CEFET/RJ
DIRETORIA DE PESQUISA E PÓS-GRADUAÇÃO

APLICAÇÃO DE MÉTODOS BASEADOS EM CONCEPT DRIFT PARA PREVISÃO DE GOLS NO FUTEBOL PROFISSIONAL

Ana Gabriela Viana de Araújo

Dissertação de Mestrado apresentada ao Programa de Pós-graduação em Ciência da Computação, Centro Federal de Educação Tecnológica Celso Suckow da Fonseca CEFET/RJ, como parte dos requisitos necessários à obtenção do grau de mestre.

Orientador:
Jorge de Abreu Soares

Rio de Janeiro,

13 de maio de 2026

Aplicação de métodos baseados em concept drift para previsão de gols no futebol profissional

Dissertação de Mestrado em Ciência da Computação, Centro Federal de Educação Tecnológica Celso Suckow da Fonseca, CEFET/RJ.

Ana Gabriela Viana de Araújo

Aprovada por:

Jorge de Abreu Soares, D.Sc. (Cefet/RJ)

Glauco Fiorott Amorim, D.Sc. (Cefet/RJ)

Pedro Henrique González Silva , D.Sc. (UFRJ)

Carlos Eduardo Ribeiro de Mello, D.Sc. (Unirio)

Rio de Janeiro,

13 de maio de 2026

Ficha catalográfica elaborada pela Biblioteca Central do CEFET/RJ

A663 Araújo, Ana Gabriela Viana de
Aplicação de métodos baseados em concept drift para previsão de gols no futebol profissional / Ana Gabriela Viana de Araújo. — 2026.
70f. : il. color. , enc.

Dissertação (Mestrado) Centro Federal de Educação Tecnológica Celso Suckow da Fonseca, 2026.
Bibliografia : f. 66-70
Orientador: Jorge de Abreu Soares

1. Aprendizado de máquina. 2. Teoria da previsão. 3. Futebol - Processamento de dados. 4. Análise de séries temporais. 5. Inteligência artificial. I. Soares, Jorge de Abreu. (Orient.). II. Título.

CDD 006.31

Elaborada pela bibliotecária Tania Mello – CRB/7 nº 5507/04

DEDICATÓRIA

Dedico esta dissertação a todos que acreditaram em mim quando eu mesma duvidei e a todos que são beneficiados pelo ensino público de qualidade, que um dia eu possa retribuir.

AGRADECIMENTOS

Ao meu orientador que compartilhou essa jornada comigo, sempre motivando e acreditando no meu potencial, serei eternamente grata. Ao Lucas Giusti por todo o apoio na pesquisa e ao Matheus Melo pela parceria no tema.

A Deus, pelas oportunidades e caminhos que me trouxeram até aqui. À minha família, sendo exemplo, amparo e motivação. Ao meu companheiro Salomão Alencar pela compreensão e apoio incondicional.

Aos amigos Fernando Fraga, Thays Leal e Vanessa Soares, que atravessaram essa jornada do mestrado comigo. Às amigas mestras que compreenderam como ninguém minhas aflições, Ana Carolina Alves e Sara Vieira.

A todos que de alguma forma contribuíram para a realização desta dissertação. Essa conquista não seria possível sem cada uma das pessoas mencionadas e tantas outras que me apoiaram ao longo dessa caminhada.

RESUMO

Aplicação de métodos baseados em concept drift para previsão de gols no futebol profissional

Ana Gabriela Viana de Araújo

Orientador:
Jorge de Abreu Soares

Resumo da Dissertação submetida ao Programa de Pós-graduação em Ciência da Computação do Centro Federal de Educação Tecnológica Celso Suckow da Fonseca CEFET/RJ como parte dos requisitos necessários à obtenção do grau de mestre.

