se muito próxima dos pontos do *cluster* roxo, indicando sobreposição entre as regiões de decisão. Já no caso com três *clusters* (gráfico à direita), o agrupamento acompanha de forma mais fiel a distribuição original dos dados, com três grupos bem definidos, não havendo evidência visual de que um quarto *cluster* seja necessário. Esse exemplo ilustra uma desvantagem da escolha inadequada do número de *clusters*. Valores de *k* baixos podem gerar agrupamentos pouco discriminativos, comprometendo a interpretação dos padrões encontrados.

Figura 2 – Formação de *clusters* pelo algoritmo K-Means.



Fonte: CRUZ, 2023.

Apesar da simplicidade computacional e da escalabilidade, o K-Means

apresenta limitações. Primeiramente, exige que o número de clusters *k* seja previamente definido. Além disso, o algoritmo é sensível a *outliers* e pressupõe que os agrupamentos tenham formas esféricas e variâncias similares, o que nem sempre se verifica em dados reais. Para contornar essas limitações, variantes como o K-Medoids, que utiliza instâncias reais como representações centrais, e o K-Modes, adaptado a dados categóricos, têm sido exploradas (HAN; PEI; TONG, 2022).

2.3.2 Métodos hierárquicos

Os métodos hierárquicos constroem uma estrutura em árvore, conhecida como dendrograma, que representa a formação progressiva dos agrupamentos. Essa abordagem pode ser dividida em dois modelos principais: o aglomerativo (*bottom-up*), que inicia considerando cada instância como um *cluster* independente e os une recursivamente, e o divisivo (*top-down*), que parte de um único grupo e o divide iterativamente (HAN; PEI; TONG, 2022).

Um fator determinante na qualidade dos agrupamentos obtidos é o critério de ligação adotado. Dentre esses, destacam-se o *single linkage* e o *ward linkage*, com características distintas que influenciam a forma dos agrupamentos gerados. O primeiro *single linkage* define a distância entre dois *clusters* como sendo a menor distância entre quaisquer dois elementos pertencentes a eles. Essa abordagem favorece a formação de estruturas encadeadas, o que pode ser problemático em conjuntos de dados com ruído. Já o segundo *ward linkage* busca minimizar o aumento da variância intra-cluster a cada fusão, o que resulta em grupos mais compactos e esféricos (RIBEIRO, 2022).

Além dos critérios de ligação já citados, outros esquemas são empregados pela literatura. Sejam C_i e C_j dois *clusters* e $d(x, y)$ uma função de distância entre instâncias. No *complete linkage*, a distância entre *clusters* é definida pelo par mais distante, conforme (2.2):

$$D_{\text{complete}}(C_i, C_j) = \max_{x \in C_i,\, y \in C_j} d(x, y). \tag{2.2}$$

Essa estratégia tende a produzir grupos mais compactos, porém pode ser sensível a *outliers*. No *average linkage*, utiliza-se a média das distâncias entre todos os pares, como definido em (2.3):

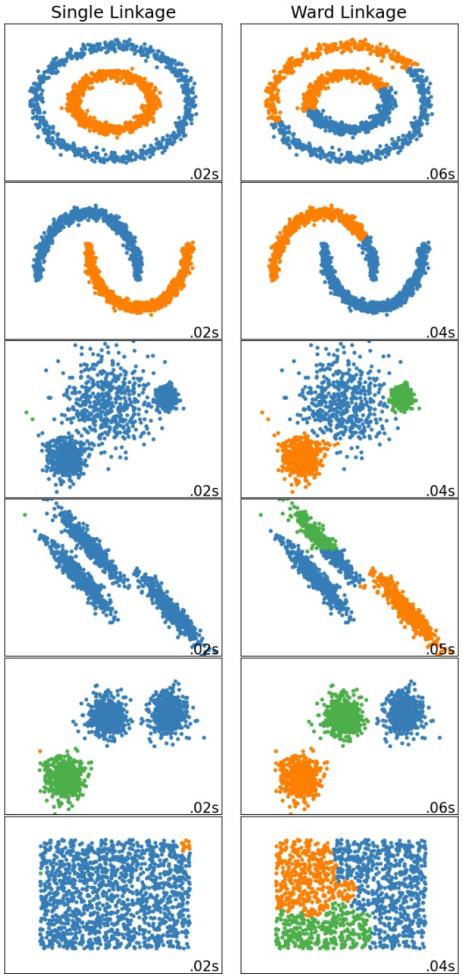
$$D_{\text{average}}(C_i, C_j) = \frac{1}{|C_i||C_j|} \sum_{x \in C_i} \sum_{y \in C_j} d(x, y). \tag{2.3}$$

Esse critério pode reduzir efeitos extremos ao comparar *clusters*. Já o método de Ward não se baseia diretamente em $d(x, y)$, mas no incremento de variância

intra-*cluster* provocado por uma fusão, favorecendo partições com menor dispersão interna e maior separação entre grupos (HAN; PEI; TONG, 2022; RIBEIRO, 2022).

A Figura 3 apresenta uma comparação visual da aplicação do algoritmo de agrupamento hierárquico aglomerativo utilizando os critérios de ligação *single* e *ward*. Cada subfigura exibe os agrupamentos resultantes em diferentes distribuições de dados, evidenciando como a escolha do critério de ligação influencia diretamente a forma e a coesão dos *clusters* formados. Os valores exibidos no canto inferior direito de cada subfigura representam o tempo de execução do algoritmo, em segundos, para aquele cenário específico.

Figura 3 – Métodos de ligação *single linkage* e *ward linkage* em diferentes conjuntos de dados.



Fonte: Adaptado de PEDREGOSA *et al.*, 2011.

O *Balanced Iterative Reducing and Clustering using Hierarchies* (BIRCH), que significa Redução Iterativa Balanceada e Agrupamento Usando Hierarquias, é um método hierárquico incremental proposto para cenários com grandes volumes de dados. O algoritmo resume os dados em uma estrutura compacta em memória, denominada *Clustering Feature Tree* (CFT), que significa Árvore de Atributos de Agrupamento. Nessa árvore, subconjuntos de instâncias são representados por estatísticas

agregadas (por exemplo, contagem e somas), o que permite inserir novos dados e atualizar os agrupamentos de forma eficiente, sem a necessidade de reprocessar todo o conjunto (ZHANG; RAMAKRISHNAN; LIVNY, 1996).

O BIRCH representa subconjuntos de pontos por um vetor de características de agrupamento *Clustering Feature* (CF), definido como N, LS, SS. N é o número de instâncias no subgrupo, $LS = \sum_{i=1}^{N} x_i$ é a soma linear dos vetores e $SS = \sum_{i=1}^{N} x_i^2$ é a soma quadrática. A partir dessas estatísticas, é possível estimar propriedades como centróide e medidas de dispersão sem armazenar todos os pontos individualmente, reduzindo custo de memória e tempo de execução.

De forma geral, o algoritmo constrói a *CF-tree* inserindo instâncias e absorvendo-as em nós folha quando um critério de proximidade e um limiar de raio e diâmetro são satisfeitos. Quando não são, novos nós são criados. Em seguida, a árvore pode ser condensada e, por fim, aplicada uma etapa de agrupamento global sobre as entradas resumidas. Essa estratégia torna o BIRCH adequado quando se deseja escalabilidade, embora o resultado possa depender do limiar adotado e, como em outros métodos, da função de distância utilizada (ZHANG; RAMAKRISHNAN; LIVNY, 1996).

Embora ofereçam uma representação rica e interpretável dos dados, os métodos hierárquicos sofrem com a alta complexidade computacional, geralmente $\mathcal{O}(n^3)$, e com a ausência de mecanismos naturais para correção de fusões ou divisões equivocadas. Além disso, sua sensibilidade a pequenas perturbações nos dados pode comprometer a estabilidade dos agrupamentos.

2.3.3 Métodos baseados em densidade

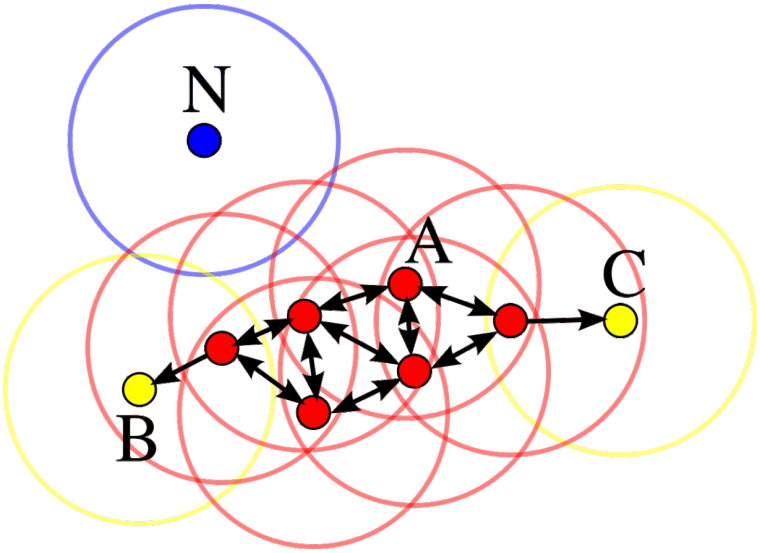
Diferentemente dos métodos anteriores, os algoritmos baseados em densidade identificam agrupamentos como regiões do espaço de atributos com alta concentração de pontos, separadas por regiões de baixa densidade. O DBSCAN é um representante dessa classe e opera com dois parâmetros fundamentais, o raio de vizinhança ε e o número mínimo de pontos *MinPts* (HAN; PEI; TONG, 2022).

No DBSCAN, um ponto é considerado *core point* quando sua vizinhança ε contém pelo menos *MinPts* pontos, caracterizando uma região densa. Um *border point* é um ponto que está dentro da vizinhança de um *core point*, mas cuja própria vizinhança não atinge *MinPts*. Pontos que não são *core* nem *border* são tratados como ruído (*noise*). Os *clusters* são formados expandindo-se a partir dos *core points*, conectando pontos alcançáveis por densidade.

A Figura 4 apresenta um exemplo de como o algoritmo DBSCAN organiza os dados com base em densidade local. Os pontos em vermelho ilustram instâncias que satisfazem o número mínimo de vizinhos dentro de um raio especificado, sendo,

portanto, considerados pontos centrais. Já os pontos amarelos representam pontos de borda, que, embora não atendam ao critério de densidade por si só, estão dentro da vizinhança de um ponto central. O ponto azul, por sua vez, caracteriza um ruído, pois não possui vizinhos suficientes e não pertence à região de influência de nenhum ponto central. A união dos pontos centrais e de borda define uma região densa, que é interpretada pelo algoritmo como um agrupamento válido.

Figura 4 – Formação de agrupamentos pelo algoritmo DBSCAN.



Fonte: WIKIPEDIA, 22/09/2025.

Entre as vantagens do DBSCAN estão a capacidade de identificar agrupamentos com formatos arbitrários e a detecção automática de *outliers*. No entanto, sua eficácia é dependente da escolha adequada dos parâmetros ε e MinPts, que podem variar entre diferentes regiões do espaço. Para superar essa limitação, o algoritmo *Ordering Points To Identify the Clustering Structure* (OPTICS), que significa Ordenação de Pontos para Identificação da Estrutura de Agrupamento, foi proposto como uma extensão do DBSCAN, permitindo a identificação de agrupamentos com diferentes densidades (SINGH; GIRDHAR; DAHIYA, 2022).

2.3.4 Métodos baseados em modelos

Nos métodos baseados em modelos, parte-se da suposição de que os dados foram gerados por uma combinação de distribuições estatísticas subjacentes, geralmente modeladas por funções de densidade de probabilidade. O *Gaussian Mixture Model* (GMM), que significa Modelo de Mistura Gaussiana, é um exemplo e assume que cada *cluster* segue uma distribuição normal multivariada. A estimativa dos parâmetros do modelo é feita por meio do algoritmo *Expectation-Maximization* (EM), que significa Maximização por Expectativa, que consiste em duas etapas, a Etapa E (Expectativa) e a Etapa M (Maximização), como descreve Hastie, Tibshirani e Friedman

(2009).

A etapa de Expectativa calcula a probabilidade de cada instância pertencer a cada componente gaussiano e a etapa de Maximização, atualiza os parâmetros (médias, covariâncias e pesos) das distribuições para maximizar a verossimilhança total.

Os modelos de mistura oferecem maior flexibilidade do que os métodos particionais, permitindo sobreposição entre grupos e representações mais precisas de distribuições assimétricas. Contudo, sua sensibilidade à inicialização e a necessidade de especificar o número de componentes são limitações práticas recorrentes.

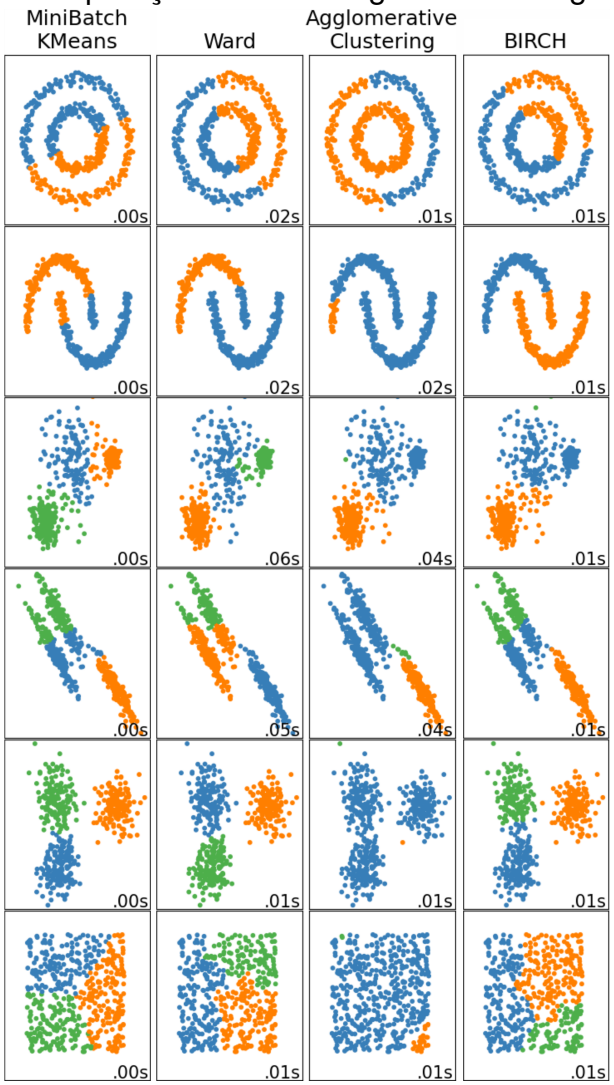
Além dos modelos probabilísticos clássicos, são citados os *Self-Organizing Maps* (SOM), que significa Mapas Auto-Organizáveis, redes neurais não supervisionadas propostas por Kohonen (2001). O SOM projeta dados de alta dimensionalidade em uma grade bidimensional, preservando relações topológicas entre instâncias. Durante o treinamento, cada entrada é mapeada para seu neurônio mais próximo, *Best Matching Unit* (BMU), que significa Unidade de Melhor Correspondência, e os pesos dos vizinhos desse neurônio são ajustados em direção à entrada, promovendo a formação de regiões coerentes que representam agrupamentos naturais dos dados.

Com o tempo, surgiram abordagens que integram aprendizado profundo e técnicas de agrupamento como o *Deep Embedded Clustering* (DEC), que significa Agrupamento Incorporado Profundo. Este combina redes neurais profundas para extrair representações latentes com mecanismos de agrupamento que refinam essas representações e as atribuições de *cluster*. Essa otimização conjunta torna o DEC adequado para dados de alta dimensionalidade e estruturas complexas (XIE; GIRSHICK; FARHADI, 2016).

Como pode ser visto na Figura 5, cada algoritmo produz agrupamentos com formas e quantidades distintas, destacando a influência do paradigma na estrutura dos dados. Os valores exibidos no canto inferior direito de cada subfigura representam o tempo de execução do algoritmo, em segundos, para aquele cenário específico.

A escolha de qual abordagem utilizar depende de múltiplos fatores: o tamanho da base de dados, a forma dos agrupamentos esperados (esféricos, arbitrários), a presença de ruído e a necessidade (ou não) de definição prévia do número de *clusters*. Como destacado por Tan *et al.* (2018), não existe um algoritmo universalmente superior. Cada técnica produz partições distintas, com *clusters* de formatos e fronteiras diferentes. Esse comportamento evidencia que não existe um algoritmo de agrupamento universalmente superior. A qualidade do resultado depende tanto das características dos dados quanto dos pressupostos de cada método, tornando necessária a escolha criteriosa da técnica em função do problema em estudo.

Figura 5 – Comparação visual de algoritimos de agrupamento.



Fonte: Adaptado de PEDREGOSA *et al.*, 2011.

2.3.5 Validação interna de agrupamentos

Em problemas não supervisionados, a avaliação da qualidade do agrupamento é desafiadora pela ausência de uma "verdade de referência" (rótulos). Nesses casos, recorre-se a métricas internas, que quantificam coesão e separação dos *clusters* utilizando apenas os dados e a partição obtida (XU; WUNSCH, 2005).

2.3.5.1 Coeficiente de Silhueta

O coeficiente de Silhueta foi proposto por Rousseeuw (1987) e mede, para cada instância i, o quão bem ela se ajusta ao seu *cluster*. Seja $a(i)$ a distância média de i para as demais instâncias do seu próprio *cluster* (coesão) e $b(i)$ a menor distância média de i para instâncias de um *cluster* diferente (separação). A silhueta de i é definida por (2.4):

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i),\, b(i)\}}. \tag{2.4}$$

O valor de $s(i)$ varia em $[-1, 1]$. Valores próximos de 1 indicam boa alocação (alta separação e boa coesão), valores próximos de 0 sugerem sobreposição entre *clusters*, e valores negativos indicam que a instância pode estar mais próxima de outro grupo do que do grupo ao qual foi atribuída (ROUSSEEUW, 1987).

Além da avaliação por instância, é comum empregar a silhueta global (ou silhueta média) como um indicador agregado da qualidade do agrupamento. A silhueta global é definida como a média de $s(i)$ sobre todas as instâncias, conforme (2.5):

$$\overline{s} = \frac{1}{n} \sum_{i=1}^{n} s(i). \tag{2.5}$$

Valores maiores de $\overline{s}$ indicam, em geral, partições com melhor separação entre *clusters* e maior coesão interna, sendo frequente utilizar essa medida para comparar diferentes parametrizações e diferentes números de grupos. Em análises diagnósticas, também pode-se calcular a média de $s(i)$ por *cluster*, a fim de identificar grupos com pior ajuste relativo (ROUSSEEUW, 1987).

2.3.5.2 Índice de Calinski–Harabasz (CH)

O índice de Calinski–Harabasz, também conhecido como *Variance Ratio Criterion* (VRC), foi proposto por Caliński e Harabasz (1974). A métrica compara a dispersão entre *clusters* com a dispersão dentro dos *clusters*. Seja k o número de *clusters*, n o número de instâncias, W_k a matriz (ou soma) de dispersão intra-*cluster* e B_k a dispersão inter-*cluster*. O índice pode ser expresso por (2.6):

$$CH = \frac{\mathrm{tr}(B_k)/(k-1)}{\mathrm{tr}(W_k)/(n-k)}. \tag{2.6}$$

em que tr denota o traço. Valores maiores indicam partições com maior separação entre *clusters* e menor dispersão interna, sendo comum selecionar a configuração que maximiza o CH ao comparar diferentes parametrizações e números de grupos (CALIŃSKI; HARABASZ, 1974).

2.3.5.3 Índice de Davies–Bouldin (DB)

O índice de Davies–Bouldin avalia a similaridade média entre cada *cluster* e o *cluster* mais parecido com ele (DAVIES; BOULDIN, 1979). Para cada *cluster* C_i, define-se uma medida de dispersão S_i (por exemplo, distância média dos pontos ao centróide) e a distância M_{ij} entre os centróides de C_i e C_j. O índice é definido por

(2.7):

$$DB = \frac{1}{k} \sum_{i=1}^{k} \max_{j \neq i} \left(\frac{S_i + S_j}{M_{ij}} \right). \tag{2.7}$$

Como DB cresce quando *clusters* são pouco separados e/ou muito dispersos, valores menores indicam melhor qualidade de agrupamento (DAVIES; BOULDIN, 1979).

2.3.6 Validação externa de agrupamentos

Além das métricas internas, existem métricas externas, que avaliam a concordância entre a partição obtida por um algoritmo de agrupamento e uma partição de referência (por exemplo, rótulos conhecidos ou uma segmentação previamente definida). Essas métricas não dependem de distâncias entre instâncias, mas da correspondência entre as atribuições de *cluster* e os rótulos de referência (XU; WUNSCH, 2005).

2.3.6.1 Matriz de contingência

A comparação entre uma partição produzida por um algoritmo de agrupamento e uma partição de referência pode ser representada por uma matriz (tabela) de contingência, que resume as frequências conjuntas entre os grupos obtidos e as classes de referência (AGRESTI, 2013).

Seja $U = \{C_1, \ldots, C_k\}$ a partição produzida pelo algoritmo de agrupamento e $V = \{L_1, \ldots, L_m\}$ a partição de referência (por exemplo, vitória, empate e derrota), ambas definidas sobre o mesmo conjunto de n instâncias. Define-se a matriz de contingência $\mathbf{N} = [n_{ij}] \in \mathbb{N}^{k \times m}$ conforme (2.8):

$$n_{ij} = \left| \{ x : x \in C_i \ \wedge \ x \in L_j \} \right|, \quad i = 1, \ldots, k, \ \ j = 1, \ldots, m. \tag{2.8}$$

A partir de $\mathbf{N}$, as somas marginais (frequências por linha e por coluna) são dadas por (2.9) e (2.10):

$$a_i = \sum_{j=1}^{m} n_{ij}, \quad i = 1, \ldots, k, \tag{2.9}$$

$$b_j = \sum_{i=1}^{k} n_{ij}, \quad j = 1, \ldots, m. \tag{2.10}$$

A matriz de contingência fornece uma visão direta da distribuição dos rótulos dentro de cada *cluster*, permitindo avaliar se determinados grupos concentram uma classe específica ou se apresentam mistura entre classes. Além disso, ela é a base para o cálculo de métricas externas de validação, como o *Adjusted Rand Index* (ARI) e a *Normalized Mutual Information* (NMI), pois tais métricas dependem das frequências conjuntas e marginais observadas (HUBERT; ARABIE, 1985; VINH; EPPS; BAILEY, 2010).

2.3.6.2 Adjusted Rand Index (ARI).

O *Rand Index* mede a concordância entre duas partições com base em comparações par-a-par, considerando se pares de instâncias são colocados no mesmo grupo em as ambas partições ou em grupos diferentes em ambas. O ARI introduz um ajuste para o acaso, de modo que partições aleatórias tenham valor esperado próximo de zero (HUBERT; ARABIE, 1985). O ARI pode ser calculado a partir da tabela de contingência, utilizando n_{ij}, a_i e b_j definidos em (2.8), (2.9) e (2.10). Sua expressão é dada por (2.11):

$$ARI = \frac{\sum_i \sum_j \binom{n_{ij}}{2} - \frac{\left(\sum_i \binom{a_i}{2}\right)\left(\sum_j \binom{b_j}{2}\right)}{\binom{n}{2}}}{\frac{1}{2}\left[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}\right] - \frac{\left(\sum_i \binom{a_i}{2}\right)\left(\sum_j \binom{b_j}{2}\right)}{\binom{n}{2}}}. \tag{2.11}$$

O ARI assume valor 1 quando as partições são idênticas. Valores próximos de 0 indicam concordância compatível com o acaso e valores negativos podem ocorrer quando há discordância sistemática em relação ao esperado aleatoriamente (HUBERT; ARABIE, 1985).