Este trabalho investiga a aplicação de técnicas de detecção de *concept drift* para a identificação antecipada de gols em partidas de futebol, com base em dados de eventos intra-partida. A abordagem trata o problema como monitoramento de mudanças na distribuição de passes intra-partida, utilizando *drift* virtual operacionalmente, isto é, detecção baseada exclusivamente em $P(X)$ sem rótulos em tempo real, com a premissa de que essas mudanças precedem alterações na probabilidade de gol. A robustez dos resultados é verificada por divisão temporal com 190 partidas de treino e 190 de teste. Foram utilizados dados da temporada 2015/2016 da La Liga: 380 partidas, agregadas em intervalos de um minuto, com análise tanto do comportamento ofensivo quanto defensivo. Três detectores de *drift* foram avaliados (Page-Hinkley, KSWIN e ADWIN) em comparação com baselines determinístico e estocástico, utilizando como sinal de entrada médias móveis da frequência de passes. A avaliação adota uma variante assimétrica do *SoftED evaluation*, que penaliza alarmes tardios por meio de uma função de pontuação linear decrescente na janela $[t - K, t]$, com $K = 10$ minutos. Os resultados indicam que o Page-Hinkley obteve o maior MCC entre os detectores avaliados, superando ambos os *baselines*; Page-Hinkley e KSWIN apresentaram F1 equivalentes, com vantagem marginal do KSWIN. A comparação com abordagem supervisionada da literatura evidencia que o método proposto, embora mais simples e sem necessidade de dados rotulados, atinge desempenho competitivo a partir da primeira partida. Discutem-se limitações da abordagem, incluindo o uso de passes como único sinal *proxy* e a restrição a uma única temporada, além de perspectivas para trabalhos futuros com variáveis multivariadas e análise longitudinal.

Palavras-chave:
Concept Drift, Predição de Gol, Futebol profissional, Análise de dados no esporte, Séries Temporais

Rio de Janeiro,

13 de maio de 2026

ABSTRACT

Application of concept drift-based methods for predicting goals in professional football

Ana Gabriela Viana de Araújo

Advisor:
Jorge de Abreu Soares

Abstract of dissertation submitted to Programa de Pós-graduação em Ciência da Computação - Centro Federal de Educação Tecnológica Celso Suckow da Fonseca CEFET/RJ as partial fulfillment of the requirements for the degree of master.

This work investigates the application of concept drift detection techniques for the early identification of goals in football matches, using in-match event data. The problem is framed as monitoring distributional changes in pass frequency time series, employing virtual drift operationally, i.e., detection based solely on $P(X)$ without real-time labels, under the premise that such changes precede shifts in goal probability. Robustness is assessed through a temporal split with 190 training and 190 test matches. Data from the 2015/2016 La Liga season: 380 matches aggregated at one-minute intervals were used, with analysis covering both offensive and defensive behavior. Three drift detectors were evaluated (Page-Hinkley, KSWIN, and ADWIN) against deterministic and stochastic baselines, using moving averages of pass frequency as input signal. Evaluation follows an asymmetric variant of SoftED evaluation, which penalizes late alarms through a linearly decreasing scoring function over the window $[t-K, t]$, with $K = 10$ minutes. Results show that Page-Hinkley achieved the highest MCC among all evaluated detectors, outperforming both baselines; Page-Hinkley and KSWIN yielded equivalent F1, with a marginal advantage for KSWIN. Comparison with a supervised approach from the literature demonstrates that the proposed method, despite being simpler and requiring no labeled data, achieves competitive performance from the very first match. Limitations are discussed, including the use of passes as the sole proxy signal and the restriction to a single season, alongside directions for future work involving multivariate features and longitudinal analysis.

Key-words:
Concept Drift, Goal Prediction, Professional Football, Sport analytics, Time Series

Rio de Janeiro,

May 13th, 2026

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Abreviações

ADWIN	*Adaptive Windowing*	7
CWA	*Competing Windows Algorithm*	17
F1	*F1 Score*	7
FN	*False Negative*	10
FP	*False Positive*	10
GPS	*Global Positioning System*	13
GVDEP	*Generalized Valuing Defense by Estimating Probabilities*	19
IC	*Intervalo de Confiança*	27
IGSOP	*In-Game Outcome Prediction*	17
KS	*Kolmogorov-Smirnov*	9
KSWIN	*Kolmogorov-Smirnov Windowing*	7
LEM	*Large Events Model*	18
MCC	*Matthew's Correlation Coefficient*	11
MLP	*Multi-Layer Perceptron*	18
OPENSTARLAB	*Open Spatio-Temporal Agent Research Lab*	18
PG	*Prediction Goal*	20
PI	*Performance Indicator*	18
PRISMA	*Preferred Reporting Items for Systematic Reviews*	15
SCORE	*Scoring COnvolution for next-Event REcognition*	19
SOFTED	*Soft Evaluation for Event Detection*	3
SRE	*Success/Score-Related Event*	17
TN	*True Negative*	11
TP	*True Positive*	10
VAEP	*Valuing Actions by Estimating Probabilities*	16
XG	*Expected Goals*	14
XGBOOST	*EXtreme Gradient Boosting*	18

Capítulo I Introdução

Futebol é o esporte mais popular do mundo, caracterizado por combinar habilidade e fatores aleatórios, o que o torna altamente imprevisível [Alves, 2025]. Essa complexidade reflete em placares geralmente baixos [Berrar et al., 2019], no qual o gol é a principal medida de efetividade ofensiva [Sarmento et al., 2018]. Além de ser relativamente raro, o gol constitui um evento crítico capaz de alterar os padrões de comportamento das equipes ao longo da partida. Estudos indicam que o primeiro gol, em particular, pode ser decisivo para o resultado final, ao oferecer uma vantagem tanto estatística quanto psicológica para a equipe que abre o placar [Liu et al., 2021; Anwar et al., 2022; Dutta et al., 2024].

De acordo com Wiechno et al. [2025], os primeiros estudos voltados à previsão de resultados em partidas de futebol remontam aos anos 1960. Por várias décadas, a modelagem baseada em placares foi dominada por métodos estatísticos tradicionais [Hubáček et al., 2022], até que, em 1996, algoritmos de *machine learning* começaram a ser incorporados, inicialmente em estudos sobre futebol americano [Purucker, 1996], e, na década seguinte, em pesquisas sobre futebol [Reed and O'Donoghue, 2005; Bunker and Susnjak, 2022]. Apesar dos avanços, os estudos sobre predição de resultados e eventos ainda se encontram em estágios iniciais, dadas as complexidades e nuances do esporte [Alves, 2025].

Nesse contexto, uma partida de futebol pode ser analisada a partir de variáveis contextuais [Lago-Peñas et al., 2011], ofensivas e defensivas [Delgado-Bordonau et al., 2013], além de aspectos físicos e psicológicos que influenciam diretamente o desempenho e os resultados [Liu et al., 2021]. As variáveis contextuais situam o jogo em seu enquadramento competitivo, abrangendo fatores como atuar em casa ou fora, a posição da equipe na tabela e o momento da temporada [Sarmento et al., 2018]. Diversos estudos têm se dedicado a investigar o chamado *home advantage* [Almeida et al., 2014; Chacón-Fernández et al., 2025; Lago-Peñas et al., 2011], a vantagem de atuar em casa, e o quanto esse fator pode influenciar o comportamento dos jogadores e, consequentemente, o desfecho das partidas.

As variáveis ofensivas, por sua vez, como o número de ataques, finalizações e chutes a gol, estão associadas à criação de oportunidades e à eficácia na conversão em gols [Anwar et al., 2022; Delgado-Bordonau et al., 2013]. Em contrapartida, as variáveis defensivas, como desarmes, interceptações e faltas, refletem os esforços empregados pelas equipes para neutralizar o adversário e minimizar riscos

durante a partida [Forcher et al., 2024]. Diversas pesquisas, inclusive, têm explorado comparativamente as relações entre métricas ofensivas e defensivas, buscando compreender como o equilíbrio entre esses dois domínios influencia o desempenho e os resultados das equipes [Anwar et al., 2022; Delgado-Bordonau et al., 2013].

A literatura sobre previsão de gols tem se concentrado principalmente em indicadores ofensivos, como finalizações e posse de bola [de Souza et al., 2019; Anwar et al., 2022]. Entretanto, alguns estudos apontam que, após marcar um gol, as equipes tendem a modificar seu estilo e estratégia de jogo [Liu et al., 2021]. Essas alterações desafiam a suposição de estabilidade nos dados ao longo do tempo e evidenciam a possibilidade de ocorrência de *drift* virtual nas métricas de jogo, isto é, mudanças na distribuição das variáveis de entrada ao longo da partida [Gama et al., 2014]. Neste trabalho, esse *drift* virtual é utilizado como *proxy* para *concept drift*: a premissa é que alterações detectáveis no padrão de passes, tanto da equipe atacante quanto da defensora, precedem mudanças na dinâmica do jogo que aumentam a probabilidade de ocorrência de gol.