2.3.6.3 Normalized Mutual Information (NMI).

A NMI quantifica a quantidade de informação compartilhada entre duas partições U e V, interpretadas como variáveis aleatórias discretas. Seja $p(i)$ a probabilidade de uma instância pertencer ao *cluster* i em U, $p(j)$ a probabilidade de pertencer ao grupo j em V, e $p(i,j)$ a probabilidade conjunta. A informação mútua é definida por (2.12):

$$MI(U,V) = \sum_i \sum_j p(i,j) \log\left(\frac{p(i,j)}{p(i)p(j)}\right). \tag{2.12}$$

Como $MI(U,V)$ depende do número de grupos e do tamanho das partições, utiliza-se uma normalização para tornar a medida comparável entre diferentes cenários. Uma forma comum é dada por (2.13):

$$NMI(U,V) = \frac{MI(U,V)}{\sqrt{H(U)H(V)}}. \tag{2.13}$$

em que $H(U)$ e $H(V)$ denotam as entropias das partições. O NMI assume valores em $[0,1]$, em que 1 indica concordância perfeita e valores próximos de 0 indicam baixa dependência entre as partições (STREHL; GHOSH, 2002; VINH; EPPS; BAILEY, 2010).

2.3.7 Redução de dimensionalidade e interpretação dos agrupamentos

Além das métricas internas e externas de validação, empregou-se uma etapa complementar de análise com o objetivo de interpretar a estrutura dos dados e dos *clusters* no espaço de atributos. Como os vetores utilizados possuem alta dimensionalidade e incluem variáveis potencialmente correlacionadas, técnicas de redução de dimensionalidade auxiliam a sintetizar a variabilidade do conjunto e a viabilizar visualizações em baixa dimensão. Neste trabalho, adotou-se a Análise em *Principal Component Analysis* (PCA), que significa Análise em Componentes Principais, como ferramenta descritiva para (i) avaliar quanta variância pode ser representada por poucas componentes, (ii) examinar a contribuição relativa dos atributos nas direções principais de variação e (iii) projetar as instâncias no plano das componentes principais, apoiando a inspeção qualitativa de padrões associados a ligas, *clusters* e ao resultado das partidas.

2.3.7.1 Análise em Componentes Principais

A Análise em PCA é uma técnica estatística utilizada para explorar a estrutura de dados multivariados e reduzir sua dimensionalidade, sobretudo quando há atributos correlacionados. Em termos gerais, a PCA substitui o conjunto original de variáveis por um conjunto menor de componentes ortogonais, construídas como combinações lineares dos atributos originais e ordenadas de modo que as primeiras concentrem a maior parte da variabilidade observada (JOLLIFFE, 2002; HASTIE; TIBSHIRANI; FRIEDMAN, 2009).

Considere uma matriz de dados $\mathbf{X} \in \mathbb{R}^{n \times p}$, com n instâncias e p atributos, após centralização e, quando apropriado, padronização para reduzir efeitos de escala. A PCA pode ser apresentada a partir da matriz de covariância (ou correlação), definida em (2.14):

$$\mathbf{S} = \frac{1}{n-1}\mathbf{X}^{\top}\mathbf{X}. \tag{2.14}$$

A decomposição espectral de $\mathbf{S}$ fornece autovalores e autovetores conforme

(2.15):

$$\mathbf{S}\mathbf{v}_i = \lambda_i \mathbf{v}_i, \tag{2.15}$$

em que λ_i representa a variância explicada pela i-ésima componente principal e $\mathbf{v}_i$ define os pesos associados aos atributos na formação dessa componente. As componentes são ordenadas do maior para o menor autovalor, de modo que as primeiras tendem a reter mais informação (variabilidade) do conjunto de dados.

A projeção das instâncias no subespaço gerado pelas primeiras q componentes é dada por (2.16):

$$\mathbf{Z} = \mathbf{X}\mathbf{V}_q, \tag{2.16}$$

em que $\mathbf{V}_q = [\mathbf{v}_1, \ldots, \mathbf{v}_q]$ e $\mathbf{Z} \in \mathbb{R}^{n \times q}$ contém os escores das componentes principais, isto é, as coordenadas de cada instância no espaço reduzido.

Uma forma comum de orientar a escolha de q é por meio da proporção de variância explicada acumulada, definida em (2.17):

$$\text{PVE}(q) = \frac{\sum_{i=1}^{q} \lambda_i}{\sum_{i=1}^{p} \lambda_i}. \tag{2.17}$$

Do ponto de vista interpretativo, a PCA permite analisar os pesos das componentes, que indicam como os atributos originais contribuem para cada direção principal de variação, e os escores, que representam as instâncias no novo espaço de baixa dimensão. Essa representação viabiliza visualizações bidimensionais, como PC1–PC2, e auxilia na inspeção de padrões e tendências no conjunto de dados (HASTIE; TIBSHIRANI; FRIEDMAN, 2009).

Considere vetores de desempenho construídos a partir de janelas de partidas contendo, entre outros atributos, posse de bola e finalizações. Em muitos cenários, essas variáveis apresentam associação, pois equipes que mantêm maior posse tendem a finalizar mais. Nesse caso, a primeira componente pode concentrar a variação comum entre esses atributos, atuando como uma síntese de comportamento ofensivo geral na janela analisada. Já uma segunda componente pode capturar um contraste entre estilos, por exemplo, equipes com alta posse e menor volume de finalizações em comparação a equipes com menor posse e maior volume de finalizações, caracterizando um jogo mais direto. Assim, ao observar os escores no plano PC1–PC2, torna-se mais simples interpretar diferenças de comportamento sem a necessidade de examinar simultaneamente um grande número de variáveis.

Em estudos que envolvem agrupamento, a projeção em poucas componentes é utilizada como apoio à análise dos resultados, pois permite visualizar a distribuição das instâncias no espaço reduzido e discutir, de forma qualitativa, como os grupos

obtidos se organizam em relação às direções de maior variação do conjunto de dados.

No campo da análise esportiva, especialmente no futebol, o agrupamento tem se mostrado uma ferramenta para revelar padrões em grandes volumes de dados. Por exemplo, Głowania, Kozak e Juszczuk (2023) utilizaram técnicas de *clustering* para agrupar partidas com base em variáveis como posse de bola, finalizações e passes certos. Essa aplicação revelou grupos de jogos com comportamentos estatísticos similares que estavam associados a diferentes probabilidades de vitória. Essa análise permitiu identificar perfis de jogo vitoriosos, mesmo em situações com estatísticas aparentemente equilibradas.

Estudos vêm utilizando técnicas de agrupamento para segmentar jogadas específicas (como escanteios ou transições rápidas), agrupar atletas por estilo de jogo, ou identificar padrões táticos espaço-temporais por meio de dados de rastreamento (MUÑOZ *et al.*, 2024; YEUNG; BUNKER; FUJII, 2024). Essas aplicações mostram que o agrupamento pode servir tanto como etapa exploratória quanto para modelagens mais complexas.

2.4 Estatísticas de futebol

A integração entre esporte e Ciência de Dados tem impulsionado uma transformação na forma como o futebol é compreendido, analisado e gerido. Por meio do uso de sensores, câmeras, sistemas de rastreamento e plataformas de coleta de dados, tornou-se possível mensurar uma gama de variáveis relacionadas ao desempenho individual, dinâmicas coletivas e aspectos contextuais das partidas. Essa evolução contribuiu para o desenvolvimento do campo da Ciência de Dados Esportiva, que se baseia em métodos estatísticos, computacionais e de Aprendizado de Máquina para analisar o jogo de maneira sistemática e quantitativa (GUDMUNDSSON; HORTON, 2017; PERIN *et al.*, 2018).

Há estudos voltados à análise de desempenho esportivo que propõem formas de categorizar as métricas utilizadas no futebol. Entre as abordagens, destaca-se a divisão entre estatísticas técnicas, físicas, táticas e contextuais (CARLING; WILLIAMS; REILLY, 2008), (HUGHES; BARTLETT, 2002), (REIN; MEMMERT, 2016), (LAGO-PEÑAS; DELLAL, 2010) e (LIU; HOPKINS; GÓMEZ, 2016).

As estatísticas técnicas envolvem ações associadas ao manuseio da bola e constituem, historicamente, o núcleo mais tradicional da análise de desempenho no futebol. Entre as principais métricas estão: número de passes (curtos, médios, longos), porcentagem de acerto nos passes, finalizações ao gol, cruzamentos, dribles bem-sucedidos, desarmes, interceptações, faltas cometidas e assistências. Essas estatísticas são extraídas a partir de vídeos, sistemas de anotação manual e, por soluções automatizadas com visão computacional e Aprendizado de Máquina. O uso

dessas métricas é central em processos como observação, análise e monitoramento *scouting* de atletas, avaliação pós-jogo e análise comparativa entre jogadores ou equipes (CARLING; WILLIAMS; REILLY, 2008; HUGHES; BARTLETT, 2002).

As estatísticas físicas quantificam os aspectos biomecânicos e fisiológicos do desempenho dos jogadores, sendo coletadas por tecnologias como GPS, acelerômetros, sensores inerciais e sistemas de rastreamento óptico. São exemplos de métricas: distância total percorrida, quantidade de acelerações e desacelerações, número de *sprints*, velocidade máxima, tempo em alta intensidade e carga metabólica estimada. Essas variáveis são fundamentais para o monitoramento da carga de trabalho, prescrição individualizada de treinamentos, planejamento de estratégias de rotação do elenco e prevenção de lesões (CARLING; WILLIAMS; REILLY, 2008; BRADLEY *et al.*, 2013). Dispositivos como os da marca Catapult[1], por exemplo, permitem o acompanhamento em tempo real da intensidade da atividade física durante jogos e treinos.

As estatísticas táticas são voltadas à análise da organização coletiva da equipe e seu comportamento espacial ao longo da partida. Envolvem métricas como ocupação de zonas do campo, compactação entre linhas (distância média entre defesa, meio e ataque), variação de formações táticas, densidade posicional em regiões estratégicas, transições ofensivas e defensivas, além da sincronia nos movimentos entre setores. A coleta dessas métricas requer dados de rastreamento posicional (*tracking data*), e técnicas como mapas de calor, grafos de rede e diagramas de Voronoi têm sido aplicadas para compreender padrões coletivos (REIN; MEMMERT, 2016). Sistemas como o SkillCorner[2] utilizam inteligência artificial para automatizar esse tipo de análise, gerando percepções sobre a estrutura tática das equipes.

Por fim, as estatísticas contextuais abrangem variáveis externas ao desempenho direto dos atletas, mas que exercem influência significativa sobre o comportamento das equipes e o resultado final do jogo. Entre as principais estão: mando de campo, situação classificatória, nível e estilo do adversário, fase do campeonato, horário da partida e condições climáticas. Incorporar essas variáveis em modelos analíticos permite construir inferências mais robustas, controlando fatores de viés que afetam o desempenho (LAGO-PEÑAS; DELLAL, 2010; LIU; HOPKINS; GÓMEZ, 2016). Tais variáveis são utilizadas em estudos preditivos e explicativos, auxiliando analistas na compreensão de flutuações no rendimento esportivo sob diferentes circunstâncias.

A correta interpretação dessas estatísticas exige uma compreensão integrada de suas interações. Conforme observam Gudmundsson e Horton (2017), o futebol é altamente dinâmico e não-linear, o que torna essencial a análise das inter-relações entre jogadores, setores do campo e eventos temporais para que os dados

[1] https://www.catapult.com/
[2] https://skillcorner.com/

sejam verdadeiramente significativos. Por isso, análises isoladas de métricas técnicas ou físicas, sem considerar o contexto tático ou situacional, podem conduzir a interpretações equivocadas.

A consolidação de plataformas especializadas como StatsBomb[3], Opta[4], WyScout[5], InStat[6], Second Spectrum[7] e SportVU[8] tem possibilitado a disponibilização de dados estruturados sobre eventos e movimentações dos jogadores. Essas fontes servem a diversos públicos: comissões técnicas, analistas de desempenho, departamentos de *scouting*, jornalistas e a comunidade acadêmica. A granularidade dos dados viabiliza análises que buscam, por exemplo, identificar padrões recorrentes de ataque, medir a efetividade defensiva em zonas específicas do campo ou estimar a contribuição individual de atletas com base em modelos probabilísticos (VIDAL-CODINA *et al.*, 2022; FERNÁNDEZ; BORNN; CERVONE, 2021).

Apesar da crescente sofisticação técnica, a análise estatística no futebol ainda enfrenta desafios metodológicos. Um dos principais refere-se à padronização dos dados. Diferentes ligas e provedores adotam critérios distintos para mensurar eventos, o que dificulta a comparação entre bases e compromete a reprodutibilidade dos estudos (PERIN *et al.*, 2018). Além disso, a variabilidade tática entre equipes, os estilos de jogo e os fatores contextuais exigem que as métricas sejam interpretadas com cautela, considerando sempre o papel tático do jogador e a função que ele desempenha no sistema da equipe.

A literatura recente tem explorado o uso de técnicas estatísticas e de Aprendizado de Máquina para investigar quais atributos se correlacionam com os resultados das partidas. Linhares (2024) analisaram a relação entre distância percorrida e desfecho do jogo utilizando regressão logística. Os resultados indicaram que essa métrica, de forma isolada, não apresentou influência estatisticamente significativa sobre o placar final. Em contrapartida, Głowania, Kozak e Juszczuk (2023) identificaram correlações positivas entre vitória e atributos como posse de bola, número de finalizações e eficiência ofensiva, especialmente quando analisados de forma contextualizada.

Esses estudos reforçam a importância de uma abordagem contextualizada na análise estatística do futebol. Como enfatiza Fernández, Bornn e Cervone (2021), a complexidade do jogo exige a integração entre dados técnicos, físicos e táticos, articulados a modelos capazes de capturar relações não lineares e dependências temporais. Assim, o uso eficaz de estatísticas no futebol não se limita à quantificação de eventos, mas à construção de representações informacionais que apoiem a tomada

[3] https://www.hudl.com/en_gb/products/statsbomb
[4] https://www.statsperform.com/opta/
[5] https://www.hudl.com/en_gb/products/wyscout
[6] https://www.hudl.com/products/instat
[7] https://www.secondspectrum.com/
[8] https://www.statsperform.com/pt-br/resource/sportvu-the-independently-validated-optical-tracking-solution/

de decisão e o desenvolvimento estratégico em campo.

2.5 Fontes de dados em estudos de futebol

A disponibilidade e qualidade das bases de dados são fatores determinantes para o sucesso de aplicações em Mineração de Dados. No domínio esportivo, a variabilidade contextual e a natureza multidimensional das partidas exigem fontes abrangentes, estruturadas e confiáveis. No futebol, bases de dados públicas têm sido empregadas em estudos acadêmicos e projetos de análise de desempenho, permitindo desde análises estatísticas descritivas até modelagens preditivas e extração de padrões. No caso do futebol, diferentes estudos têm utilizado fontes variadas de dados, muitas delas públicas, consolidando um conjunto de bases exploradas na literatura científica recente.

Um exemplo é a base disponibilizada pelo site Football-Data.co.uk[9], utilizada em estudos acadêmicos por conter estatísticas de partidas de ligas europeias, incluindo dados como gols, mandos de campo, resultados parciais, *odds* de casas de apostas e datas dos jogos. A simplicidade e o formato estruturado dos arquivos facilitam seu uso em análises exploratórias e preditivas, como demonstrado por Bunker e Thabtah (2019), que utilizaram essa base para prever resultados de partidas a partir de variáveis básicas e contextuais. Outros estudos, como os de Ashraf (2021) e Smith (2020), também adotaram essa fonte de dados em modelagens estatísticas aplicadas ao futebol, reforçando sua relevância para investigações nessa área.

Outra fonte referenciada é a plataforma Kaggle[10], que hospeda conjuntos de dados relacionados ao futebol, organizados por competições, temporadas e tipos de eventos. Dentre os exemplos conhecidos estão o European Soccer Database[11], que reúne informações de ligas europeias entre 2008 e 2016, e conjuntos de dados que contêm atributos técnicos e táticos de partidas específicas. Embora nem todos os dados tenham origem institucional padronizada, sua acessibilidade, documentação e comunidade ativa tornam a plataforma uma referência para experimentação e prototipagem em Ciência de Dados aplicada ao esporte.

Dentre as bases com maior detalhamento (granularidade), destaca-se a StatsBomb Open Data[12], que disponibiliza dados de eventos detalhados, como passes, finalizações, ações defensivas e posicionamento dos jogadores ao longo das partidas. Essa base é utilizada em pesquisas que demandam modelagens táticas, espaciais ou temporais, permitindo análises aprofundadas sobre a dinâmica do jogo. Estudos como o de Fernández, Bornn e Cervone (2021) e Vidal-Codina *et al.* (2022)

[9] https://www.football-data.co.uk/
[10] https://www.kaggle.com/
[11] https://www.kaggle.com/datasets/hugomathien/soccer
[12] https://github.com/statsbomb/open-data

demonstram sua aplicação para análise automatizada de ações e extração de padrões tático-comportamentais. A base cobre competições como a Copa do Mundo Feminina da FIFA, a FA Women's Super League e outros torneios com dados rotulados manualmente e validados.

Apesar das contribuições dessas fontes, ainda há limitações. A falta de padronização entre ligas e fornecedores de dados, somada à inconsistência na taxonomia e na codificação dos eventos (por exemplo, diferentes critérios para registrar finalizações, cruzamentos ou posse de bola), dificulta a comparação direta entre competições e pode introduzir vieses nas análises. Além disso, a presença de campos ausentes ou parcialmente preenchidos em algumas ligas reduz a confiabilidade de certas estatísticas e exige etapas adicionais de limpeza, integração e seleção de atributos antes da aplicação das técnicas de mineração de dados.

Cabe mencionar que nem todos os estudos revisados revelam com clareza as bases utilizadas ou as tornam publicamente disponíveis, o que representa uma limitação metodológica relevante no campo. Contudo, observa-se uma tendência crescente à adoção de dados abertos, tanto por questões de transparência quanto pela necessidade de promover reprodutibilidade e comparabilidade entre pesquisas.

2.6 Estado da arte

As seções a seguir apresentam um panorama atualizado de estudos que utilizam técnicas de agrupamento aplicadas ao futebol, ressaltando seus métodos, abordagens e contribuições. Quando pertinente, recorreu-se a exemplos relevantes em esportes cognatos, como o basquete, para reforçar a fundamentação metodológica.

2.6.1 Estudos de agrupamento de dados no futebol

A literatura apresenta trabalhos que buscam identificar padrões, prever resultados, classificar jogadas e avaliar o desempenho de jogadores e equipes por meio de dados estatísticos e modelos computacionais. Essa tendência reflete o uso da Ciência de Dados como campo aplicado ao esporte, apresentando uma área de investigação em contínuo crescimento, com abordagens metodológicas e desafios teóricos Perin *et al.* (2018) e Fernández, Casals *et al.* (2024).

Akhanli e Hennig (2023) propuseram um modelo de agrupamento *fuzzy*, aplicado a jogadores da temporada 2014–2015 de oito ligas europeias. O modelo utilizou uma métrica de dissimilaridade e validação baseada em múltiplos índices, resultando em agrupamentos úteis tanto para composição de elenco quanto para identificação de pares de jogadores com estilos similares.

Akhanli e Hennig (2023) apresentaram uma aplicação de agrupamento de jogadores de futebol baseada em dados de desempenho e índices de validação agregados. O trabalho traz a importância da escolha de medidas de dissimilaridade e da validação cruzada de resultados, contribuindo para a consolidação de boas práticas na análise de agrupamentos esportivos.

Kalt (2024) aplicou análise de agrupamento hierárquico em equipes de ligas europeias na temporada 2022–2023, utilizando distância euclidiana e método de Ward. O estudo identificou dois agrupamentos principais: equipes de elite claramente distintas das demais, sugerindo a utilidade de abordagens hierárquicas para segmentação de estilos táticos conforme o nível competitivo.

Mais recentemente, Demir, Şahin e Üre (2025) propuseram uma abordagem de DEC para revelar estilos de jogo de equipes com base em dados de eventos, combinando aprendizado profundo e agrupamento para capturar padrões táticos complexos. O estudo demonstrou a capacidade de modelos híbridos em identificar dinâmicas coletivas.

2.6.2 Contribuições e lacunas identificadas

A literatura voltada ao futebol apresenta exemplos de aplicação de agrupamento, com destaque para modelos hierárquicos, *fuzzy* e, mais recentemente, híbridos e baseados em aprendizado profundo. Entretanto, observa-se menor incidência de pesquisas explorando integração de variáveis contextuais ou modelos capazes de generalizar entre diferentes competições.

Apesar dos avanços, ainda existem lacunas relevantes, especialmente quanto à padronização das bases de dados e à replicabilidade dos métodos empregados. A revisão conduzida por Fernández, Casals *et al.* (2024) destaca a carência de transparência metodológica e a heterogeneidade nos critérios de validação de resultados, reforçando a necessidade de estudos que unam rigor estatístico e reprodutibilidade.

Com base nos fundamentos discutidos, observa-se que a interseção entre Mineração de Dados, análise estatística e técnicas de agrupamento fornece uma base sólida para a identificação de padrões em contextos esportivos. A seguir, este trabalho apresenta a metodologia proposta para a aplicação dessas técnicas na análise de partidas de futebol.

3 METODOLOGIA

Este capítulo descreve os procedimentos metodológicos adotados para o desenvolvimento do presente trabalho. A estrutura metodológica está orientada pela combinação de abordagens exploratórias, descritivas e quantitativas, com o objetivo de investigar padrões estatísticos em dados de partidas de futebol por meio de agrupamento. São apresentados a seguir os critérios de classificação metodológica da pesquisa, os materiais e ferramentas utilizados, bem como a organização do processo de trabalho.

3.1 Classificação metodológica

A presente pesquisa caracteriza-se, segundo a tipologia proposta por Gerhardt e Silveira (2009), como sendo de natureza aplicada, pois visa empregar conhecimentos e técnicas da Ciência de Dados e da Mineração de Dados para a resolução de um problema, a identificação de padrões recorrentes em desempenho de equipes de futebol a partir de estatísticas de partidas. Embora não esteja centrada no desenvolvimento de um produto ou sistema computacional, os resultados obtidos possuem aplicação prática no contexto esportivo, especialmente no apoio à tomada de decisão baseada em dados.

Quanto aos seus objetivos, a pesquisa é classificada como exploratória, uma vez que busca aprofundar a compreensão sobre a aplicação de técnicas de Mineração de Dados ao futebol. A investigação tem como finalidade levantar hipóteses iniciais e identificar possíveis padrões em conjuntos de dados, sem, no entanto, estabelecer relações de causalidade. Complementarmente, assume também um caráter descritivo, ao se propor a organizar, apresentar e interpretar as características dos dados analisados por meio de recursos computacionais (GERHARDT; SILVEIRA, 2009).

Do ponto de vista dos procedimentos metodológicos, trata-se de uma pesquisa predominantemente documental e bibliográfica, fundamentada na análise de publicações acadêmicas, bem como em bases de dados compostas por estatísticas de partidas de futebol. Conforme apontam Laville e Dionne (1999), esse tipo de abordagem visa articular teorias e dados empíricos previamente produzidos, sendo adequada para estudos que não envolvem intervenção direta do pesquisador sobre os fenômenos investigados. Ademais, realiza-se um estudo de caso computacional, ao aplicar técnicas de mineração sobre um conjunto específico de dados (como temporadas ou competições) com o intuito de identificar padrões relevantes. A pesquisa também pode ser classificada como ex-post-facto, pois trabalha com dados retrospectivos, sem interferência sobre os eventos analisados.

Por fim, adota-se uma abordagem quantitativa, conforme definido por

(GERHARDT; SILVEIRA, 2009), uma vez que o tratamento e a análise dos dados se baseiam em informações numéricas. Essa abordagem permite realizar inferências estatísticas e identificar tendências, relações e agrupamentos com base em evidências empíricas extraídas dos dados.

3.2 Metodologia Scrum e Metodologia Analítica

A condução deste trabalho foi estruturada com base nos princípios da metodologia Scrum, adaptada ao contexto acadêmico. Essa abordagem, originalmente desenvolvida para a gestão ágil de projetos de *software*, tem se mostrado eficaz também em ambientes de pesquisa e ensino, por proporcionar ciclos curtos de planejamento, execução e revisão contínua (SCHWABER; SUTHERLAND, 2020). O projeto foi desenvolvido com entregas semanais (*sprints*) e reuniões regulares de alinhamento com o professor orientador, garantindo a evolução incremental, o acompanhamento próximo do progresso e a flexibilidade para ajustes conforme as descobertas e desafios surgiram.

O processo de desenvolvimento seguiu as boas práticas de gestão de projetos ágeis, com a definição de entregáveis intermediários, revisões parciais e documentação contínua das decisões técnicas e analíticas. A organização em ciclos curtos permitiu revisar os resultados, alinhar a execução às expectativas acadêmicas e manter a transparência no avanço da pesquisa.

No que tange à metodologia analítica, foi adotado um processo de Mineração de Dados com foco em agrupamento. Com a seleção e preparação da base de dados, foi aplicado um algoritmo de agrupamento adequado às características dos dados coletados, visando identificar padrões e estruturas em partidas com comportamentos estatísticos semelhantes. O objetivo foi compreender como certas configurações estatísticas de partidas estão associadas a desfechos ou perfis de jogo, contribuindo para a análise exploratória do futebol sob uma perspectiva quantitativa.

3.3 Materiais e tecnologias

Nesta seção são detalhados os materiais, métodos e procedimentos adotados para realizar a pesquisa.

3.3.1 Escolha da base de dados

A seleção da base de dados utilizada neste trabalho foi conduzida de forma sistemática, a partir do levantamento de diferentes conjuntos de dados voltados para o futebol, citados na literatura e disponíveis em repositórios públicos. Entre as al-

ternativas inicialmente identificadas, destacaram-se conjuntos com diferentes níveis de granularidade, abrangência temporal e riqueza de atributos, tais como resultados históricos de ligas específicas, bases nacionais e coleções de eventos detalhados de partida.

Para a escolha do conjunto principal, foram considerados critérios metodológicos relacionados à compatibilidade com os objetivos do estudo: (i) disponibilidade de estatísticas de desempenho em nível de partida (como gols, finalizações, posse de bola, escanteios, faltas, cartões e cruzamentos); (ii) cobertura de múltiplas ligas e temporadas, permitindo a comparação entre contextos competitivos distintos; (iii) consistência estrutural e documental, de modo a favorecer a reprodutibilidade; e (iv) acesso público e licenciamento adequado para uso acadêmico. A partir desses critérios, o European Soccer Database[1], disponibilizado na plataforma Kaggle[2], foi selecionado como fonte de dados para este estudo, por oferecer um equilíbrio favorável entre detalhamento estatístico, número de ligas contempladas e período histórico coberto.

Embora outras bases, como coleções focadas exclusivamente na Premier League, conjuntos nacionais (por exemplo, campeonatos brasileiros) e dados de eventos disponibilizados pela StatsBomb, também apresentassem potencial de uso, elas foram consideradas menos aderentes ao escopo deste trabalho. Em geral, essas alternativas ou careciam de algumas das estatísticas necessárias para a análise de padrões de desempenho, ou exigiriam um esforço de integração e padronização de múltiplos arquivos além do previsto para o cronograma deste estudo. Dessa forma, a adoção do *European Soccer Database* constitui uma decisão metodológica fundamentada em critérios de qualidade, completude e viabilidade prática, detalhados e comparados no Capítulo 4.

3.3.2 Ambiente de desenvolvimento e ferramentas de software

As implementações realizadas neste trabalho foram desenvolvidas em linguagem Python, em um ambiente voltado à análise de dados, mineração de dados e experimentação de algoritmos de agrupamento. Todas as rotinas de criação de vetores de atributos, execução dos algoritmos de agrupamento e geração de gráficos foram implementadas em scripts e *notebooks* Python, organizados em um repositório versionado.

A seguir são listadas as principais ferramentas e bibliotecas utilizadas:

- Python, versão 3.11.9 (https://www.python.org/);
- VSCode (*Visual Studio Code*) como ambiente de desenvolvimento integrado (https://code.visualstudio.com/);

[1] https://www.kaggle.com/datasets/hugomathien/soccer
[2] https://www.kaggle.com/

- Pandas, versão 2.3.3, para manipulação e análise de dados tabulares por meio de *DataFrames* (https://pandas.pydata.org/);
- NumPy, versão 2.4.0, para operações vetoriais e matriciais (https://numpy.org/);
- Scikit-learn, versão 1.7.2, para padronização de dados, aplicação dos algoritmos de agrupamento e cálculo de métricas (https://scikit-learn.org/);
- Matplotlib, versão 3.10.8, para a construção de gráficos e visualizações dos resultados (https://matplotlib.org/);
- Biblioteca padrão *sqlite3* do Python e ferramentas gráficas como DB Browser for SQLite, para acesso, inspeção e validação do banco de dados relacional em formato SQLite;
- Biblioteca *xml.etree.ElementTree*, para leitura e processamento dos campos em formato XML;
- Git e GitHub, para versionamento, controle de histórico e backup do código-fonte (https://github.com/).

O ambiente de desenvolvimento foi configurado em um notebook com sistema operacional Microsoft Windows 11. As especificações de hardware utilizadas foram:

- Processador: Intel Core i7-11370H;
- Memória RAM: 16 GB;
- GPU: NVIDIA GeForce GTX 1650;
- Armazenamento: SSD de 500 GB.

Os experimentos foram executados em linha de comando a partir da estrutura de diretórios do projeto, permitindo reproduzir o fluxo de processamento desde a leitura da base de dados em SQLite, passando pelas etapas de construção dos vetores de atributos, aplicação dos algoritmos de agrupamento e, por fim, geração das tabelas e gráficos apresentados no Capítulo 4.

Os scripts de extração, transformação de dados, construção dos vetores de atributos e execução dos experimentos de agrupamento foram desenvolvidos em Python e estão disponíveis em um repositório público no GitHub[3], de modo a favorecer a reprodutibilidade dos experimentos realizados neste trabalho.

3.4 Engenharia de atributos

Após a seleção e consolidação da base de dados, foi realizada uma etapa de Engenharia de Atributos com o objetivo de transformar os eventos brutos de partida em vetores numéricos capazes de representar o desempenho das equipes ao longo do tempo. Essa etapa foi fundamental para que as técnicas de mineração de dados e

[3] https://github.com/iurirodrigues17/tcc-futebol-agrupamento

agrupamento operem sobre variáveis que reflitam padrões táticos e comportamentais, em linha com recomendações recentes da literatura de análise de desempenho no futebol (CARPITA; CIAVOLINO; PASCA, 2019; BUNKER; THABTAH, 2019; YI *et al.*, 2019).

A seleção das ligas considerou, prioritariamente, a completude dos eventos técnicos disponíveis (gols, finalizações, finalizações no alvo, posse de bola, escanteios, cruzamentos, faltas e cartões). Estudos mostram que essas estatísticas estão entre os indicadores mais utilizados para caracterizar estilos de jogo, carga física e eficácia ofensiva em diferentes contextos competitivos (YI *et al.*, 2019; LAGO-PEÑAS; DELLAL, 2010; COLLET, 2013; SAPP; SPANGENBURG; IRVING, 2018). Adicionalmente, trabalhos que exploram a própria *European Soccer Database*, utilizada neste estudo, evidenciam o potencial desse tipo de informação para modelar o desempenho coletivo e os desfechos das partidas (CARPITA; CIAVOLINO; PASCA, 2019). Com base nesses critérios, optou-se por manter apenas as seis ligas com cobertura consistente desses eventos, excluindo competições com registros escassos ou incompletos.

No que se refere às variáveis, foram priorizadas estatísticas reportadas em estudos de predição de resultados e análise de desempenho, tais como gols marcados e sofridos, número de finalizações e finalizações no alvo, posse de bola, escanteios, cruzamentos e indicadores disciplinares (faltas cometidas, cartões amarelos e vermelhos) (LAGO-PEÑAS; DELLAL, 2010; COLLET, 2013; YI *et al.*, 2019). Esses atributos têm sido associados tanto à probabilidade de vitória quanto à caracterização de estilos de jogo mais associativos ou diretos, além de capturarem aspectos de agressividade e disciplina tática (SAPP; SPANGENBURG; IRVING, 2018; BUNKER; THABTAH, 2019).

Para incorporar a dimensão temporal e a noção de forma recente das equipes, as variáveis de eventos foram agregadas em janelas deslizantes de partidas. Inicialmente, foram construídos vetores baseados no histórico das últimas cinco partidas, incluindo contagens e médias de desempenho. Em seguida, foram desenvolvidos vetores estatísticos em janelas de 3, 4 e 5 partidas, calculando-se, para cada atributo, médias e desvios-padrão nesse intervalo. Essa estratégia busca representar simultaneamente o nível médio de desempenho e a variabilidade recente das equipes, o que é coerente com abordagens de modelagem que incorporam indicadores de forma em horizontes de poucas partidas (CARPITA; CIAVOLINO; PASCA, 2019; YI *et al.*, 2019). Estudos recentes de revisão sobre predição de resultados em esportes coletivos destacam justamente o uso de agregações temporais e indicadores derivados de desempenho técnico como entrada para modelos de aprendizado de máquina (BUNKER; THABTAH, 2019).

Ao final dessa etapa, obtiveram-se duas famílias principais de vetores: (i) vetores de histórico de cinco partidas, com atributos agregados a partir de sequências fixas de jogos, e (ii) vetores estatísticos em janelas deslizantes de 3, 4 e 5 partidas,

incorporando medidas de tendência central e dispersão. Essas representações constituem a base para os experimentos de agrupamento, bem como para a análise em componentes principais.

4 RESULTADOS

Este capítulo apresenta os resultados obtidos com a aplicação das técnicas de Mineração de Dados propostas neste trabalho à base de partidas de futebol construída. Na Seção 4.1, descrevem-se a seleção e a preparação do conjunto de dados, incluindo as etapas de limpeza, normalização e verificação de consistência. A Seção 4.2 apresenta os vetores de atributos construídos com janelas de cinco partidas e os resultados iniciais de agrupamento por liga. Em seguida, a Seção 4.3 discute vetores baseados em janelas deslizantes de 3, 4 e 5 jogos e os efeitos de diferentes combinações de atributos nos resultados dos agrupamentos. Por fim, a Seção 4.4 apresenta a Análise em Componentes Principais como apoio à exploração e à interpretação dos perfis de desempenho identificados.

4.1 Seleção e preparação da base de dados

Esta seção descreve o processo de seleção e preparação da base de dados utilizada neste estudo, desde a comparação entre diferentes conjuntos de dados de futebol disponíveis na literatura e em repositórios públicos até a definição da base utilizada. Inicialmente, são analisadas as alternativas consideradas, destacando-se a cobertura temporal, o nível de detalhamento das estatísticas de partida e a acessibilidade dos dados. Em seguida, são apresentadas as etapas de limpeza, normalização e reestruturação do banco, incluindo a exclusão de ligas com registros incompletos de eventos, a transformação dos campos em formato XML em tabelas relacionais e a verificação de consistência do processo de *Extract, Transform, Load* (ETL), que significa Extração, Transformação e Carga, garantindo a integridade das informações utilizadas nas análises posteriores.

4.1.1 Seleção do dataset

A Tabela 1 apresenta um comparativo entre os principais conjuntos de dados considerados. Cada coluna indica a presença ou ausência de determinados tipos de informação nos *datasets* analisados. A coluna Ligas informa se a base registra a qual campeonato ou competição cada partida pertence (por exemplo, diferentes ligas nacionais ou continentais), enquanto Equipes indica se há identificação dos times envolvidos em cada jogo. A coluna Resultados representa os desfechos das partidas (vitória, empate ou derrota, bem como o placar), e Datas indica se o conjunto de dados contém o registro temporal dos jogos. As demais colunas referem-se a estatísticas das partidas, tais como gols, posse de bola, número de escanteios, faltas cometidas, cartões aplicados, cruzamentos e finalizações. Dessa forma, a tabela sintetiza quais

atributos estão disponíveis em cada fonte, permitindo avaliar em que medida cada *dataset* atende às necessidades deste estudo.

Tabela 1 – Bases de dados analisadas para seleção do conjunto principal.

Dataset	**Fonte**	**Ligas**	**Equipes**	**Resultados**	**Datas**	**Gols**	**Posse**	**Escanteios**	**Faltas**	**Cartões**	**Cruzamentos**	**Finalizações**
European Soccer Database	Kaggle	X	X	X	X	X	X	X	X	X	X	X
EPL Results	Kaggle	X	X	X								
StatsBomb Open Data	GitHub	X	X	X	X	X	X		X		X	X
Football-Data.co.uk	football-data.co.uk	X	X	X	X		X		X			
Brazilian Soccer Database	Kaggle	X	X	X	X				X			

Fonte: Elaborado pelo autor, 2025.

Dentre as alternativas levantadas, optou-se pela utilização do European Soccer Database, disponível na plataforma Kaggle. Essa base se destacou por apresentar estatísticas detalhadas de partidas, abrangendo onze ligas europeias (como a inglesa, alemã, italiana, espanhola e portuguesa), no período de 2008 a 2016. Os atributos disponíveis incluem dados como gols, posse de bola, escanteios, cruzamentos, faltas e cartões, elementos essenciais para análises de desempenho. A base também apresenta estrutura consistente, acesso gratuito e compatibilidade com bibliotecas populares de ciência de dados, o que favoreceu sua escolha.

Outras bases, como a EPL Results 1993–2018, embora com longo alcance temporal, apresentam apenas dados básicos como placar, datas e nomes dos times, o que limita sua utilidade para análise tática ou comportamental.

A StatsBomb Open Data, embora rica e frequentemente utilizada em estudos recentes, exige esforço adicional de transformação e padronização de dados, o que pode ser um entrave em projetos com escopo limitado de tempo. Isso ocorre porque a base é disponibilizada em múltiplos arquivos, em geral no formato JSON, organizados em nível de competições, partidas, escalações e eventos, e não em uma única tabela já agregada por jogo. Para que os dados sejam utilizados no tipo de análise proposto neste trabalho, seria necessário integrar esses arquivos, tratar chaves e identificadores, normalizar os campos e construir, por meio de rotinas de pré-processamento, uma estrutura tabular com as estatísticas de interesse para cada partida. A Football-Data.co.uk oferece informações sobre ligas, equipes, resultados e

gols, mas dispõe de um conjunto mais restrito de estatísticas de partida, o que reduz seu potencial para a identificação de padrões de desempenho.

Bases nacionais também foram consideradas, como o Brazilian Soccer Database, que cobre o Campeonato Brasileiro entre 2003 e 2022. No entanto, apesar de sua ampla cobertura temporal, os atributos oferecidos são mais restritos em relação à proposta desta pesquisa.

4.1.2 Limpeza e remoção de ligas inconsistentes

A base de dados escolhida, em sua forma original, apresentava redundâncias, atributos desnecessários e campos armazenados em formato *Extensible Markup Language* (XML) que significa Linguagem de Marcação Extensível, o que inviabilizava sua utilização direta para fins analíticos e para a aplicação das técnicas propostas neste trabalho. Dessa forma, tornou-se necessário realizar um processo de normalização e transformação com o objetivo de reestruturar o banco, eliminar inconsistências e adaptar sua modelagem às boas práticas de bancos de dados relacionais.

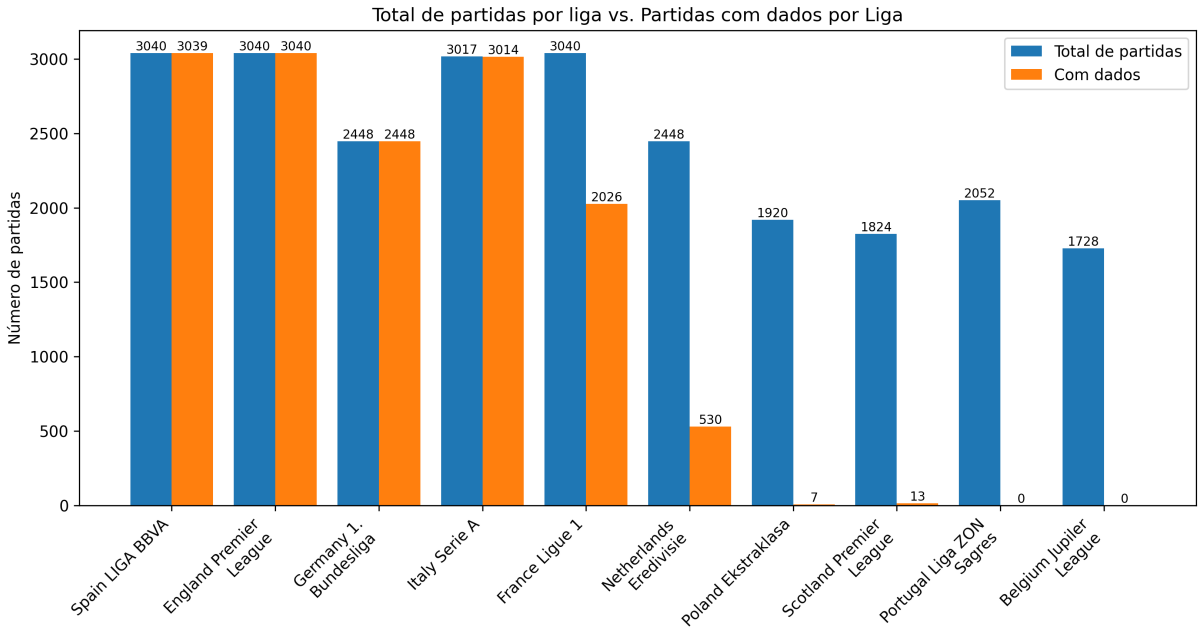
Esse processo envolveu etapas de limpeza, padronização e reorganização dos dados, assegurando a integridade e a coerência entre as tabelas. Inicialmente, foram identificadas e removidas colunas com informações não essenciais ao escopo do estudo, como os campos de *odds* referentes às casas de apostas, que não possuíam relevância para os objetivos propostos.

Em seguida, procedeu-se à avaliação da completude e consistência das ligas presentes na base. Como pode ser visto na Figura 6, onde determinadas competições como a Liga Polonesa (*Poland Ekstraklasa*), Liga Escocesa (*Scotland Premier League*), Liga Portuguesa (*Portugal Liga ZON Sagres*) e a Liga Belga (*Belgium Jupiler League*), apresentavam registros escassos ou incompletos nos campos de eventos (como *gols (goal), chutes a gol (shoton), chutes fora do gol (shotoff), faltas cometidas (foulcommit), cartões (card), cruzamentos (cross), escanteios (corner) e posse de bola (possession)*). A fim de evitar distorções estatísticas e preservar a qualidade das análises, optou-se por excluir essas ligas e suas respectivas partidas da base de dados. Com isso, as 6 ligas escolhidas para o desenvolvimento deste trabalho foram a Liga Inglesa (*England Premier League*), Liga Francesa (*France Ligue 1*), Liga Alemã (*Germany 1. Bundesliga*), Liga Italiana (*Italy Serie A*), Liga Holandesa (*Netherlands Eredivisie*) e a Liga Espanhola (*Spain LIGA BBVA*).

4.1.3 Processo de ETL dos XMLs e validação

Após essa etapa, foi implementado um processo de *Extract, Transform, Load* (ETL), que significa Extração, Transformação e Carga, destinado aos campos

Figura 6 – Total de partidas por liga vs. Partidas com dados por Liga.

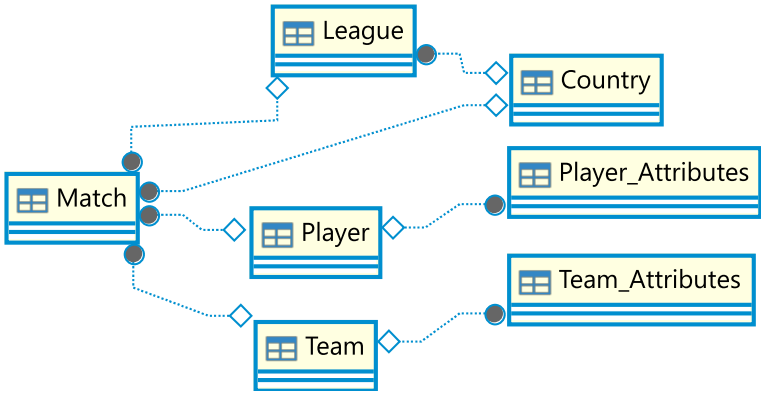


Fonte: Elaborado pelo autor, 2025.

que continham dados no formato XML da tabela *Match*. Essa transformação permitiu converter estruturas hierárquicas e não relacionais em tabelas normalizadas, estabelecendo vínculos diretos com a tabela principal por meio de chaves estrangeiras.

O processo de ETL foi desenvolvido com uso da linguagem Python, utilizando a biblioteca *xml* para leitura, análise e transformação dos dados. Esse processo foi responsável por converter os elementos XML contidos de forma agrupada na coluna *events* da base original, em estruturas relacionais normalizadas, de modo a preservar as informações e garantir consistência com a tabela principal. A Figura 7 mostra a estrutura da base de dados antes do processo de ETL.

Figura 7 – Diagrama da base de dados antes do processo de ETL.