Do ponto de vista aplicado, a antecipação de gols tem valor direto para a tomada de decisão tática durante a partida. Treinadores dispõem de um conjunto limitado de intervenções em tempo real, cuja efetividade depende de serem executadas com antecedência suficiente para alterar o curso do jogo antes que o equilíbrio se rompa. Uma substituição, por exemplo, exige que o treinador identifique a necessidade, comunique a instrução e aguarde a próxima interrupção do jogo para efetivá-la, o que torna alarmes com menos de cinco minutos de antecedência frequentemente tardios demais para qualquer ação relevante em campo. Esse requisito operacional motiva a adoção de uma janela de tolerância de $K = 10$ minutos na avaliação proposta neste trabalho.

Esta dissertação propõe investigar a hipótese de que mudanças na distribuição do comportamento de jogo podem funcionar como alarme preditor para a ocorrência de gols. A análise opera na escala da partida, avaliando cada tempo (primeiro ou segundo) de cada jogo individualmente por janelas temporais de 1 minuto. A pergunta de pesquisa que orienta este trabalho é: **em que medida detectores de *concept drift* não supervisionados, aplicados a séries temporais de passes intra-partida, são capazes de antecipar a ocorrência de gols no futebol profissional?** A hipótese investigada é que mudanças estatisticamente detectáveis no padrão de passes, tanto da equipe atacante quanto da defensora, precedem eventos de gol, configurando um problema de *drift* virtual que pode ser monitorado sem necessidade de dados rotulados. Para responder a essa questão, é proposto um arcabouço metodológico que detecta variações nos padrões ofensivos e defensivos em dados intra-partida da La Liga, de forma não supervisionada, avaliado tanto em protocolo *in-sample* quanto em divisão temporal com 190 partidas de treino e 190 de teste.

Esta dissertação traz como contribuições: (i) a formulação do problema de detecção antecipada de gols como detecção de drift virtual em séries de passes intra-partida; (ii) a adaptação da métrica

Soft Evaluation for Event Detection (SoftED) para avaliação assimétrica de alarmes antecipados, valorizando detecções temporalmente próximas ao evento; e (iii) a avaliação empírica de três detectores não supervisionados sobre dados reais de futebol profissional, com validação *out-of-sample* por divisão temporal.

Além da Introdução, o texto se organiza da seguinte forma: O Capítulo II apresenta o referencial teórico, abordando análise de desempenho no futebol, os fundamentos de *concept drift* em fluxos de dados, seu uso para detecção de eventos, os algoritmos de detecção utilizados na pesquisa e a metodologia de avaliação *SoftED evaluation*. O Capítulo III discute os trabalhos relacionados, situando esta pesquisa no contexto da literatura existente e estabelecendo um paralelo com abordagens supervisionadas de previsão intra-partida. O Capítulo IV detalha a metodologia proposta, incluindo o tratamento de dados, a análise exploratória, os algoritmos utilizados e o método de avaliação dos resultados. Por sua vez, os resultados dos experimentos e a discussão sobre comparação entre detectores, análise por perspectiva ofensiva e defensiva, perspectiva de time da casa e time visitante, além do contraste com a abordagem supervisionada da literatura são explicitados no Capítulo V. Por fim, o Capítulo VI traz as considerações finais, destacando as limitações da pesquisa e possíveis direções para evolução do tema.

Capítulo II Referencial Teórico

Este capítulo está organizado em três seções. A Seção II.1 apresenta os fundamentos de *concept drift* em fluxos de dados, abordando sua definição formal, tipos, uso para detecção de eventos e os algoritmos de detecção utilizados nesta pesquisa. A Seção II.2 apresenta a metodologia de avaliação adotada, descrevendo as métricas clássicas de detecção e a abordagem *SoftED evaluation* utilizada para mensurar a qualidade dos alarmes em relação à janela temporal de ocorrência dos gols. A Seção II.3 detalha as principais variáveis de jogo analisadas, explicando sua relevância para o desempenho das equipes e sua relação com a ocorrência de gols.