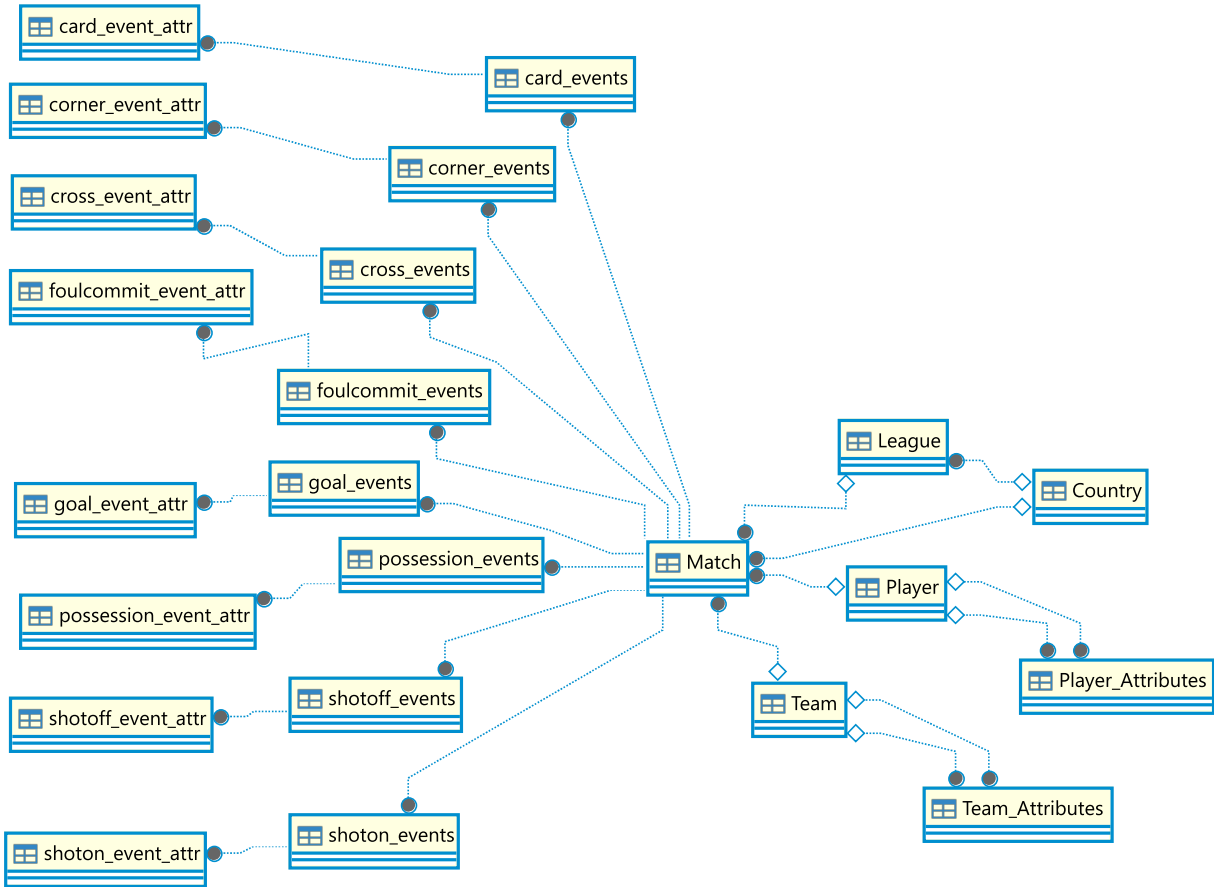


Fonte: Elaborado pelo autor, 2025.

A Figura 8 mostra a estrutura da base de dados após o término da transformação. Foram criadas novas tabelas específicas para cada tipo de

evento: *goal_events, shoton_events, shotoff_events, foulcommit_events, card_events, cross_events, corner_events e possession_events*. Essas tabelas armazenam apenas os campos considerados essenciais de cada evento, permitindo consultas diretas e eficientes. Complementarmente, foram criadas as tabelas auxiliares **_event_attr*, responsáveis por armazenar os atributos adicionais encontrados nos XMLs e não contemplados nos campos principais. Essas tabelas seguem um formato de armazenamento chave–valor, utilizando os campos *attr_key* e *attr_value*, garantindo que nenhum dado do XML seja perdido, mesmo que sua estrutura varie entre partidas ou ligas.

Figura 8 – Diagrama da base de dados após o processo de ETL.



Fonte: Elaborado pelo autor, 2025.

Cada tabela de eventos foi projetada com uma chave estrangeira (FK) referenciando *Match(id)*, assegurando sua integridade referencial. Além disso, foi incluída a coluna *raw_tag* em cada tabela de eventos, destinada a armazenar o XML original de cada registro individual. Essa coluna garante rastreabilidade e auditabilidade, permitindo a revalidação do processo de transformação ou o reprocessamento futuro dos dados, caso seja necessário.

Após a execução do processo de Extração, Transformação e Carga, foi realizada uma etapa de validação da consistência dos dados extraídos. Essa validação tem como objetivo garantir que todos os eventos originalmente presentes nos campos

XML da tabela *Match* fossem devidamente convertidos para as novas tabelas relacionais. Para isso, elaborou-se uma consulta SQL de verificação que contabilizou, para cada tipo de evento, o número total de ocorrências da *tag <event>* dentro dos campos XML e comparou esse valor com a quantidade de registros inseridos nas respectivas tabelas de destino criadas pelo ETL.

Os resultados obtidos evidenciaram correspondência exata entre as contagens, conforme apresentado na Tabela 2, o que comprova a integridade e a completude do processo de extração.

Tabela 2 – Eventos XML e os registros extraídos pelo processo de ETL

Tipo de evento	**Eventos extraídos**
goal	39.929
shoton	93.565
shotoff	95.076
foulcommit	218.638
card	62.001
cross	284.059
corner	87.629
possession	34.745

Fonte: Elaborado pelo autor, 2025.

A partir da base consolidada[1], contendo apenas as ligas com registros completos de eventos e com os dados em formato XML transformados em tabelas relacionais, foram construídos diferentes vetores de atributos para representar o desempenho recente das equipes. Com esses vetores, foram conduzidos experimentos de agrupamento em múltiplas fases, incluindo testes preliminares com atributos brutos diretamente derivados das estatísticas de partida, experimentos principais com janelas deslizantes e variações de conjuntos de atributos, além de análises por liga e por meio da técnica PCA. As subseções seguintes descrevem esses experimentos e discutem os resultados obtidos.

4.2 Vetores de atributos baseados em janelas de cinco partidas

A primeira etapa de experimentação concentrou-se na construção de vetores de atributos derivados de janelas temporais de cinco partidas por equipe. O objetivo foi avaliar se estatísticas recentes de desempenho, agrupadas em blocos de jogos, seriam capazes de induzir uma estrutura de agrupamento consistente no espaço de atributos relacionada ao resultado das partidas.

[1] https://www.kaggle.com/datasets/iurirodrigues17/tcc-futebol-agrupamento-base

4.2.1 Vetor inicial com janela rígida e atributos brutos

O primeiro vetor construído, denominado *team_windows_5_strict*[2], foi obtido a partir da agregação das estatísticas das cinco partidas imediatamente anteriores para cada observação. Nesse vetor, cada linha representa uma janela de cinco jogos de uma mesma equipe, e os atributos numéricos são organizados por partida, com sufixos que indicam a posição temporal na sequência (*match_1*, *match_2*, . . . , *match_5*). Para cada partida individual foram considerados 19 indicadores técnicos 13 derivados do processo de ETL, incluindo gols marcados e sofridos, número de finalizações e finalizações no alvo, posse de bola, escanteios, cruzamentos, faltas cometidas, cartões amarelos e cartões vermelhos, entre outros. Ao replicar esse conjunto de 19 variáveis para cada uma das cinco partidas da janela, obtém-se um vetor final com $19 \times 5 = 95$ atributos por observação.

Antes da aplicação dos algoritmos de agrupamento, verificando a tabela *team_windows_5_strict*[3] pôde-se observar que a mesma continha 33.310 janelas e 101 colunas (incluindo identificadores e atributos).

Foram avaliados três métodos de agrupamento utilizados na literatura: *K-Means*, *Agglomerative Clustering* (com ligação do tipo *ward*) e *Birch*. Para cada algoritmo, testaram-se diferentes valores de k (número de clusters), especificamente $k \in \{3, 5, 7, 9\}$. A Tabela 3 apresenta as métricas internas de qualidade de cluster para cada configuração: índice de Silhouette, índice de Calinski–Harabasz (CH) e índice de Davies–Bouldin (DB).

Os resultados da Tabela 3 indicam que as melhores configurações, em termos de Silhouette e Calinski–Harabasz, concentram-se em valores baixos de k, particularmente em $k = 3$. Entre os algoritmos testados, o *K-Means* com $k = 3$ apresenta um índice de Silhouette de aproximadamente $0{,}23$ e um valor de Calinski–Harabasz superior a 10.000, sugerindo a existência de alguma estrutura de agrupamento no espaço de atributos quando consideradas janelas de cinco partidas.

Com base nesse resultado, adotou-se o *K-Means* com $k = 3$ como configuração de referência para esse vetor. A distribuição das janelas entre os três *clusters* resultantes é apresentada na Tabela 4.

Para tornar os grupos mais interpretáveis, foi calculada, para cada *cluster*, a média das principais estatísticas de desempenho nas cinco partidas da janela, conforme resumido na Tabela 5. Nessa tabela as colunas GP, GC, C e CA correspondem, respectivamente, a gols pró, gols contra, chutes e chutes no alvo. Os valores indicam que os Clusters 1 e 2 reúnem janelas em que as equipes apresentam valores elevados de chutes, posse de bola, faltas e cartões, ao passo que o Cluster 0 concentra

[2] https://www.kaggle.com/datasets/iurirodrigues17/tcc-futebol-agrupamento-data/
[3] https://www.kaggle.com/datasets/iurirodrigues17/tcc-futebol-agrupamento-data/

Tabela 3 – Resultados de agrupamento para o vetor *team_windows_5_strict*.

Algoritmo	k	**Silhouette**	**Calinski–Harabasz**	**Davies–Bouldin**
K-Means	3	0,2309	10.031,88	3,38
K-Means	5	0,0693	6.147,31	3,22
K-Means	7	0,0723	4.514,58	3,08
K-Means	9	0,0259	3.423,13	3,24
Agglomerative (ward)	3	0,2169	10.111,12	2,42
Agglomerative (ward)	5	0,1117	5.899,96	2,82
Agglomerative (ward)	7	0,1062	4.201,76	2,93
Agglomerative (ward)	9	0,0611	3.322,61	3,63
Birch	3	0,2169	10.111,12	2,42
Birch	5	0,1117	5.899,84	2,82
Birch	7	0,1061	4.201,68	2,93
Birch	9	0,0611	3.322,54	3,63

Fonte: Elaborado pelo autor, 2025.

Tabela 4 – Distribuição das janelas por *cluster* para $k = 3$.

Cluster	**Número de janelas**
0	17.123
1	8.000
2	8.187

Fonte: Elaborado pelo autor, 2025.

janelas com estatísticas muito baixas nesses quesitos. Uma inspeção da distribuição por liga revelou ainda que o Cluster 0 é composto majoritariamente por janelas das ligas francesa, alemã, italiana, holandesa e espanhola, enquanto a liga inglesa tende a se concentrar nos Clusters 1 e 2.

Tabela 5 – Resumo interpretável dos clusters para o vetor *team_windows_5_strict*.

Cluster	**GP**	**GC**	**C**	**CA**	**Posse (%)**	**Faltas**	**Amarelos**	**Vermelhos**
0	1,35	1,39	0,69	0,33	3,20	1,03	1,44	0,04
1	1,34	1,37	10,32	5,12	46,56	12,09	2,09	0,06
2	1,44	1,30	10,92	5,45	47,84	12,02	2,02	0,06

Fonte: Elaborado pelo autor, 2025.

Apesar de essa configuração inicial evidenciar padrões distintos de intensidade de jogo, ela ainda não incorpora explicitamente o resultado da partida (*result_current*) na avaliação da qualidade dos agrupamentos. Isso motivou a construção de vetores alternativos, em que o histórico de cinco jogos é representado de forma mais explícita, permitindo comparar diretamente a estrutura de clusters com o desfecho das partidas.

4.2.2 Vetor histórico completo de cinco partidas

Na segunda etapa, foi construído um vetor de atributos que preserva de forma mais detalhada o histórico recente de cada equipe. Esse vetor, denominado *team_windows_hist5*[4], contém, para cada janela de cinco partidas, uma representação das estatísticas de cada jogo da sequência. Para cada uma das cinco partidas anteriores são armazenadas 20 variáveis técnicas, incluindo, entre outras, gols marcados e sofridos, número de finalizações e finalizações no alvo, posse de bola, escanteios, cruzamentos, faltas cometidas e cartões amarelos e vermelhos, organizadas em blocos identificados por sufixos que indicam a posição temporal na janela (*match_1*, *match_2*, . . . , *match_5*). Dessa forma, o vetor *team_windows_hist5* resulta em $5 \times 20 = 100$ atributos numéricos por observação, descrevendo o comportamento da equipe ao longo das cinco partidas mais recentes.

Analizando o vetor construído foi possível ver que o mesmo continha 33.121 janelas e 107 colunas. A variável *result_current*, que codifica o resultado da partida atual como derrota (-1), empate (0) ou vitória (1), foi utilizada na etapa de avaliação, não sendo fornecida aos algoritmos de agrupamento durante o treinamento.

A distribuição dos rótulos reais no conjunto foi a seguinte: 12.324 derrotas, 8.393 empates e 12.404 vitórias, indicando um balanço relativamente equilibrado entre vitórias e derrotas, com menor proporção de empates. Nesta etapa, nenhuma técnica de clusterização foi aplicada. Os rótulos reais foram utilizados como uma partição teórica, com o objetivo de avaliar se as classes apresentam separação intrínseca no espaço de atributos. As métricas internas obtidas foram:

- Silhouette (rótulo real): $-0{,}0033$;
- Calinski–Harabasz (rótulo real): $31{,}25$;
- Davies–Bouldin (rótulo real): $42{,}40$.

O índice de Silhouette ligeiramente negativo, aliado ao valor elevado de Davies–Bouldin, indica que, mesmo quando se considera o rótulo supervisionado como referência, as classes de derrota, empate e vitória não formam grupos naturalmente bem separados no espaço de atributos definido pelo vetor *hist5*.

Na sequência, foi aplicado o algoritmo *K-Means* com $k = 3$ clusters. As métricas internas obtidas para essa configuração foram:

- Silhouette (K-Means): $0{,}2190$;
- Calinski–Harabasz (K-Means): $9.217{,}98$;
- Davies–Bouldin (K-Means): $3{,}45$.

Esses valores sugerem a existência de alguma estrutura de agrupamento no vetor *hist5*, ainda que menos pronunciada do que no caso do vetor inicial *team_windows_5_strict*. Para avaliar a relação entre os *clusters* encontrados e o des-

[4] https://www.kaggle.com/datasets/iurirodrigues17/tcc-futebol-agrupamento-data

fecho das partidas, foram calculados ARI e NMI entre os rótulos de cluster e a variável *result_current*, resultando em:

- ARI (clusters vs. rótulo real): $0{,}0007$;
- NMI (clusters vs. rótulo real): $0{,}0011$.

No caso do ARI, valores próximos de 0 correspondem à expectativa média de um particionamento aleatório, enquanto valores mais elevados (por exemplo, acima de $0{,}2$ ou $0{,}3$) indicam que há alguma concordância estrutural entre os agrupamentos e as classes de referência. Assim, um ARI de $0{,}0007$ é, na prática, indistinguível do que se obteria ao atribuir as janelas de partida a *clusters* de forma aleatória. De forma análoga, a NMI varia no intervalo $[0, 1]$, em que 0 indica independência entre as partições e 1 indica coincidência perfeita; uma NMI de $0{,}0011$ significa que saber a qual *cluster* uma observação pertence praticamente não reduz a incerteza sobre o resultado da partida (vitória, empate ou derrota). Em conjunto, esses valores evidenciam que a relação entre os *clusters* obtidos e os rótulos reais de resultado é, do ponto de vista estatístico, praticamente nula.

Essa conclusão é consistente com a matriz de contingência entre clusters e *result_current*, apresentada na Tabela 6.

Tabela 6 – Matriz de contingência entre clusters e resultado para o vetor *hist5*.

Cluster	**Derrota (-1)**	**Empate (0)**	**Vitória (1)**
0	6.562	4.354	6.159
1	3.102	2.018	2.998
2	2.660	2.021	3.247

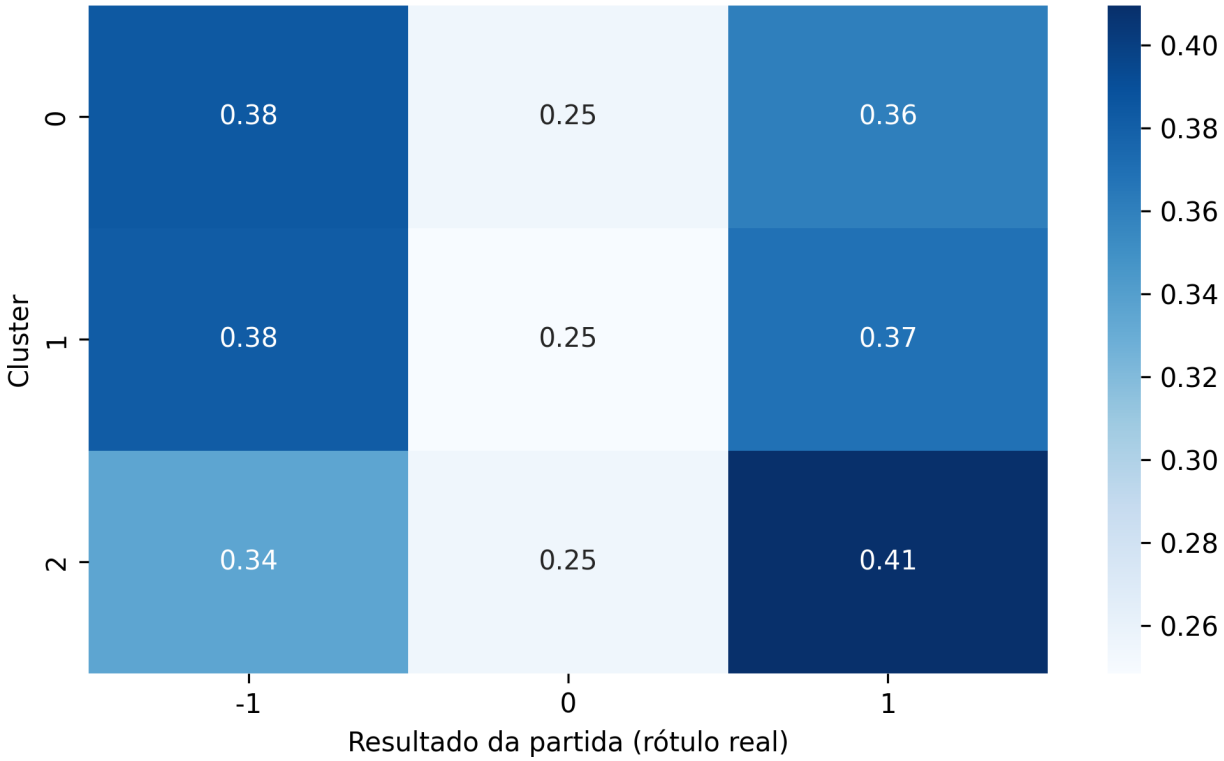
Fonte: Elaborado pelo autor, 2025.

A Tabela 6 mostra que os três clusters apresentam combinações significativas de derrotas, empates e vitórias, sem que nenhum deles se destaque como um grupo “puro” de um único tipo de resultado. Em termos proporcionais, cada cluster contém entre aproximadamente 33% e 38% de derrotas, entre 25% e 26% de empates e entre 36% e 41% de vitórias.

A Figura 9 apresenta, em forma de mapa de calor, a matriz de proporções de vitórias (1), empates (0) e derrotas (-1) em cada *cluster* obtido com o vetor *hist5*. Observa-se que, em todas as linhas, as proporções de resultados são muito próximas, sem que nenhum *cluster* concentre de forma clara um tipo específico de desfecho. Essa visualização reforça a evidência numérica da Tabela 6, indicando que, mesmo com o histórico de cinco partidas, os agrupamentos não se alinham a classes bem definidas de resultado, o que é coerente com os baixos valores de ARI e NMI reportados nesta subseção.

Em síntese, o vetor *hist5* melhora a capacidade de modelar a estrutura interna dos dados em relação à separação induzida diretamente pelo rótulo de resul-

Figura 9 – Histórico de 5 partidas: distribuição dos resultados por *cluster* ($k = 3$).



Fonte: Elaborado pelo autor, 2026.

tado, mas os *clusters* encontrados pelo *K-Means* permanecem pouco alinhados com as classes de derrota, empate e vitória. Essa evidência motivou a construção de um vetor reduzido, com agregações específicas, buscando compactar a informação relevante e reduzir possível redundância entre atributos.

4.2.3 Vetor reduzido com histórico e agregados

Na terceira etapa, buscou-se reduzir a dimensionalidade e a redundância do vetor histórico de cinco partidas, preservando ao mesmo tempo informações consideradas relevantes pela literatura, como médias de gols, finalizações, posse de bola e cartões no horizonte recente de jogos. Para isso, foi construído o vetor *team_windows_hist5_reduced*[5], obtido em duas etapas complementares. Primeiro, selecionou-se um subconjunto de indicadores técnicos por partida, como gols marcados e sofridos, número de finalizações e finalizações no alvo, escanteios, cruzamentos, faltas cometidas e cartões (amarelos e vermelhos), além da posse de bola, que são mantidos explicitamente para cada uma das cinco partidas anteriores. Esses atributos, replicados ao longo da janela temporal, resultam em $5 \times 10 = 50$ variáveis que registram o comportamento da equipe jogo a jogo. Em seguida, foram adicionadas estatísticas agregadas calculadas sobre o horizonte das cinco partidas, incluindo,

[5] https://www.kaggle.com/datasets/iurirodrigues17/tcc-futebol-agrupamento-data

por exemplo, totais e médias de gols pró e contra, número médio de finalizações, posse de bola média e indicadores disciplinares médios nesse período. Essas agregações contribuem com mais 15 atributos, totalizando 65 variáveis numéricas no vetor *team_windows_hist5_reduced*.

O experimento *clustering_hist5_reduced.py*[6] utilizou a tabela *team_windows_hist5_reduced*[7], contendo 33.121 janelas e 72 colunas. A distribuição de rótulos de resultado manteve-se idêntica à observada no vetor *hist5*: 12.324 derrotas, 8.393 empates e 12.404 vitórias.

Assim como nos experimentos anteriores, avaliou-se inicialmente a qualidade da separação imposta pelos rótulos reais no espaço de atributos. As métricas obtidas foram:

- Silhouette (rótulo real): $-0{,}0071$;
- Calinski–Harabasz (rótulo real): $67{,}76$;
- Davies–Bouldin (rótulo real): $29{,}79$.

Os valores obtidos reforçam a ausência de estrutura de agrupamento bem definida quando se consideram as classes de derrota, empate e vitória como partições no espaço vetorial. O índice de Silhouette igual a $-0{,}0071$, em uma escala que varia de -1 a 1, indica que, em média, as observações estão tão próximas (ou até mais próximas) de instâncias de outras classes quanto das instâncias da sua própria classe, o que é incompatível com fronteiras entre os três grupos. De forma consistente, o índice de Davies–Bouldin em torno de $29{,}79$, cujo ideal seria valores próximos de zero, indicando *clusters* compactos e bem separados, revela sobreposição entre as distribuições de derrota, empate e vitória nesse espaço de atributos. Já o valor de Calinski–Harabasz $(67{,}76)$ é modesto e, isoladamente, pouco informativo, mas, em conjunto com a Silhouette negativa e o Davies–Bouldin elevado, corrobora a interpretação de que as classes de resultado não se organizam como grupos naturalmente separados no vetor reduzido.