II.1 Concept Drift

O desenvolvimento de modelos preditivos em ambientes reais exige considerar a possibilidade de mudanças na distribuição dos dados ao longo do tempo [Webb et al., 2016]. Em partidas de futebol, essas mudanças são frequentes e modelos treinados sob a suposição de estacionariedade tendem a perder desempenho [Iwashita and Papa, 2019] à medida que o jogo evolui. Mesmo quando o fluxo da partida parece inalterado, as distribuições de variáveis-chave e as relações entre essas variáveis e a probabilidade de ocorrência de um gol podem variar ao longo do jogo. Essas variações dão origem ao fenômeno conhecido como *concept drift* [Gama et al., 2014]. Em termos gerais, o *concept drift* corresponde a uma diferença estatisticamente significativa entre a distribuição conjunta das variáveis de entrada e de saída observada em diferentes amostras da partida.

Em diversos problemas de aprendizado de máquina, os dados não são disponibilizados de forma estática, mas chegam continuamente ao longo do tempo, caracterizando um fluxo ou *data stream* [Hoens et al., 2012]. Nesses cenários, o modelo é treinado apenas com observações passadas, mas precisa atuar sobre dados futuros que podem ser gerados por um processo diferente daquele observado inicialmente [Webb et al., 2016]. Essa característica torna os fluxos de dados desafiadores, pois ambientes reais costumam apresentar mudanças estruturais, como surgimento de novos padrões e alteração nas relações entre variáveis [Iwashita and Papa, 2019]. Modelos tradicionais, treinados em regime *batch* e sob a suposição de estacionariedade, tendem a perder desempenho quando essa distribuição evolui ao longo do tempo [Webb et al., 2016].

Técnicas de *concept drift* buscam aprender padrões em *data streams* que podem vir a mudar

ao longo do tempo [Iwashita and Papa, 2019]. Formalmente, o *concept drift* entre dois pontos no tempo (t e $t+1$) pode ser definido como uma diferença na distribuição conjunta entre o conjunto de variáveis de entrada (X) e a variável alvo (y) nesse período. Ou seja, *p(X, y)* no tempo t é diferente de *p(X, y)* no tempo $t+1$ [Gama et al., 2014].

$$\exists X : p_{t_0}(X, y) \neq p_{t_1}(X, y) \tag{II.1}$$

Este fenômeno ocorre em ambientes dinamicamente mutáveis e não estacionários, nos quais a distribuição subjacente dos dados evolui de maneira imprevisível [Webb et al., 2016]. Consequentemente, torna-se necessária a adoção de estratégias de aprendizado adaptativo, já que modelos preditivos treinados em dados históricos tendem a se tornar progressivamente obsoletos diante dessas mudanças [Lu et al., 2020].

O *concept drift* pode ser categorizado em função de qual componente da distribuição conjunta está em transformação, bem como pela sua natureza temporal. Com base na mudança da distribuição, identificam-se dois tipos: o *concept drift* real e o *concept drift* virtual. A II.1 exemplifica visualmente esses conceitos.

O *concept drift* real refere-se a mudanças na distribuição condicional $P(y|X)$, ou seja, na relação entre as características de entrada e a variável alvo. Esta alteração afeta a fronteira de decisão, tornando o modelo preditivo anterior obsoleto e exigindo adaptação. Tais mudanças podem ocorrer com ou sem alteração em $P(X)$. Por outro lado, o *concept drift* virtual ocorre quando a distribuição dos dados de entrada $P(X)$ muda, mas essa alteração não afeta a distribuição condicional $P(y|X)$. O *drift* virtual, embora represente uma mudança na distribuição dos dados, não altera o conceito alvo em si, mas pode levar a mudanças na fronteira de decisão [Gama et al., 2014].