Em seguida, foi aplicado o *K-Means* com $k = 3$ clusters. As métricas internas do agrupamento foram:

- Silhouette (K-Means): $0{,}1523$;
- Calinski–Harabasz (K-Means): $8.567{,}90$;
- Davies–Bouldin (K-Means): $2{,}42$.

Comparado ao vetor *hist5*, observa-se que o vetor *hist5_reduced* apresenta um índice de Silhouette menor, porém com melhoria no índice de Davies–Bouldin, indicando *clusters* ligeiramente mais compactos. Para avaliar a relação entre os *clusters* e o resultado das partidas, foram novamente calculados ARI e NMI:

- ARI (*clusters* vs. rótulo real): $0{,}0023$;

[6] https://github.com/iurirodrigues17/tcc-futebol-agrupamento
[7] https://www.kaggle.com/datasets/iurirodrigues17/tcc-futebol-agrupamento-data

- NMI (*clusters* vs. rótulo real): $0{,}0019$.

Embora esses valores sejam ligeiramente superiores aos obtidos com o vetor *hist5*, eles permanecem próximos de zero, indicando concordância praticamente nula entre a estrutura de clusters e as classes de resultado. A Tabela 7 apresenta a matriz de contingência entre os clusters do *K-Means* e o rótulo *result_current*.

Tabela 7 – Matriz de contingência entre clusters e resultado para o vetor *hist5_reduced*.

Cluster	**Derrota (-1)**	**Empate (0)**	**Vitória (1)**
0	3.212	2.260	3.526
1	5.399	3.795	5.948
2	3.713	2.338	2.930

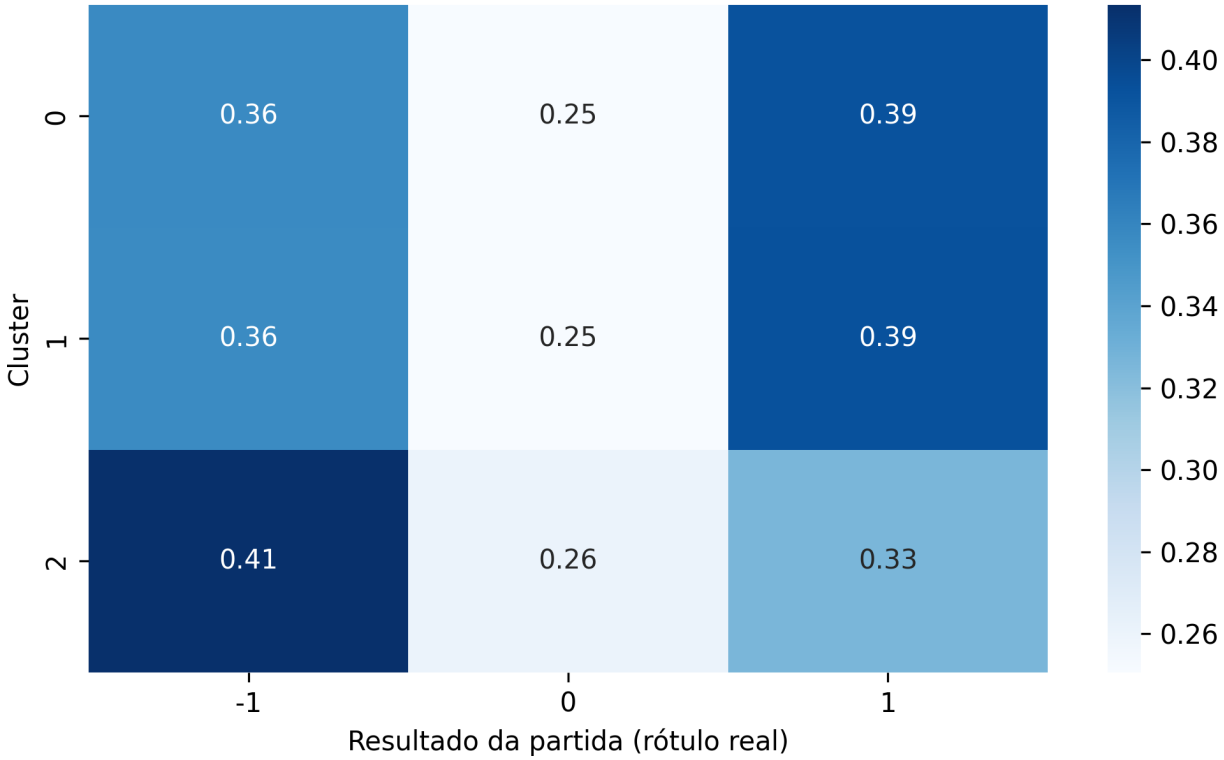
Fonte: Elaborado pelo autor, 2025.

Em termos proporcionais, os Clusters 0 e 1 apresentam distribuições semelhantes, com aproximadamente $35{,}7\%$ de derrotas, 25% de empates e 39% de vitórias, enquanto o Cluster 2 concentra uma proporção maior de derrotas (cerca de $41{,}3\%$) e menor de vitórias (em torno de $32{,}6\%$). Ainda assim, nenhum dos clusters se caracteriza como um grupo fortemente associado a um único tipo de resultado, o que reforça a dificuldade de alinhar a estrutura de agrupamento descoberta com as categorias de derrota, empate e vitória.

A Figura 10 mostra o mapa de calor da matriz de proporções de resultados para o vetor *hist5* reduzido. De modo semelhante ao experimento anterior, as linhas da matriz apresentam proporções muito próximas entre si, sem que um *cluster* se destaque como dominante em vitórias ou derrotas. Isso indica que a seleção de um subconjunto de atributos, incluindo agregados como médias de posse de bola e cartões, melhora a compacidade interna dos grupos, mas não produz *clusters* claramente associados ao desfecho das partidas. Novamente, o comportamento visual é consistente com os valores de ARI e NMI observados.

A combinação de métricas internas e externas sugere que o vetor *hist5_reduced* captura padrões de desempenho recentes que geram alguma estrutura geométrica nos dados (índices de Silhouette positivos para o *K-Means*), mas que tais padrões não são dominados pelo desfecho da partida. Para investigar se a variabilidade entre ligas poderia estar influenciando essa estrutura, foi conduzida uma análise específica por competição, utilizando o mesmo vetor *hist5_reduced*, cujos resultados são apresentados na subseção seguinte.

Figura 10 – Histórico de 5 partidas (vetor reduzido): resultados por *cluster* ($k = 3$).



Fonte: Elaborado pelo autor, 2026.

4.2.4 Análise por liga com o vetor reduzido

Os resultados apresentados nas subseções anteriores mostraram que, mesmo utilizando vetores de atributos baseados no histórico recente de cinco partidas (*hist5* e *hist5_reduced*), os clusters obtidos pelo algoritmo *K-Means* não se alinham de forma consistente com o resultado da partida (derrota, empate ou vitória). Uma hipótese natural é que a grande heterogeneidade entre ligas, em termos de estilo de jogo, intensidade, número médio de gols e padrão de marcação de faltas e cartões, possa estar mascarando estruturas mais claras quando todas as competições são analisadas conjuntamente.

Para investigar essa possibilidade, foi conduzido um experimento adicional no qual o algoritmo *K-Means* foi aplicado separadamente em cada liga, utilizando o vetor reduzido *team_windows_hist5_reduced*. Esse procedimento foi implementado no *script clustering_hist5_reduced_by_league.py*[8].

Para cada liga, foram utilizados como entrada os mesmos 65 atributos numéricos do vetor *hist5_reduced* detalhadas anteriormente, restringindo-se as janelas às observações daquela competição. Em seguida, calculou-se, em dois cenários distintos:

- as métricas internas (Silhouette, Calinski–Harabasz e Davies–Bouldin)

[8] https://github.com/iurirodrigues17/tcc-futebol-agrupamento

associadas à separação induzida diretamente pelo rótulo real (*result_current* $\in \{-1, 0, 1\}$), tratando derrota, empate e vitória como três grupos fixos;
- as métricas internas e externas para o agrupamento obtido pelo *K-Means* com $k = 3$ *clusters*, treinado sem utilizar o rótulo de resultado.

Em todas as ligas, a separação imposta pelos rótulos reais apresentou índice de Silhouette negativo, com valores entre aproximadamente $-0{,}0057$ e $-0{,}0103$, e índices de Davies–Bouldin relativamente altos (por exemplo, $23{,}73$ na *Premier League* e $25{,}61$ na *LIGA BBVA*). Esses resultados corroboram o achado global: mesmo quando consideradas individualmente, as classes de derrota, empate e vitória não formam grupos naturalmente bem definidos no espaço de atributos construído a partir do histórico de cinco partidas.

Ao aplicar o *K-Means* com $k = 3$ em cada liga, observou-se uma melhora nas métricas internas em comparação com a partição induzida pelo rótulo real. A Tabela 8 apresenta um resumo das principais métricas por competição, considerando o agrupamento não supervisionado.

Tabela 8 – Métricas de agrupamento por liga com *hist5_reduced* e *K-Means* com $k = 3$.

Liga	**ID**	n **janelas**	**Silhouette**	**CH**	**DB**	**ARI / NMI**
Inglesa	1729	5.910	0,0420	400,81	4,01	0,0220 / 0,0200
Francesa	4769	5.905	0,3881	2.276,62	2,71	-0,0015 / 0,0023
Alemã	7809	4.746	0,1060	838,74	2,71	0,0043 / 0,0047
Italiana	10257	5.874	0,1287	1.130,35	2,87	0,0054 / 0,0042
Holandesa	13274	4.771	0,1606	2.179,81	1,87	0,0147 / 0,0107
Espanhola	21518	5.915	0,1633	1.215,54	2,36	0,0074 / 0,0146

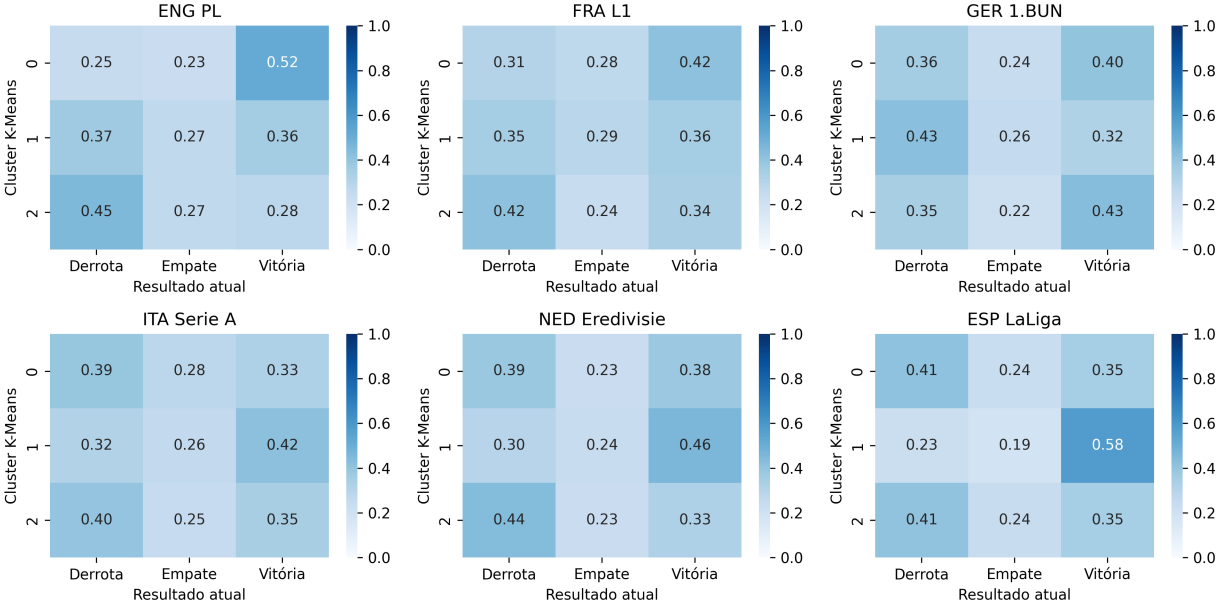
Fonte: Elaborado pelo autor, 2025.

A Tabela 8 mostra que:
- em todas as ligas, o *K-Means* encontra alguma estrutura de agrupamento, com índices de Silhouette positivos variando de $0{,}0420$ (Premier League) a $0{,}3881$ (Ligue 1);
- ligas como *Netherlands Eredivisie* e *Spain LIGA BBVA* apresentam combinações de Silhouette mais elevadas e baixos valores de Davies–Bouldin, sugerindo clusters relativamente mais compactos e bem separados no espaço de atributos;
- apesar disso, os índices externos ARI e NMI, que medem a concordância entre clusters e o rótulo de resultado, permanecem próximos de zero em todas as ligas (por exemplo, ARI $= 0{,}0220$ e NMI $= 0{,}0200$ na Premier League; ARI $= 0{,}0147$ e NMI $= 0{,}0107$ na Eredivisie).

Essa ausência de alinhamento entre clusters e desfecho da partida também é visto ao se inspecionar as matrizes de contingência *cluster vs.* resultado em cada liga. Na *England Premier League*, por exemplo, um dos clusters (Cluster 0) apresenta uma proporção maior de vitórias (cerca de $51{,}7\%$), mas ainda reúne derrotas ($24{,}9\%$) e empates ($23{,}4\%$) em quantidade significativa. Outro grupo (Cluster 2) concentra uma fração mais elevada de derrotas (cerca de $44{,}9\%$), mas continua agregando empates e vitórias em proporções não desprezíveis. Padrões semelhantes são observados nas demais ligas, em que nenhum cluster se caracteriza como um grupo "puro" de vitórias, derrotas ou empates.

A Figura 11 sintetiza esse comportamento, apresentando, para cada liga, a distribuição proporcional de derrotas, empates e vitórias em cada um dos três *clusters* obtidos a partir do vetor *team_windows_hist5_reduced*. Observa-se que, em todas as ligas, os *clusters* mantêm combinações de diferentes desfechos, reforçando a inexistência de grupos "puros" associados exclusivamente a vitórias, empates ou derrotas.

Figura 11 – Resultados por *cluster* em cada liga.



Fonte: Elaborado pelo autor, 2026.

A análise por liga com o vetor *hist5_reduced* confirma que:

- há evidências de que as janelas de desempenho recente se organizam em *clusters* com alguma coerência interna em cada competição, conforme indicado pelas métricas internas de qualidade de agrupamento;
- essa estrutura não está fortemente determinada pelo resultado final da partida, uma vez que ARI e NMI permanecem muito baixos, e as matrizes de contingência revelam uma mistura consistente de derrotas, empates e vitórias em todos os *clusters*.

Esses resultados reforçam a conclusão de que, mesmo ao controlar o efeito de heterogeneidade entre ligas, os vetores baseados apenas em histórico de cinco

partidas não são suficientes para induzir agrupamentos que se alinhem de forma clara aos desfechos dos jogos. Esse resultado motivou a etapa seguinte da pesquisa, em que foram construídos vetores de atributos baseados em médias e desvios-padrão em janelas deslizantes de diferentes tamanhos (3, 4 e 5 partidas), buscando representar de forma mais robusta o perfil estatístico recente das equipes.

4.3 Vetores estatísticos em janelas deslizantes de 3, 4 e 5 partidas

Os experimentos apresentados anteriormente mostraram que vetores baseados no histórico das cinco últimas partidas, ainda que incluam contagens agregadas e médias simples de eventos, não foram suficientes para produzir agrupamentos fortemente associados ao resultado das partidas. A partir desses achados, e motivado pela literatura de predição de resultados em futebol, optou-se por construir uma nova família de vetores estatísticos, baseada em janelas deslizantes de diferentes tamanhos.

Enquanto os vetores anteriores utilizavam contagens acumuladas nas últimas partidas, a nova abordagem faz uso de estatísticas de síntese (média e desvio-padrão) calculadas sobre janelas de 3, 4 e 5 jogos. A hipótese é que esse tipo de representação possa capturar o "perfil" recente de desempenho da equipe, por exemplo, médias de gols marcados, volume de finalizações, posse de bola e disciplina (faltas e cartões), bem como a variabilidade desses indicadores ao longo da sequência recente de confrontos.

4.3.1 Construção da tabela team_windows_stats_3_4_5

A partir das tabelas de eventos geradas no processo de ETL, foi construída a tabela derivada *team_windows_stats_3_4_5*[9], que serve de base para os experimentos desta seção. Essa tabela consolida, para cada equipe e partida, estatísticas de desempenho calculadas sobre janelas deslizantes de tamanho 3, 4 e 5 jogos imediatamente anteriores ao confronto atual.

O processo de geração desse vetor foi implementado em um conjunto de rotinas que percorrem, para cada equipe, a sequência cronológica de partidas em cada liga. Para cada jogo, são identificadas as k partidas anteriores ($k \in \{3, 4, 5\}$) em que a equipe atuou, respeitando a ordem temporal. A partir dessas janelas, são calculadas, para o lado "a favor" da equipe, as seguintes quantidades para cada valor de k:

- gols: média e desvio-padrão de gols marcados;
- finalizações: média e desvio-padrão de chutes totais e chutes no alvo;

[9] https://www.kaggle.com/datasets/iurirodrigues17/tcc-futebol-agrupamento-data

- posse de bola: média e desvio-padrão da posse;
- escanteios e cruzamentos: médias e desvios-padrão de escanteios e cruzamentos a favor;
- disciplina: médias e desvios-padrão de faltas cometidas, cartões amarelos e vermelhos.

Ao todo, considerando as três janelas temporais (3, 4 e 5 partidas) e todos os tipos de eventos selecionados, a tabela *team_windows_stats_3_4_5*[10] passa a contar com 54 atributos numéricos derivados de médias e desvios-padrão. Esse número decorre da seguinte combinação: foram consideradas nove categorias de eventos de jogo (gols marcados, finalizações totais, finalizações no alvo, posse de bola, escanteios, cruzamentos, faltas cometidas, cartões amarelos e cartões vermelhos). Para cada uma dessas nove categorias foram calculadas duas estatísticas (média e desvio-padrão) em três horizontes de forma recente (últimos 3, 4 e 5 jogos). Assim, cada categoria de evento gera $2 \times 3 = 6$ atributos, e o conjunto completo resulta em $9 \times 6 = 54$ atributos numéricos. Além desses atributos, a tabela inclui colunas de identificação e contexto (*team_id*, *current_match_id*, *current_date*, *league_id*, *season*, *opponent_id_current* e *result_current*). A versão utilizada nos experimentos de agrupamento contém 33.121 janelas de partida, distribuídas entre as seis ligas europeias consideradas.

Um aspecto importante desse processo foi o tratamento das partidas iniciais de cada equipe em uma temporada. Como o cálculo das janelas requer um número mínimo de jogos anteriores (por exemplo, cinco partidas para a janela de tamanho 5), partidas sem histórico suficiente são descartadas da tabela *team_windows_stats_3_4_5*. Essa decisão segue a prática adotada na literatura de predição de resultados, que estabelece um número mínimo de jogos prévios para que as estatísticas recentes sejam consideradas confiáveis.

A construção desse vetor abre duas linhas principais de investigação, exploradas nas próximas subseções:

- avaliar se os *clusters* obtidos a partir dessas estatísticas em janelas deslizantes apresentam maior alinhamento com o rótulo de resultado (*result_current*) do que nos experimentos preliminares; e
- analisar em que medida os agrupamentos refletem diferenças estruturais entre ligas, isto é, se as janelas de desempenho tendem a se organizar em “tipos de liga” com perfis estatísticos semelhantes.

Nas subseções seguintes, são apresentados os resultados dos experimentos considerando diferentes combinações de atributos, bem como análises complementares por liga e uma investigação baseada em PCA para interpretar a importância relativa dos atributos na formação dos clusters.

[10] https://www.kaggle.com/datasets/iurirodrigues17/tcc-futebol-agrupamento-data

4.3.2 Experimentos globais com K-Means ($k = 3$) e combinações de atributos

Com a tabela *team_windows_stats_3_4_5*[11] construída, foram conduzidos experimentos de agrupamento globais utilizando o algoritmo *K-Means* com $k = 3$ *clusters*. O número de grupos foi mantido alinhado ao rótulo de resultado disponível na base (*result_current* $\in \{-1, 0, 1\}$, representando derrota, empate e vitória), o que facilita a comparação posterior entre os clusters obtidos e o desfecho real das partidas.

Em todos os experimentos, os atributos derivados de médias e desvios-padrão foram padronizados (*standard scaling*) antes da aplicação do *K-Means*. Além disso, foram avaliadas quatro combinações de atributos, denominadas *full*, *ofensivo+faltas*, *ofensivo+cartões* e *ofensivo*, construídas a partir de diferentes níveis de inclusão de variáveis relacionadas à disciplina (faltas e cartões) e inspiradas tanto na literatura quanto nos testes preliminares descritos anteriormente.

A primeira combinação, *full*, utiliza todos os atributos estatísticos disponíveis (54 atributos), contemplando as médias e os desvios-padrão de gols, finalizações (totais e no alvo), posse de bola, escanteios, cruzamentos, faltas cometidas e cartões amarelos e vermelhos. Essa configuração busca maximizar a informação fornecida ao algoritmo, incluindo aspectos ofensivos e disciplinares.

A segunda combinação, *ofensivo+faltas*, mantém os atributos ofensivos (gols, finalizações, posse de bola, escanteios e cruzamentos) e inclui faltas, mas remove os atributos de cartões amarelos e vermelhos (42 atributos). O objetivo é avaliar a contribuição das faltas como indicador disciplinar, sem incorporar variáveis associadas a punições.

A terceira combinação, *ofensivo+cartões*, mantém os mesmos atributos ofensivos e inclui cartões amarelos e vermelhos, mas exclui as faltas (48 atributos). Essa alternativa permite analisar se medidas relacionadas a sanções (cartões) diferenciam padrões de desempenho de forma distinta da frequência de infrações (faltas).

Por fim, a quarta combinação, *ofensivo*, utiliza apenas atributos ofensivos (gols, finalizações, posse de bola, escanteios e cruzamentos), excluindo faltas e cartões (36 atributos). Essa configuração concentra a análise no comportamento de criação de chances e produção ofensiva, evitando variáveis ligadas à disciplina.

A Tabela 9 resume as principais métricas de qualidade dos agrupamentos obtidos para cada configuração: índice de silhueta, índice de Calinski–Harabasz (CH), índice de Davies–Bouldin (DB) e as medidas de concordância entre clusters e resultado real (ARI e NMI).