É importante notar que, neste trabalho, o termo *drift* virtual é empregado em sentido operacional, e não estritamente conforme a definição de Gama et al. [2014]. Na definição formal, o *drift* virtual implica que a mudança em $P(X)$ não afeta $P(y|X)$, ou seja, a relação entre passes e probabilidade de gol permaneceria inalterada. A premissa subjacente desta dissertação, contudo, é distinta: assume-se que a mudança detectável em $P(X)$ (o padrão de passes) é um *precursor* de uma mudança iminente em $P(y|X)$, isto é, na probabilidade de ocorrência de gol. O fenômeno é tratado como *drift* virtual operacionalmente porque a detecção se baseia exclusivamente em $P(X)$, sem acesso a rótulos em tempo real, mas a motivação preditiva pressupõe que essa mudança em $P(X)$ antecede uma alteração em $P(y|X)$, o que é conceitualmente mais próximo de um precursor de *drift* real do que de *drift* virtual puro.

Considerando-se o padrão temporal, encontram-se quatro possíveis classificações: *Sudden Drift*, *Gradual Drift*, *Incremental Drift* e *Reocurring Concepts*. O *Sudden Drift* caracteriza-se pela substituição rápida de um conceito por outro em um curto período de tempo, frequentemente por meio

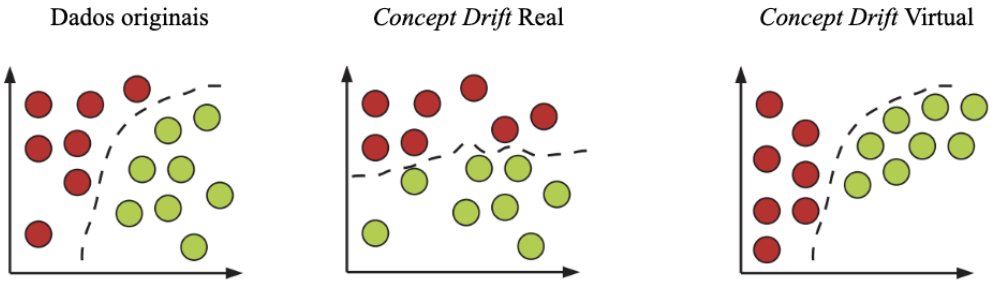

Figura II.1: Tipos de *concept drift* por mudança. Adaptado de [Gama et al., 2014].

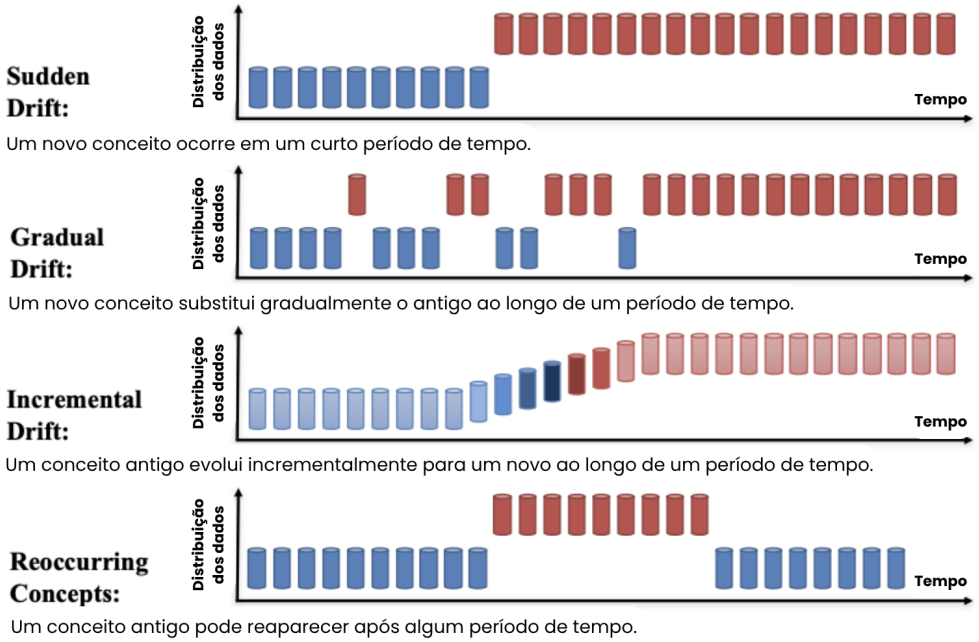

Figura II.2: Tipos de *concept drift* por padrão temporal. Adaptado de Lu et al. [2020].

de uma mudança súbita. No *Gradual Drift*, o novo conceito substitui o antigo de forma lenta, ao longo de um período extenso, com diversos conceitos intermediários envolvidos na transição. No *Incremental Drift*, o conceito antigo evolui gradualmente para um novo conceito durante um período, e os conceitos intermediários representam uma mistura progressiva dos conceitos inicial e final. Por fim, com os *Reocurring Concepts*, conceitos previamente observados podem reaparecer após um período de tempo. A Figura II.2 explica visualmente tais conceitos.