Os resultados indicam dois comportamentos complementares:

- do ponto de vista geométrico do espaço de atributos, os valores de silhueta e Calinski–Harabasz são relativamente altos, especialmente quando

[11] https://www.kaggle.com/datasets/iurirodrigues17/tcc-futebol-agrupamento-data

Tabela 9 – Resultados globais do K-Means ($k = 3$) em janelas de 3, 4 e 5 partidas

Configuração	Silhouette	CH	DB	ARI	NMI
(a) Todos os atributos	0,3571	18 275,23	1,46	0,0007	0,0005
(b) Com faltas, sem cartões	0,4652	31 278,50	1,15	0,0007	0,0005
(c) Com cartões, sem faltas	0,3236	15 945,21	1,59	0,0006	0,0005
(d) Sem faltas e sem cartões	0,4355	28 225,68	1,21	0,0007	0,0005

Fonte: Elaborado pelo autor, 2025.

se considera a configuração (b), com faltas e sem cartões, e a configuração (d), sem faltas nem cartões. Isso sugere que os vetores em janelas de 3, 4 e 5 partidas formam grupos internamente coesos e bem separados entre si no espaço de características;

- por outro lado, do ponto de vista supervisionado, as medidas ARI e NMI apresentam valores próximos de zero em todas as configurações, indicando que a rotulagem induzida pelo *K-Means* guarda pouca relação com o resultado atual da partida (*result_current*).

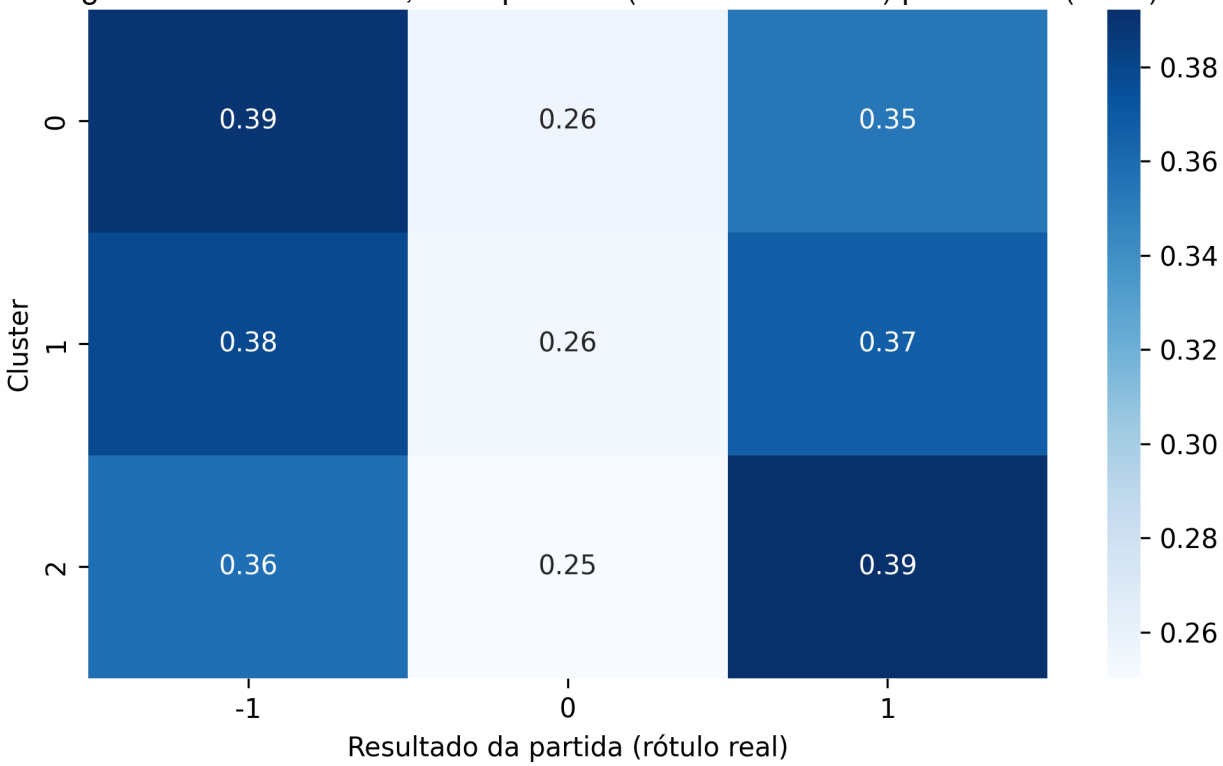
Essa conclusão é reforçada pela análise das contagens cruzadas entre clusters e resultados. Em todas as quatro configurações, a proporção de vitórias, empates e derrotas dentro de cada cluster permanece próxima à distribuição global da base, com valores típicos em torno de 35%–41% de vitórias, 25%–26% de empates e 35%–41% de derrotas em cada grupo. Em outras palavras, nenhum cluster se caracteriza como um "cluster de vitórias" ou "cluster de derrotas". Os grupos formados pelo algoritmo estão separados por padrões estatísticos de desempenho recente, mas esses padrões não se traduzem em uma separação clara dos desfechos das partidas.

Do ponto de vista da engenharia de atributos, os experimentos também sugerem que o uso ou não de estatísticas disciplinares (faltas e cartões) não altera de forma substantiva a capacidade dos clusters em recuperar o rótulo de resultado. Embora os valores de Silhueta e de CH oscilem entre as configurações, a concordância com *result_current* permanece praticamente inalterada (ARI e NMI ≈ 0). Isso reforça a ideia de que, na base considerada e para o tipo de vetor construído, a estrutura principal de agrupamento está relacionada a perfis gerais de desempenho (volume ofensivo, posse, estilo de jogo, disciplina), e não a uma separação direta entre vitórias, empates e derrotas.

A Figura 12 apresenta o mapa de calor da matriz de proporções de vitórias, empates e derrotas por *cluster* para o vetor estatístico em janelas deslizantes de 3, 4 e 5 partidas (médias e desvios-padrão). Assim como nos experimentos com o vetor *hist5*, as proporções por linha são bastante similares, sem um *cluster* claramente associado a "times em boa fase" ou "times em má fase". Essa evidência visual complementa as métricas de ARI e NMI próximas de zero e sustenta a conclusão de que

os agrupamentos extraídos a partir das estatísticas recentes capturam principalmente diferenças gerais de estilo e intensidade de jogo, mas não se traduzem em grupos bem separados em termos do resultado final da partida.

Figura 12 – Janelas de 3, 4 e 5 partidas (médias e desvios) por *cluster* ($k = 3$).



Fonte: Elaborado pelo autor, 2026.

Esses achados justificam, do ponto de vista metodológico, a investigação adicional realizada nas subseções seguintes: (i) a análise dos agrupamentos por liga, buscando verificar se os clusters capturam diferenças de estilo entre as competições; (ii) experimentos com agrupamento em "tipos de liga" (por exemplo, $k = 6$ clusters associados às seis ligas remanescentes); e (iii) uma análise de componentes principais (PCA) para interpretar quais atributos têm maior peso na formação dos grupos, independentemente do resultado das partidas.

4.3.3 Análise por liga dos agrupamentos com vetores em janelas 3–4–5

Para investigar se a relação entre o desempenho recente das equipes e o resultado das partidas se torna mais evidente dentro de cada competição específica, os experimentos com o vetor em janelas de 3, 4 e 5 partidas (*team_windows_stats_3_4_5*[12]) foram repetidos de forma estratificada por liga. Ou seja, para cada uma das seis ligas remanescentes, foi considerado apenas o subconjunto de janelas daquela liga, mantendo-se a mesma engenharia de atributos (54 mé-

[12] https://drive.google.com/file/d/1c1SINRCPTfOICN057HnI7Rj9As0ooF40/view?usp=sharing

dias e desvios-padrão de gols, finalizações, posse de bola, escanteios, cruzamentos, faltas e cartões). Em cada liga, aplicou-se o algoritmo *K-Means* com $k = 3$ clusters. A Tabela 10 apresenta um resumo dessas métricas por liga.

Tabela 10 – Resultados do K-Means ($k = 3$) por liga com vetores em janelas 3–4–5.

Liga	n **janelas**	**Silhouette**	**CH**	**ARI**	**NMI**
ENG PL (ID 1729)	5.910	0,14	771,37	0,0137	0,0143
FRA L1 (ID 4769)	5.905	0,32	5 033,99	$-0{,}0003$	0,0006
GER 1.BUN (ID 7809)	4.746	0,29	2 200,69	0,0002	0,0008
ITA Serie A (ID 10257)	5.874	0,29	2 680,17	0,0025	0,0025
NED Eredivisie (ID 13274)	4.771	0,70	3 966,53	0,0000	0,0028
ESP LaLiga (ID 21518)	5.915	0,29	2 698,71	0,0057	0,0044

Fonte: Elaborado pelo autor, 2025.

A partir desses resultados, observa-se inicialmente que, em todas as ligas, o índice de silhueta do *K-Means* é positivo e, em diversos casos, assume valores moderados (aproximadamente 0,28–0,32 para França, Alemanha, Itália e Espanha). Isso indica que, mesmo com a análise restrita a uma única competição, os vetores em janelas de 3, 4 e 5 partidas formam agrupamentos internamente coesos e relativamente bem separados no espaço de atributos.

Na Eredivisie (ID 13274), o coeficiente de Silhueta atinge cerca de 0,70, sugerindo alguma separação entre três regimes de desempenho recente dentro da liga. Ainda assim, essa separação geométrica não se traduz em uma segmentação nítida dos resultados das partidas, pois os grupos não se associam de modo consistente a vitórias, empates ou derrotas.

Essa limitação também é refletida pelas métricas externas. As medidas de concordância entre *clusters* e o resultado atual (ARI e NMI) permanecem muito próximas de zero em todas as ligas. O maior valor de ARI ocorre na Premier League (aproximadamente 0,014), e a NMI mantém-se tipicamente na faixa de 0,00–0,01, sugerindo dependência fraca entre a partição produzida e o rótulo *result_current*.

A inspeção das contagens cruzadas entre *clusters* e *result_current* reforça essa interpretação. Embora alguns grupos apresentem leve predominância de vitórias ou derrotas, as proporções internas não se afastam de forma expressiva da distribuição global da liga. Na Premier League, por exemplo, observa-se um *cluster* com proporção um pouco maior de vitórias (cerca de 48%), mas ainda contendo uma parcela de empates e derrotas, o que impede caracterizá-lo como um "*cluster* de vitórias".

Em síntese, os vetores em janelas de 3, 4 e 5 partidas capturam padrões recentes de desempenho dentro de cada liga e geram agrupamentos com separação geométrica razoável. Entretanto, esses padrões não se alinham fortemente ao desfecho pontual da partida seguinte, indicando que os *clusters* tendem a refletir regimes

ou perfis de atuação (como maior volume ofensivo, maior posse ou maior intensidade disciplinar) mais do que uma partição direta entre vitória, empate e derrota.

Esse comportamento é consistente com os achados dos experimentos preliminares e reforça a hipótese de que, na base considerada e com o tipo de vetor temporal construído, o agrupamento de janelas de desempenho está mais associado às diferenças de estilo entre equipes e ligas do que à predição direta do desfecho das partidas. A seção seguinte aprofunda essa análise ao comparar explicitamente os clusters globais com os rótulos de liga, bem como ao avaliar um agrupamento específico em "tipos de liga" com $k = 6$ clusters.

4.3.4 Clusters globais vs ligas e agrupamento em "tipos de liga"

Os resultados apresentados nas subseções anteriores mostram que, mesmo após a construção de vetores de desempenho recente em janelas de 3, 4 e 5 partidas, os agrupamentos obtidos por *K-Means* não reproduzem, de forma consistente, os rótulos de resultado das partidas (*result_current*). Por outro lado, as análises exploratórias sugeriam que os clusters pareciam seguir, em boa medida, as diferenças de estilo entre ligas. Para investigar essa hipótese de forma sistemática, foram conduzidos dois experimentos adicionais: (i) comparação direta entre os clusters globais ($k = 3$) e os rótulos de liga; e (ii) um agrupamento específico em "tipos de liga" com $k = 6$ clusters.

4.3.4.1 Clusters globais ($k = 3$) versus rótulos de liga

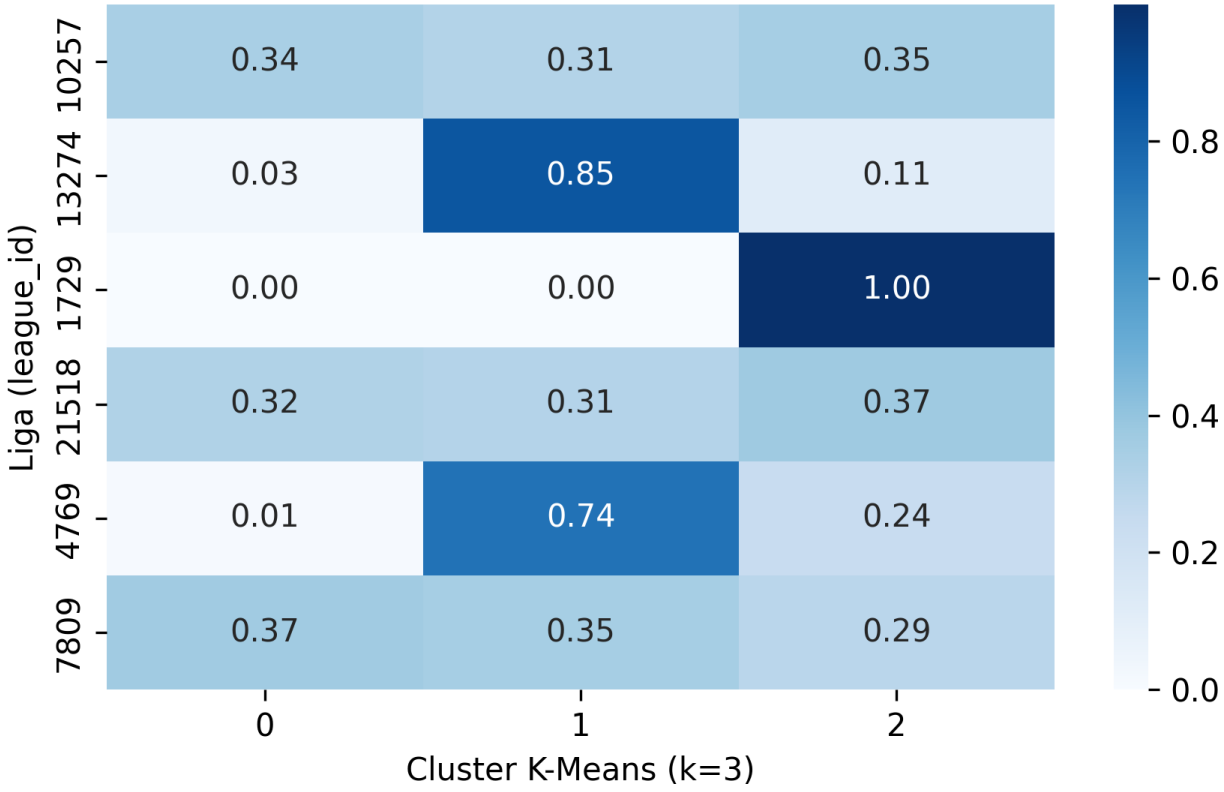
No primeiro experimento desta seção, foi utilizado o conjunto *team_windows_stats_3_4_5*[13]. Sobre esse conjunto, já havia sido ajustado o *K-Means* com $k = 3$ clusters usando as 54 médias e desvios-padrão de gols, finalizações, posse de bola, escanteios, cruzamentos, faltas e cartões. Inicialmente, a concordância entre esses clusters e o resultado atual da partida (*result_current*) mostrou-se praticamente nula, com ARI em torno de $0{,}0007$ e NMI próxima de $0{,}0005$.

Em seguida, os mesmos clusters foram comparados aos rótulos de liga (*league_id*), por meio do script *clustering_stats_3_4_5_cluster_vs_league.py*[14]. Nesse caso, observou-se uma mudança de escala nas métricas de concordância: o ARI passou para $0{,}1573$ e a NMI para $0{,}2174$. Embora esses valores ainda não indiquem uma correspondência perfeita, eles são duas ordens de grandeza maiores do que aqueles obtidos em relação ao resultado de partida, sugerindo uma associação mais forte entre clusters e ligas do que entre clusters e desfechos (*vitória, empate, derrota*).

[13] https://www.kaggle.com/datasets/iurirodrigues17/tcc-futebol-agrupamento-data
[14] https://github.com/iurirodrigues17/tcc-futebol-agrupamento

A Figura 13 sintetiza essa relação por meio de um mapa de calor com as proporções de janelas de cada liga em cada um dos três clusters globais. Cada linha representa uma liga e cada coluna um cluster do *K-Means*.

Figura 13 – Clusters globais por liga para o vetor de 3, 4 e 5 partidas.



Fonte: Elaborado pelo autor, 2026.

Os valores indicam a fração de janelas daquela liga atribuída a cada grupo. A partir dessa matriz de distribuição, observam-se alguns padrões marcantes. Em termos de proporção por liga, tem-se, por exemplo:

- na Premier League (ID 1729), quase todas as janelas (aproximadamente 99,9%) pertencem ao mesmo cluster (cluster 2), indicando que, no espaço de atributos construído, os vetores dessa liga formam um grupo extremamente dominante e coeso;
- na Ligue 1 francesa (ID 4769), cerca de 74% das janelas pertencem ao cluster 1, enquanto na Eredivisie holandesa (ID 13274) aproximadamente 85% das janelas também se concentram nesse mesmo cluster. Isso sugere que essas duas ligas compartilham um “regime” de desempenho recente similar, capturado pelo algoritmo de agrupamento;
- a Bundesliga alemã (ID 7809), a Serie A italiana (ID 10257) e a La Liga espanhola (ID 21518) apresentam distribuições um pouco mais equilibradas entre os três clusters, ainda que com leve predominância de um ou outro grupo. Nessas ligas, as janelas de desempenho parecem ocupar regiões mais diversas no espaço de atributos.

De forma geral, os clusters globais com $k = 3$ não conseguem separar adequadamente vitórias, empates e derrotas, mas exibem uma estrutura alinhada às características de cada liga. Em outras palavras, o algoritmo tende a agrupar janelas de ligas com estilos de jogo semelhantes (por exemplo, maior volume ofensivo, maior número de cartões ou posse de bola típica) muito mais do que a separar diretamente os desfechos das partidas.

4.3.4.2 Agrupamento em "tipos de liga" com $k = 6$

No segundo experimento, buscou-se explorar a hipótese de que existiriam perfis distintos de ligas a partir dos vetores em janelas de 3, 4 e 5 partidas. Para isso, foi aplicado o *K-Means* com $k = 6$ sobre o mesmo conjunto de atributos, de forma a obter seis clusters que pudessem ser interpretados como diferentes "tipos de liga". Os resultados, obtidos via o script *clustering_stats_3_4_5_kmeans6_leagues.py*[15], são resumidos na Tabela 11.

Tabela 11 – Resultados do K-Means ($k = 6$) em "tipos de liga" em janelas 3–4–5

Configuração	**Silhouette**	**CH**	**DB**	**ARI / NMI vs. liga**
Clusters globais ($k = 3$) vs. liga	–	–	–	0,1573 / 0,2174
"Tipos de liga" ($k = 6$)	0,21	11 267,66	1,73	0,1708 / 0,2418

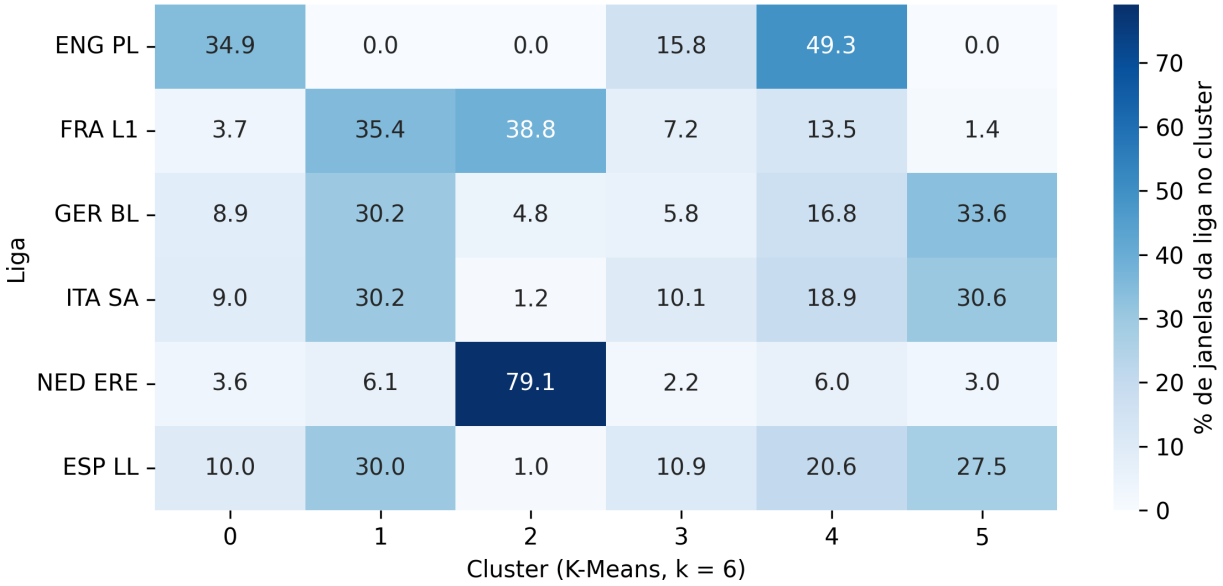
Fonte: Elaborado pelo autor, 2025.

Algumas observações se destacam a partir desses resultados. O valor de silhueta em torno de 0,21 indica que os seis *clusters* formados apresentam separação geométrica razoável, ainda que menos pronunciada do que a observada em alguns casos de análise por liga específica. Além disso, as métricas de concordância com o rótulo de liga aumentam ligeiramente em relação ao caso $k = 3$, atingindo ARI = 0,1708 e NMI = 0,2418, o que sugere que a divisão em seis grupos captura de forma mais refinada diferenças entre as ligas.

A Figura 14 sintetiza a distribuição das janelas de desempenho por liga e por cluster, considerando o *K-Means* com $k = 6$. Observa-se que algumas competições são fortemente associadas a um único perfil: na Eredivisie (NED ERE), por exemplo, cerca de 79% das janelas concentram-se no mesmo cluster, enquanto a Premier League (ENG PL) se distribui principalmente entre dois grupos, com predominância dos clusters 4 e 0. Em contraste, ligas como La Liga (ESP LL), Serie A (ITA SA) e Bundesliga (GER BL) apresentam uma combinação mais equilibrada de três ou mais clusters. Esse padrão visual reforça a interpretação de que os seis grupos funcionam como "tipos de liga", isto é, perfis de desempenho compartilhados em diferentes proporções por cada campeonato.

[15] https://github.com/iurirodrigues17/tcc-futebol-agrupamento

Figura 14 – Distribuição das janelas por liga e cluster (*K-Means* com $k = 6$).



Fonte: Elaborado pelo autor, 2026.

Essa estrutura de mistura indica que os seis clusters resultantes podem ser interpretados como perfis de desempenho característicos, compartilhados em diferentes proporções pelas ligas. Algumas competições aparecem associadas a um único perfil (como a Eredivisie), enquanto outras combinam múltiplos perfis (como Espanha, Itália e Alemanha), sugerindo maior diversidade interna de estilos.

4.3.4.3 Síntese interpretativa

A comparação entre os diferentes cenários permite sintetizar três pontos para a interpretação dos resultados deste trabalho. Em primeiro lugar, os vetores em janelas de 3, 4 e 5 partidas, construídos a partir de estatísticas mencionadas anteriormente, geram agrupamentos com coerência interna razoável, como indicado por valores consistentes de silhueta e Calinski–Harabasz. Esse comportamento é observado tanto na análise global quanto quando a análise é restrita por liga.

Em segundo lugar, esses agrupamentos apresentam concordância praticamente nula com o resultado da partida seguinte, com ARI e NMI próximos de zero. Em contrapartida, observa-se uma associação mais forte com os rótulos de liga, tanto no cenário global com $k = 3$ quanto no experimento específico de “tipos de liga” com $k = 6$.

Por fim, esses achados indicam que, para a base utilizada e para a engenharia de atributos adotada, o agrupamento de janelas de desempenho recente tende a capturar diferenças de estilo e intensidade entre ligas e, em menor grau, padrões internos a cada liga. Assim, os *clusters* obtidos se mostram mais adequados para caracterizar perfis de atuação do que para discriminar diretamente vitórias, empates e

derrotas.

Essa constatação é coerente com a literatura de predição de resultados, segundo a qual desempenhos preditivos costumam requerer modelos supervisionados e a combinação de múltiplas fontes de informação, como contexto de tabela, mando de campo, características do elenco e variáveis contextuais adicionais. Assim, os resultados obtidos reforçam que, neste estudo, o agrupamento desempenha um papel predominantemente exploratório e descritivo, ao identificar perfis de desempenho entre ligas e equipes. Em seguida, a próxima subseção aprofunda a análise do espaço de atributos, examinando quais variáveis mais contribuem para a estrutura dos *clusters* por meio de técnicas de redução de dimensionalidade e de avaliação de importância de atributos, tornando mais transparente a contribuição de cada estatística na formação dos agrupamentos.

4.4 Análise em Componentes Principais

Com o objetivo de investigar a estrutura dos vetores de atributos construídos a partir das janelas deslizantes e apoiar a interpretação dos agrupamentos obtidos, foi aplicada uma PCA sobre o mesmo conjunto de variáveis utilizado nos experimentos de *clustering*.

A matriz de entrada da PCA foi formada pelas 54 estatísticas de média e desvio padrão calculadas nas janelas de 3, 4 e 5 partidas anteriormente, padronizadas via *standardization* (subtração da média e divisão pelo desvio padrão). Dessa forma, todas as variáveis passam a contribuir em uma mesma escala, evitando que atributos com valores absolutos maiores dominem a decomposição.

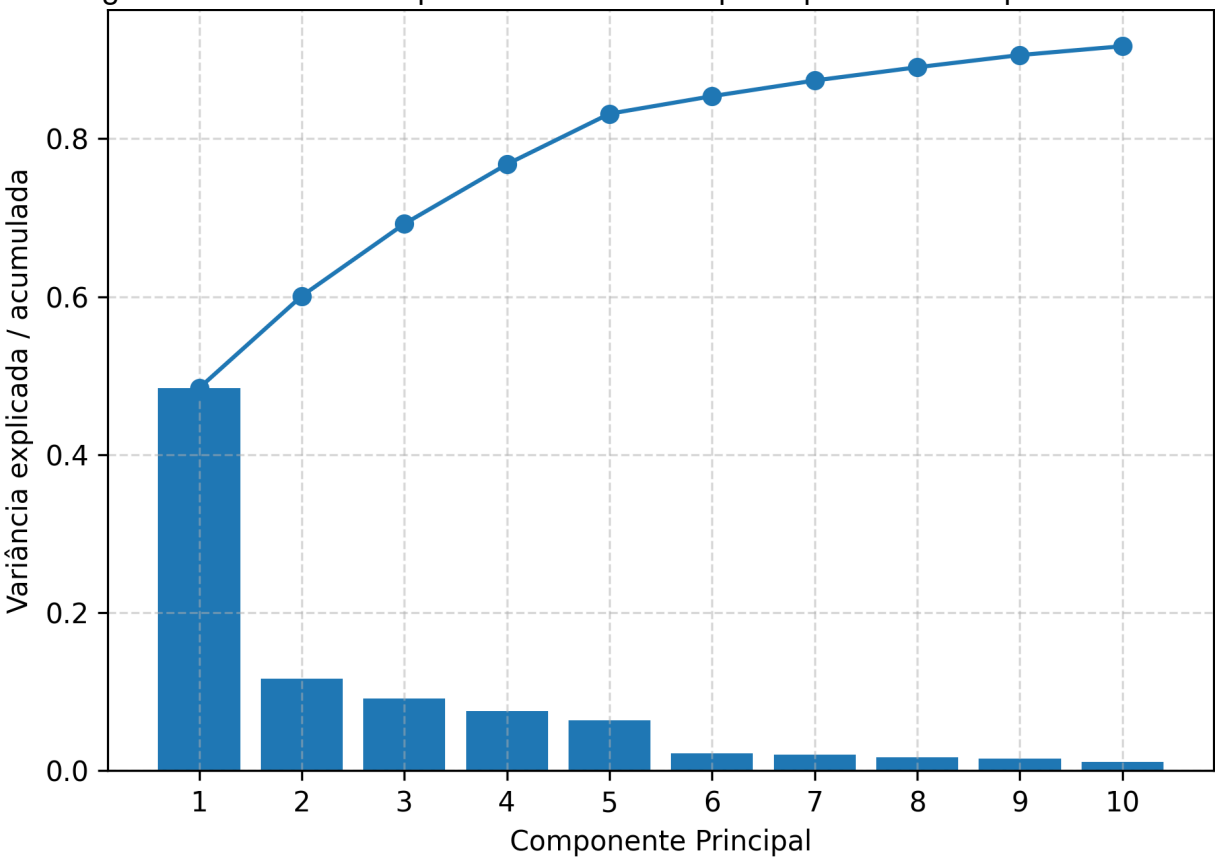
4.4.1 Variância explicada pelas componentes principais

A Figura 15 apresenta o gráfico de sedimentação (*scree plot*) da PCA, no qual as barras indicam a variância explicada por componente e a linha representa a variância acumulada pelas primeiras componentes.

As componentes principais são formadas como combinações lineares dos 54 atributos, de modo que cada componente corresponde a uma direção no espaço de atributos que maximiza a variância dos dados sob a restrição de ser ortogonal às anteriores. Assim, a PC1 representa a direção de maior variabilidade. A PC2 captura a maior variabilidade remanescente, independente de PC1 e assim sucessivamente. A variância explicada por cada componente é associada ao respectivo autovalor, o que se reflete diretamente na altura das barras do *scree plot*.

Observa-se que a primeira componente (PC1) concentra aproximadamente 48% da variância total dos atributos, e a segunda (PC2) adiciona cerca de 12%, to-

Figura 15 – Variância explicada e acumulada pelas primeiras componentes.



Fonte: Elaborado pelo autor, 2025.

talizando aproximadamente 60% nas duas primeiras componentes. A partir da quinta componente, a variância acumulada atinge em torno de 83%, e as componentes seguintes passam a contribuir com parcelas individuais relativamente pequenas, caracterizando ganhos marginais de explicação. Esse padrão sugere correlação entre as estatísticas consideradas, de modo que uma representação em baixa dimensionalidade preserva grande parte da variabilidade do conjunto de dados.

Do ponto de vista metodológico, esses resultados justificam o uso da PCA como apoio descritivo e para visualizações no plano PC1–PC2. Nas subseções seguintes, analisam-se as contribuições dos atributos nas componentes mais relevantes e as projeções das janelas nesse espaço reduzido, relacionando-as aos *clusters*, às ligas e ao resultado das partidas.

4.4.2 Importância dos atributos segundo o PCA

Além da análise da variância explicada, foi calculada uma medida de importância para cada atributo a partir dos módulos dos carregamentos (*loadings*) nas três primeiras componentes principais. Intuitivamente, quanto maior o carregamento de uma variável em PC1, PC2 e PC3, maior a sua contribuição para a estrutura de

variância capturada pela PCA.

A Tabela 12 apresenta os dez atributos com maior importância segundo esse critério, calculado sobre o vetor de estatísticas com médias e desvios-padrão nas janelas de 3, 4 e 5 partidas. Note-se que os valores de "Importância PCA" são bastante próximos entre si (diferenças da ordem de $4{,}7 \times 10^{-4}$, ou cerca de 2,5% em relação a um valor típico em torno de 0,018). Isso indica que o PCA não aponta para um atributo isolado claramente dominante, mas para um conjunto de variáveis com contribuições semelhantes nas primeiras componentes.

Tabela 12 – Atributos com maior contribuição para as três primeiras componentes.

Atributo	**Importância PCA**
mean_yellows_for_last4	0.018142
mean_shots_for_last4	0.017842
mean_yellows_for_last5	0.017815
mean_shots_for_last5	0.017808
mean_possession_for_last4	0.017786
mean_goals_for_last4	0.017766
mean_possession_for_last5	0.017727
std_yellows_for_last4	0.017692
std_corners_for_last4	0.017677
std_crosses_for_last4	0.017673

Fonte: Elaborado pelo autor, 2025.

Mesmo com essa proximidade numérica, é possível observar que os atributos que mais aparecem no topo pertencem a três grupos principais: (i) cartões amarelos (*mean_yellows_for_last4*, *mean_yellows_for_last5*, *std_yellows_for_last4*), (ii) volume ofensivo medido por finalizações (*mean_shots_for_last4*, *mean_shots_for_last5*) e gols (*mean_goals_for_last4*), e (iii) controle do jogo por meio da posse de bola (*mean_possession_for_last4*, *mean_possession_for_last5*) e de ações em bola parada (*std_corners_for_last4*, *std_crosses_for_last4*). Assim, mais do que destacar um único atributo, o PCA sugere que as componentes principais são influenciadas, em conjunto, por indicadores de disciplina (cartões), volume ofensivo e controle territorial da partida.

Esse padrão é coerente com trabalhos da literatura de predição de resultados de partidas, que normalmente combinam indicadores de desempenho ofensivo (chutes, gols), controle de posse e disciplina (faltas e cartões) como preditores. No contexto deste estudo, a PCA reforça que as informações agregadas nas janelas de 4 e 5 jogos sobre esses aspectos são as que mais explicam a variação global entre as janelas, o que justifica sua inclusão tanto nos experimentos de *clustering* quanto nas análises comparativas entre ligas e resultados.

Por outro lado, a distribuição relativamente próxima dos valores de importância sugere que não há um único atributo dominante; em vez disso, a estrutura dos dados é determinada por uma combinação de múltiplas estatísticas correlacionadas. Esse resultado é consistente com a dificuldade observada em separar vitórias, empates e derrotas por meio de agrupamentos não supervisionados, conforme discutido nas seções anteriores.

4.4.3 Visualização das componentes principais por clusters, ligas e resultados

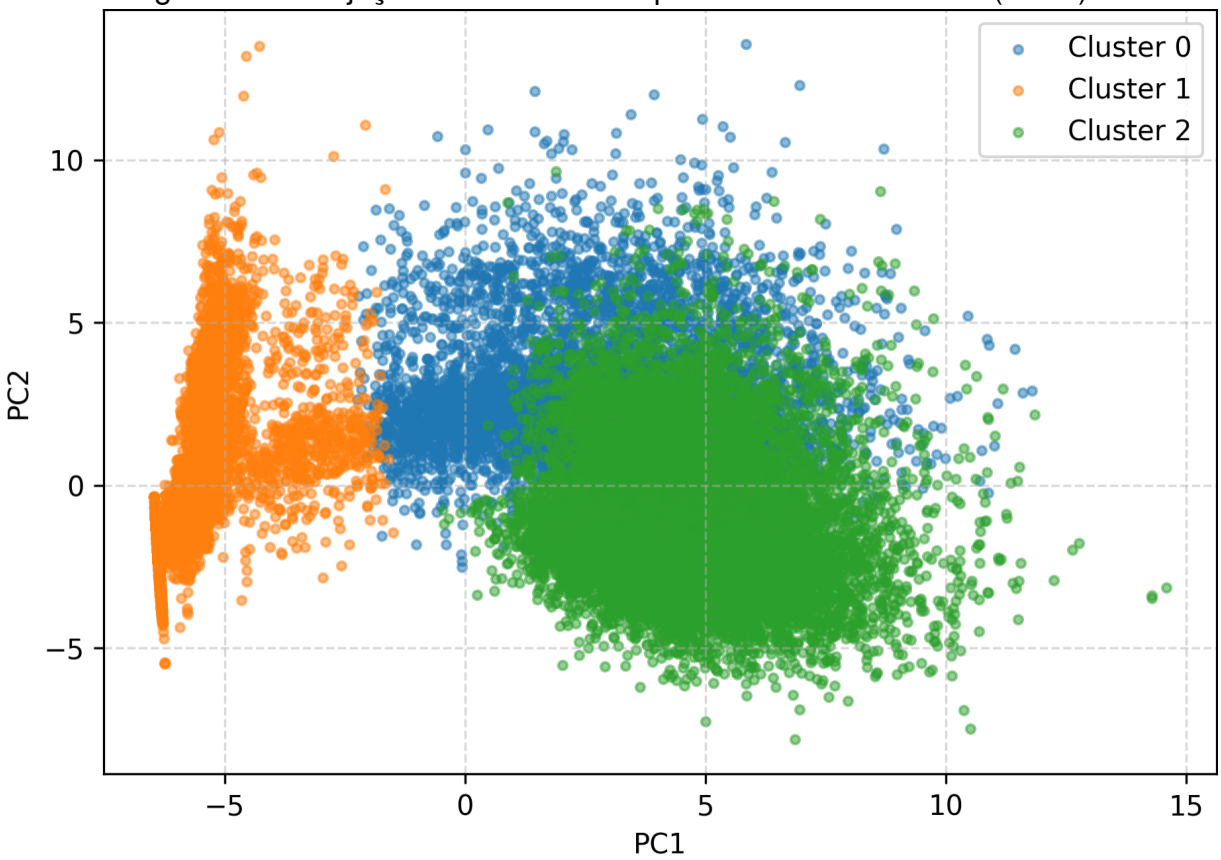
Com base na tabela *clustering_stats_3_4_5_pca_scores*, que armazena os *scores* das componentes principais juntamente com os rótulos de resultado (*result_current*), de liga (*league_id*) e de cluster (*cluster_kmeans_3_stats*), foi construída uma projeção bidimensional dos dados nas duas primeiras componentes principais. Como visto anteriormente, PC1 e PC2 respondem, em conjunto, por aproximadamente 60% da variância total das janelas de desempenho, de modo que essa visualização captura a maior parte da estrutura global dos dados.

A Figura 16 apresenta a projeção das janelas em PC1 e PC2, coloridas de acordo com os três clusters obtidos pelo K-Means ($k = 3$). Observa-se que o espaço bidimensional é ocupado por uma nuvem contínua de pontos, sem fronteiras nítidas entre os grupos, mas com regiões de maior concentração associadas a cada cluster. Um dos agrupamentos forma uma faixa mais compacta em valores negativos de PC1, enquanto os outros dois se distribuem em uma região mais espalhada, com PC1 predominante positivo. Essa organização é consistente com a interpretação da PCA apresentada anteriormente: PC1 resume principalmente o nível de intensidade ofensiva e de controle de jogo (finalizações, posse de bola, cartões), de modo que os clusters refletem variações nesses padrões médios de desempenho ao longo das janelas deslizantes.

Na Figura 17, a mesma projeção é colorida por liga. Embora os pontos de diferentes campeonatos se sobreponham amplamente, nota-se que certas regiões do plano PC1–PC2 são dominadas por ligas específicas. Essa observação está em consonância com as métricas de associação entre clusters e ligas discutidas anteriormente, em especial com o índice de Rand ajustado (ARI $\approx 0{,}16$) e a informação mútua normalizada (NMI $\approx 0{,}22$) obtidos na comparação *cluster vs. liga*. Em termos práticos, isso indica que os padrões médios de desempenho capturados por PC1 e PC2 tendem a agrupar janelas de partidas de uma mesma liga, sugerindo diferenças estruturais de estilo de jogo entre campeonatos (por exemplo, ligas com maior intensidade ofensiva e número de cartões versus ligas com jogos mais equilibrados e menos faltas).

Por fim, a Figura 18 mostra a distribuição das janelas no plano PC1–PC2 colorida de acordo com o resultado da partida (vitória, empate ou derrota). Diferen-

Figura 16 – Projeção em PC1 e PC2 por cluster de K-Means ($k = 3$).



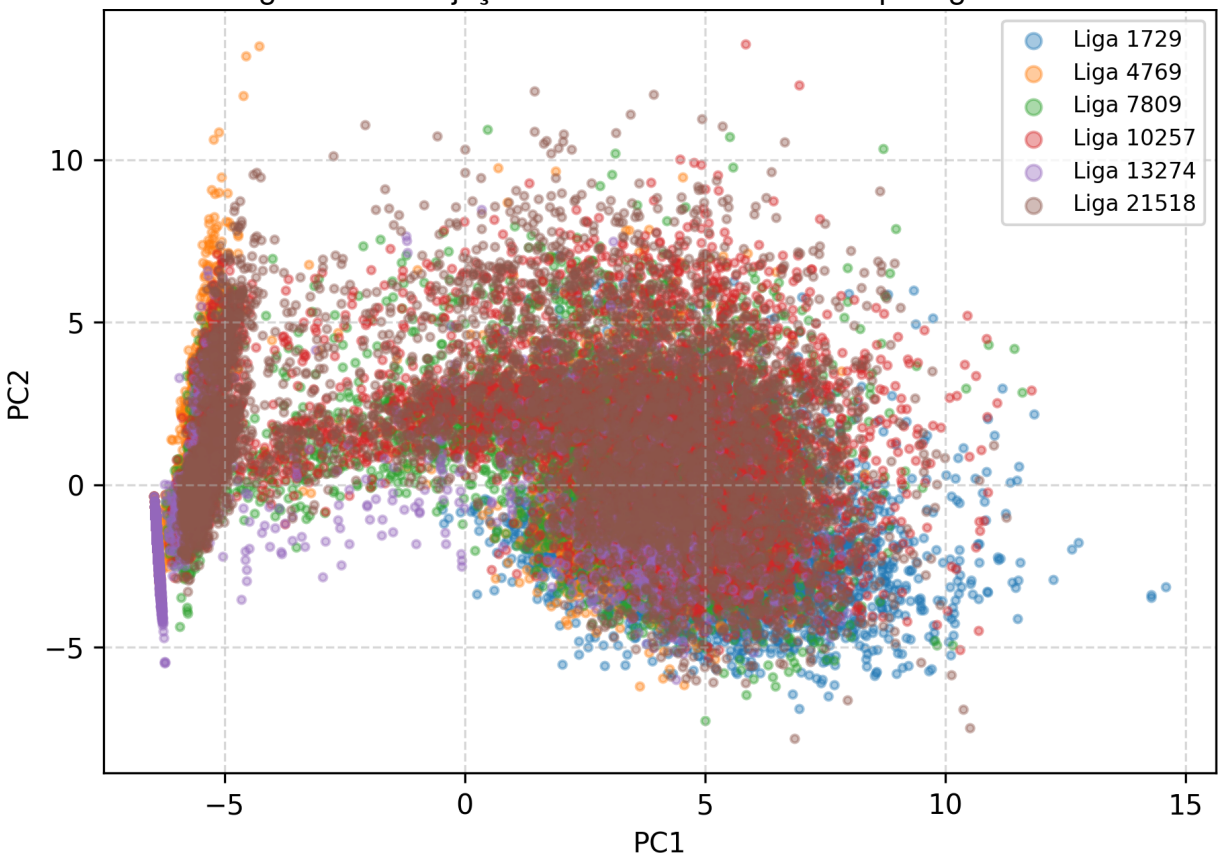
Fonte: Elaborado pelo autor, 2025.

temente do que se observa na visualização por liga, as três classes de resultado se encontram misturadas em praticamente todo o espaço, sem regiões dominadas por apenas um tipo de desfecho. Esse comportamento é compatível com os baixos valores de ARI e NMI obtidos para a comparação entre clusters e rótulos reais de resultado, bem como com as matrizes de confusão (*crosstabs*) apresentadas anteriormente, nas quais as proporções de vitórias, empates e derrotas em cada cluster permanecem próximas de um terço.

Em conjunto, essas visualizações reforçam a evidência empírica obtida nesta etapa: as componentes principais extraídas a partir das estatísticas agregadas descrevem bem variações de estilo e intensidade entre competências (e, consequentemente, entre clusters), mas não se alinham de forma clara com os rótulos de vitória, empate e derrota. Isso ajuda a explicar por que a matriz de confusão dos agrupamentos não se aproxima de uma diagonal “ideal”, mesmo após a incorporação de atributos inspirados na literatura de predição de resultados.

De forma geral, a análise com PCA reforça o quadro observado ao longo desta seção: apesar de as estatísticas agregadas de desempenho apresentarem correlação entre si, permitindo reduzir a dimensionalidade do problema a poucas componentes principais, os agrupamentos obtidos a partir desses vetores não se alinham de

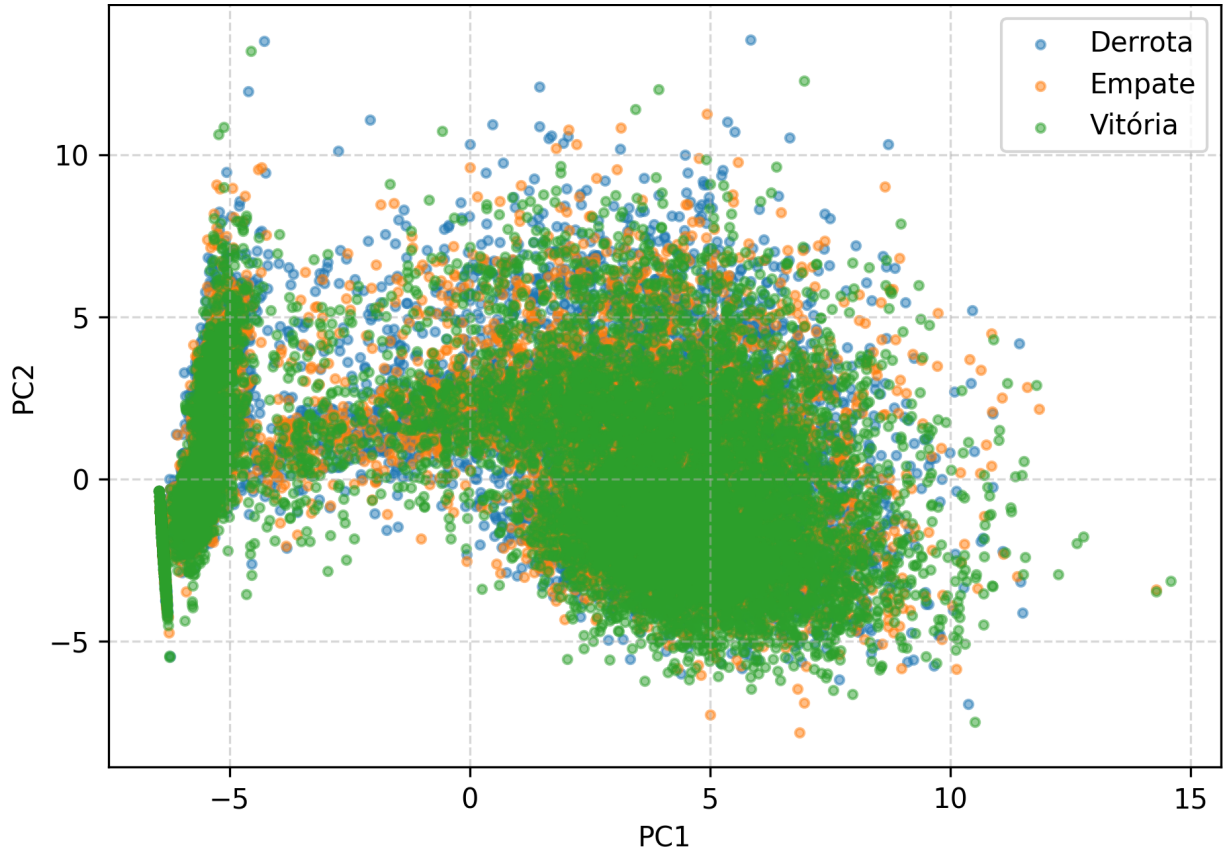
Figura 17 – Projeção em PC1 e PC2 colorida por liga.



Fonte: Elaborado pelo autor, 2025.

maneira clara com os rótulos de resultado das partidas. As primeiras componentes (especialmente PC1 e PC2) capturam sobretudo diferenças de intensidade ofensiva, posse de bola e disciplina (cartões e faltas), o que favorece a separação de estilos entre ligas e grupos de janelas com perfis semelhantes de desempenho, mas não produz fronteiras bem definidas entre vitórias, empates e derrotas. Essa evidência, combinada às matrizes de confusão e aos baixos valores de ARI e NMI em relação ao rótulo *result_current*, indica que os atributos considerados são mais adequados para caracterizar estilos de jogo do que para discriminar diretamente o desfecho das partidas.

Figura 18 – Projeção em PC1 e PC2 colorida por resultado.



Fonte: Elaborado pelo autor, 2025.

5 CONCLUSÃO

Este trabalho aplicou técnicas de Mineração de Dados para identificar padrões em partidas de futebol a partir de estatísticas de desempenho em janelas de jogos. Foram avaliadas diferentes configurações de vetores de atributos (histórico de cinco partidas, versões reduzidas e janelas deslizantes de 3, 4 e 5 jogos, com combinações com e sem variáveis disciplinares), sobre as quais se realizou agrupamento e análise descritiva dos grupos obtidos.

Os resultados mostraram que os vetores propostos permitem formar *clusters* com coerência interna, capturando regimes recorrentes de desempenho recente associados a perfis e intensidade de atuação (por exemplo, maior volume ofensivo, maior posse ou maior intensidade disciplinar). A Análise em Componentes Principais contribuiu para interpretar essa estrutura ao evidenciar que as direções de maior variabilidade do espaço de atributos estão ligadas a estatísticas de estilo de jogo, além de apoiar visualizações e comparações entre ligas e padrões internos a cada competição.

Como abordagem exploratória, investigou-se o alinhamento entre *clusters* e o rótulo *result_current* (vitória, empate ou derrota) por meio de tabelas de contingência e das métricas ARI e NMI. Observou-se associação muito fraca com o desfecho da partida, com proporções semelhantes de resultados entre grupos e valores de ARI/NMI próximos de zero, o que é compatível com a natureza contextual do futebol. Em síntese, as técnicas e os vetores propostos mostraram-se mais adequados para caracterizar padrões de desempenho e estilos de jogo do que para induzir, de forma não supervisionada, uma partição alinhada ao resultado das partidas.

5.1 Trabalhos futuros

Considerando extensões e aprimoramentos deste estudo, uma primeira continuidade natural consiste em empregar os vetores de atributos construídos como entrada para modelos supervisionados de predição de resultados. Enquanto aqui os agrupamentos foram explorados em um cenário não supervisionado, avaliando-se o alinhamento dos *clusters* ao rótulo *result_current*, trabalhos futuros podem treinar classificadores (por exemplo, regressão logística multinomial e redes neurais) para discriminar vitória, empate e derrota. Essa abordagem permitiria quantificar o potencial preditivo dos atributos propostos e comparar o desempenho de um cenário supervisionado com as limitações observadas no agrupamento.

Além disso, recomenda-se uma investigação sistemática da escolha do número de *clusters* no *K-Means*. Em particular, pode-se variar k em um intervalo x e avaliar simultaneamente a curva de inércia (soma dos quadrados intra-*cluster*, utilizada no método do cotovelo) e a silhueta média global. A análise conjunta dessas

curvas pode fornecer um outro critério para a escolha de um valor de k ao vetor avaliado.

Uma vez definido um k, torna-se relevante aprofundar a interpretação dos grupos formados a partir de variáveis de desempenho agregadas. Como os resultados indicaram baixa concordância direta com vitória/empate/derrota, uma alternativa é analisar se os *clusters* se relacionam com dinâmicas de desempenho ao longo do campeonato, tais como variações na posição da equipe na tabela ou no acúmulo de pontos em janelas de rodadas. Por exemplo, pode-se definir, para cada janela, a variação de posição (subida, queda ou estabilidade) e avaliar sua associação aos *clusters* por meio de tabelas de contingência e métricas externas (ARI/NMI) ou por modelos explicativos simples.

Outra linha de extensão é investigar a distribuição temporal das janelas de um mesmo time ao longo da temporada. Como uma equipe pode apresentar janelas atribuídas a diferentes *clusters*, é possível avaliar se determinados grupos ocorrem no início, no meio ou no fim do campeonato, bem como estudar transições entre *clusters* ao longo das rodadas. Essa análise exige engenharia de atributos relacionada ao índice temporal da partida (rodada), segmentação da temporada e, possivelmente, métricas de estabilidade de agrupamento, mas pode revelar mudanças estruturais de desempenho ao longo do tempo.

Por fim, recomenda-se ampliar a análise de sensibilidade quanto ao tamanho das janelas. Embora este trabalho tenha considerado janelas de 3, 4 e 5 partidas, estudos futuros podem avaliar comprimentos adicionais, comparando como a granularidade temporal afeta a separação geométrica (silhueta, Calinski–Harabasz e Davies–Bouldin) e a interpretação dos grupos. Em conjunto, essas extensões podem contribuir para uma compreensão mais completa do papel dos atributos e das janelas temporais na caracterização de perfis de desempenho em futebol.

REFERÊNCIAS

AGRESTI, A. **Categorical Data Analysis**. 3. ed. Hoboken: John Wiley & Sons, 2013. ISBN 9781118710944.

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. Disponível em: https://doi.org/10.1515/jqas-2022-0037.

AMERICAN SOCCER ANALYSIS. **2025 MLS Analytics Survey**. 2025. Disponível em: https://www.americansocceranalysis.com/home/2025-mls-analytics-survey. Acesso em: 25/10/2025.

ASHRAF, Q. **Data analytics in football**. 2021. Disponível em: https://mqamarulashraf.medium.com/data-analytics-in-football-part-1-project-formulation-data-acquisition-and-preparation-6219a1d24ab2. Acesso em: 24/11/2025.

BRADLEY, P. S. *et al.* Match performance and physical capacity of players in the top three competitive standards of English professional soccer. **Human Movement Science**, v. 32, n. 4, p. 808–821, 2013. Disponível em: https://doi.org/10.1016/j.humov.2013.06.002.

BUNKER, R.; THABTAH, F. A machine learning framework for sport result prediction. **Applied Computing and Informatics**, v. 15, n. 1, p. 27–33, 2019. DOI: 10.1016/j.aci.2017.09.005.

CALIŃSKI, T.; HARABASZ, J. A dendrite method for cluster analysis. **Communications in Statistics**, v. 3, n. 1, p. 1–27, 1974. Disponível em: https://doi.org/10.1080/03610927408827101.

CAO, C. **Sports data mining technology used in basketball outcome prediction**. 2012. Dissertação (Mestrado) – Dublin Institute of Technology, Dublin. Disponível em: https://arrow.tudublin.ie/scschcomdis/39/. Acesso em: 24/11/2025.

CARLING, C.; WILLIAMS, A. M.; REILLY, T. **Handbook of soccer match analysis:** a systematic approach to improving performance. London: Routledge, 2008.

CARPITA, M.; CIAVOLINO, E.; PASCA, P. Exploring and modelling team performances of the Kaggle European Soccer database. **Statistical Modelling**, v. 19, n. 1, p. 74–101, 2019. Disponível em: https://doi.org/10.1177/1471082X18810971.

COLLET, C. The possession game? A comparative analysis of ball retention and team success in European and International football, 2007–2010. **Journal of Sports Sciences**, v. 31, n. 2, p. 123–136, 2013. Disponível em: https://doi.org/10.1080/02640414.2012.727455.

CRUZ, M. A. **Análise e classificação de partidas de futebol com técnicas de mineração de dados**. 2023. Trabalho de Conclusão de Curso (Graduação em Engenharia da Computação) – Instituto Federal de Minas Gerais, Bambuí, 2023.

DAVIES, D. L.; BOULDIN, D. W. A Cluster Separation Measure. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, PAMI-1, n. 2, p. 224–227, 1979. Disponível em: https://doi.org/10.1109/TPAMI.1979.4766909.

DELLO IACONO, A. *et al.* Data analytics practices and reporting strategies in senior football: insights into athlete health and performance from over 200 practitioners worldwide. **Science and Medicine in Football**, 2025. Disponível em: https://doi.org/10.1080/24733938.2025.2476478.

DEMIR, E.; ŞAHIN, Y. H.; ÜRE, N. K. How Do Football Teams Play? A Deep Embedded Clustering Approach to Reveal Playing Styles. In: DONG, J.-s. *et al.* (Ed.). **Sports Analytics**. Cham: Springer, 2025. v. 15925. (Lecture Notes in Computer Science), p. 53–68. Disponível em: https://doi.org/10.1007/978-3-032-06167-6_4.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From data mining to knowledge discovery in databases. **AI Magazine**, v. 17, n. 3, p. 37–54, 1996.

FERNÁNDEZ, D.; CASALS, M. *et al.* Reporting of clustering techniques in sports sciences: a scoping review. **Electronic Journal of Applied Statistical Analysis**, v. 17, n. 3, p. 653–675, 2024. DOI: 10.1285/i20705948v17n3p653.

FERNÁNDEZ, J.; BORNN, L.; CERVONE, D. A framework for the fine-grained evaluation of the instantaneous expected value of soccer possessions. **Machine Learning**, v. 110, n. 2, p. 389–422, 2021. Disponível em: https://doi.org/10.1007/s10994-021-05989-6.

FERRAZ, A. *et al.* Tracking devices and physical performance analysis in team sports: a comprehensive framework for research—trends and future directions. **Frontiers in Sports and Active Living**, v. 5, p. 1284086, 2023. DOI: 10.3389/fspor.2023.1284086.

GERHARDT, T. E.; SILVEIRA, D. T. **Métodos de pesquisa**. Porto Alegre: Editora da UFRGS, 2009. Material didático da Universidade Aberta do Brasil – UAB/UFRGS. Disponível em: https://www.lume.ufrgs.br/handle/10183/52806. Acesso em: 23/12/2025.

GŁOWANIA, S.; KOZAK, J.; JUSZCZUK, P. Knowledge discovery in databases for a football match result. **Electronics**, v. 12, n. 12, p. 2712, 2023. Disponível em: https://doi.org/10.3390/electronics12122712.

GOMES, N. C. d. M. **Avaliação de modelos preditivos em estatísticas esportivas ao vivo: uma abordagem de dados em tempo real**. 2024. Trabalho de Conclusão de Curso (Graduação em Ciência da Computação) – Universidade Federal da Paraíba, João Pessoa.

GUDMUNDSSON, J.; HORTON, M. Spatio-temporal analysis of team sports. **ACM Computing Surveys**, v. 50, n. 2, p. 1–34, 2017. Disponível em: https://doi.org/10.1145/3054132.

GYARMATI, L.; STANOJEVIĆ, R. QPass: a merit-based evaluation of soccer passes. **arXiv preprint arXiv:1608.03532**, 2016. Publicado em 8 de agosto de 2016. Disponível em: https://doi.org/10.48550/arXiv.1608.03532.

HAN, J.; PEI, J.; TONG, H. **Data mining: concepts and techniques**. 4. ed. Cambridge: Morgan Kaufmann, 2022.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. **The elements of statistical learning: data mining, inference, and prediction**. 2. ed. New York: Springer, 2009. Disponível em: https://web.stanford.edu/~hastie/ElemStatLearn/. Acesso em: 23/12/2025.

HOCHKAMP, F.; SCHEIDLER, A. A.; RABE, M. Review of Maturity Models for Data Mining and Proposal of a Data Preparation Maturity Model Prototype for Data Mining. **Computers**, v. 14, n. 4, p. 146, 2025. Disponível em: https://doi.org/10.3390/computers14040146.

HUBERT, L.; ARABIE, P. Comparing partitions. **Journal of Classification**, v. 2, n. 1, p. 193–218, 1985. Disponível em: https://doi.org/10.1007/BF01908075.

HUGHES, M.; BARTLETT, R. The use of performance indicators in performance analysis. **Journal of Sports Sciences**, v. 20, n. 10, p. 739–754, 2002. DOI: 10.1080/026404102320675602.

JAMES, G. *et al.* **An introduction to statistical learning**. 2. ed. New York: Springer, 2021. Disponível em: https://www.statlearning.com/. Acesso em: 23/12/2025.

JOLLIFFE, I. T. **Principal Component Analysis**. 2. ed. New York: Springer, 2002.

KALT, C. T. **Cluster analysis on football teams performance data**. 2024. Trabalho de Conclusão de Curso (Bachelor's thesis) – Uppsala University, Uppsala. Disponível em: https://uu.diva-portal.org/smash/get/diva2:1833447/FULLTEXT01.pdf. Acesso em: 23/12/2025.

KOHONEN, T. **Self-organizing maps**. 3. ed. Berlin: Springer, 2001. v. 30. (Springer Series in Information Sciences). Disponível em: https://doi.org/10.1007/978-3-642-56927-2.

LAGO-PEÑAS, C.; DELLAL, A. Game-related statistics that discriminated winning, drawing and losing teams from the Spanish soccer league. **Journal of Sports Science & Medicine**, v. 9, n. 2, p. 288–293, 2010.

LAROSE, D. T. **Discovering knowledge in data:** an introduction to data mining. 2. ed. Hoboken, NJ: John Wiley & Sons, 2014. ISBN 0470908742.

LAVILLE, C.; DIONNE, J. **A construção do saber:** manual de metodologia da pesquisa em ciências humanas. Porto Alegre; Belo Horizonte: Artmed; Editora UFMG, 1999. Tradução de Heloísa Monteiro e Francisco Settineri.

LESKOVEC, J.; RAJARAMAN, A.; ULLMAN, J. D. **Mining of massive datasets**. 3. ed. Cambridge: Cambridge University Press, 2020. Disponível em: http://www.mmds.org.

LINHARES, L. S. **Aplicação de técnicas de mineração de dados para análise de partidas de futebol**. 2024. Trabalho de Conclusão de Curso (Graduação em Engenharia de Computação) – Universidade Tecnológica Federal do Paraná, Pato Branco.

LIU, H.; HOPKINS, W. G.; GÓMEZ, M.-A. Modelling relationships between match events and match outcome in elite football. **European Journal of Sport Science**, v. 16, n. 5, p. 516–525, 2016. DOI: 10.1080/17461391.2015.1042527.

MAIMONE, V. M.; YASSERI, T. Football is becoming more predictable: network analysis of 88 thousand matches in 11 major leagues. **Royal Society Open Science**, v. 8, n. 12, p. 210617, 2021. Disponível em: https://doi.org/10.1098/rsos.210617.

MITCHELL, T. M. **Machine learning**. New York: McGraw-Hill, 1997. ISBN 9780070428072.

MUÑOZ, O. *et al.* Automated discovery of successful strategies in association football. **Applied Sciences**, v. 14, n. 4, p. 1403, 2024. Disponível em: https://doi.org/10.3390/app14041403.

NARAYANAN, S. *et al.* Object detection and tracking for football data analytics. In: PROCEEDINGS of the 1st International Conference on Artificial Intelligence, Communication, IoT, Data Engineering and Security (IACIDS 2023). Lavasa, Pune, India, 2023. Disponível em: https://doi.org/10.4108/eai.23-11-2023.2343216.

PAPPALARDO, L. *et al.* PlayeRank: data-driven performance evaluation and player ranking in soccer via machine learning. **ACM Transactions on Intelligent Systems and Technology (TIST)**, v. 10, n. 5, p. 59:1–59:27, 2019. Disponível em: https://doi.org/10.1145/3343172.

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

PERIN, C. *et al.* State of the art of sports data visualization. **Computer Graphics Forum**, v. 37, n. 3, p. 663–686, 2018. Disponível em: https://doi.org/10.1111/cgf.13447.

PIETRASZEWSKI, P. *et al.* The Role of Artificial Intelligence in Sports Analytics: A Systematic Review and Meta-Analysis of Performance Trends. **Applied Sciences**, v. 15, n. 13, p. 7254, 2025. DOI: 10.3390/app15137254.

PLAKIAS, S. *et al.* A Multivariate and cluster analysis of diverse playing styles across European Football Leagues. **Journal of Physical Education and Sport**, v. 23, n. 7, p. 1631–1641, 2023. Disponível em: https://doi.org/10.7752/jpes.2023.07200.

PROVOST, F.; FAWCETT, T. **Data science for business:** what you need to know about data mining and data-analytic thinking. Sebastopol, CA: O'Reilly Media, 2013.

RAI, P.; SINGH, S. A survey of clustering techniques. **International Journal of Computer Applications**, v. 7, n. 12, p. 1–5, 2010. Disponível em: https://doi.org/10.5120/1326-1808.

RANA, A. S. S. Event detection in football using graph convolutional networks. **arXiv preprint arXiv:2301.10052**, 2023. Disponível em: https://doi.org/10.48550/arXiv.2301.10052.

REIN, R.; MEMMERT, D. Big data and tactical analysis in elite soccer: future challenges and opportunities for sports science. **SpringerPlus**, v. 5, n. 1, p. 1410, 2016. Disponível em: https://doi.org/10.1186/s40064-016-3108-2.

REUTERS. **With holograms, AI and big data, football front and centre in tech race**. 2024. Disponível em: https://www.reuters.com/sports/soccer/with-holograms-ai-big-data-football-front-centre-tech-race-2024-07-05/. Acesso em: 24/11/2025.

RIBEIRO, M. R. **Big data avançado e mineração de dados**. 1. ed. Belo Horizonte: Instituto Federal de Minas Gerais, 2022. Material didático em acesso aberto do IFMG. ISBN 978-65-5876-025-2. Disponível em: http://hdl.handle.net/20.500.14387/1811. Acesso em: 24/11/2025.

ROUSSEEUW, P. J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. **Journal of Computational and Applied Mathematics**, v. 20, p. 53–65, 1987. Disponível em: https://doi.org/10.1016/0377-0427(87)90125-7.

SAPP, R. M.; SPANGENBURG, E. E.; IRVING, B. A. Trends in teams' technical performance in the English Premier League: 2008–2016. **International Journal of Performance Analysis in Sport**, v. 18, n. 6, p. 945–956, 2018. Disponível em: https://doi.org/10.1080/24748668.2018.1539385.

SCHWABER, K.; SUTHERLAND, J. **O guia do Scrum:** o guia definitivo para o Scrum: as regras do jogo. 2020. Disponível em: https://scrumguides.org/scrum-guide-2020.html. Acesso em: 09/12/2025.

SINGH, H. V.; GIRDHAR, A.; DAHIYA, S. A literature survey based on DBSCAN algorithms. In: p. 751–758. Disponível em: https://doi.org/10.1109/ICICCS53718.2022.9788440.

SMITH, S. **Premier League tables with Football-Data.co.uk**. 2020. Disponível em: https://rstudio-pubs-static.s3.amazonaws.com/597789_e8d4a716dbf34b15850becf0d0ded72d.html. Acesso em: 24/11/2025.

STREHL, A.; GHOSH, J. Cluster ensembles - a knowledge reuse framework for combining multiple partitions. In: p. 93–99. Disponível em: https://www.jmlr.org/papers/v3/strehl02a.html. Acesso em: 13/12/2025.

TAN, P.-N. *et al.* **Introduction to data mining**. 2. ed. Boston, MA: Pearson, 2018.

VIDAL-CODINA, F. *et al.* Automatic event detection in football using tracking data. **arXiv preprint arXiv:2202.00804**, 2022. Disponível em: https://doi.org/10.48550/arXiv.2202.00804.

VINH, N. X.; EPPS, J.; BAILEY, J. Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. **Journal of Machine Learning Research**, v. 11, p. 2837–2854, 2010. Disponível em: https://jmlr.org/papers/v11/vinh10a.html. Acesso em: 13/01/2026.

WIKIPEDIA. **DBSCAN**. Disponível em: https://en.wikipedia.org/wiki/DBSCAN. Acesso em: 22/09/2025.

XIE, J.; GIRSHICK, R.; FARHADI, A. Unsupervised deep embedding for clustering analysis. In: v. 48, p. 478–487. Disponível em: http://proceedings.mlr.press/v48/xieb16.html. Acesso em: 23/12/2025.

XU, R.; WUNSCH, D. C. Survey of clustering algorithms. **IEEE Transactions on Neural Networks**, v. 16, n. 3, p. 645–678, 2005. Disponível em: https://doi.org/10.1109/TNN.2005.845141.

YEUNG, C.; BUNKER, R.; FUJII, K. Unveiling multi-agent strategies: a data-driven approach for extracting and evaluating team tactics from football event and freeze-frame data. **Journal of Robotics and Mechatronics**, v. 36, n. 3, p. 603–617, 2024. DOI: https://doi.org/10.20965/jrm.2024.p0603.

YI, Q. *et al.* Technical and physical match performance of teams in the FIFA World Cup 2018: Effects of match outcome and team quality. **International Journal of Performance Analysis in Sport**, v. 19, n. 6, p. 1043–1056, 2019. Disponível em: https://doi.org/10.1080/24748668.2019.1689753.

ZHANG, T.; RAMAKRISHNAN, R.; LIVNY, M. BIRCH: an efficient data clustering method for very large databases. In: p. 103–114. Disponível em: https://doi.org/10.1145/235968.233324.

APÊNDICES

Tabela 13 – Atributos considerados em cada partida no vetor team_windows_5_strict.

Atributo	**Descrição**
goals_for	Gols marcados pela equipe na partida
goals_against	Gols sofridos pela equipe na partida
shots_for	Finalizações realizadas pela equipe
shots_against	Finalizações sofridas pela equipe
shots_on_target_for	Finalizações no alvo a favor da equipe
shots_on_target_against	Finalizações no alvo contra a equipe
corners_for	Escanteios a favor da equipe
corners_against	Escanteios contra a equipe
crosses_for	Cruzamentos a favor da equipe
crosses_against	Cruzamentos contra a equipe
fouls_for	Faltas cometidas pela equipe
fouls_against	Faltas sofridas pela equipe
offsides_for	Impedimentos da equipe
offsides_against	Impedimentos dos adversários
yellows_for	Cartões amarelos recebidos pela equipe
yellows_against	Cartões amarelos recebidos pelos adversários
reds_for	Cartões vermelhos recebidos pela equipe
reds_against	Cartões vermelhos recebidos pelos adversários
possession_for	Percentual de posse de bola da equipe

Fonte: Elaborado pelo autor, 2026.