Compreender as diferentes formas de *concept drift* é essencial, pois cada tipo de variação requer abordagens específicas de detecção e adaptação, e a escolha adequada do método impacta diretamente a eficácia das análises preditivas desenvolvidas ao longo deste estudo.

II.1.1 Detecção de *Concept Drift* como Detecção de Eventos

A detecção de *concept drift* pode ser empregada diretamente como mecanismo de detecção de eventos em séries temporais. Essa perspectiva é possível porque uma mudança persistente na distribuição dos dados, o *drift*, sinaliza que o processo subjacente entrou em um novo regime, enquanto eventos pontuais (anomalias) correspondem a desvios transitórios após os quais o fluxo retorna ao estado anterior [Hinder et al., 2024; Li and Müller, 2022]. A distinção operacional, portanto, é temporal: *drift* implica que a nova distribuição se mantém estável por um período prolongado, ao passo que a anomalia é de curta duração [Li and Müller, 2022].

Esse dualismo é explorado em diferentes domínios. Hinder et al. [2024] apontam aplicações de monitoramento em infraestrutura crítica, como a detecção de vazamentos em redes de distribuição de água, nas quais a identificação de um *drift* na distribuição dos sensores equivale à detecção do evento operacional de interesse. Na área médica, Kore et al. [2024] demonstram que a detecção direta de *drift* nos dados de entrada, e não apenas o monitoramento do desempenho do modelo, foi o único método capaz de identificar a emergência da COVID-19 em radiografias torácicas, evento que não produziu degradação imediata nas métricas agregadas do classificador. Ambos os estudos reforçam que a detecção baseada em dados é mais sensível a mudanças no processo gerador do que o monitoramento de performance *a posteriori*.

Em contexto industrial, Tavares et al. [2025] demonstram essa conexão de forma sistemática: utilizando o *dataset* 3W, um *benchmark* de eventos críticos e raros em poços de petróleo, os autores interpretam o início de um *drift* na distribuição dos dados como o ponto de partida de um evento operacional, como a formação de hidratos ou mudanças abruptas no fluxo de produção. Nessa abordagem, detectar a mudança de conceito equivale a antecipar o evento.

Os estudos citados adotam a distinção entre *drift* real e virtual [Gama et al., 2014; Hinder et al., 2024]: enquanto o *drift* real implica mudança na relação condicional $P(Y \mid X)$, o *drift* virtual refere-se a alterações na distribuição das variáveis de entrada $P(X)$ sem que a relação com o alvo necessariamente se altere. Em ambientes de monitoramento não supervisionado, apenas o *drift* virtual precisa ser considerado [Hinder et al., 2024], e a detecção baseada exclusivamente em propriedades estatísticas de $P(X)$ é a alternativa viável quando rótulos em tempo real são custosos ou indisponíveis [Kore et al., 2024; Tavares et al., 2025].

Quanto à avaliação, Tavares et al. [2025] adotam o *SoftED* [Salles et al., 2024] como *framework* de métricas, reconhecendo que detectores podem antecipar ou atrasar levemente o instante exato do evento e que uma tolerância temporal é necessária para uma avaliação justa. Entre os detectores individuais avaliados (*Adaptive Windowing* (ADWIN), *Kolmogorov-Smirnov Windowing* (KSWIN) e Page-Hinkley), o Page-Hinkley apresentou o melhor equilíbrio entre *F1 Score* (F1) e cobertura (F1 = 0,46; cobertura = 90%), sugerindo que detectores determinísticos e de baixa complexidade

podem ser competitivos mesmo em ambientes ruidosos e de alta dimensionalidade.

Este trabalho segue a mesma perspectiva: o *drift* virtual na distribuição de passes é tratado como precursor de gols, e a avaliação adota uma variante assimétrica do *SoftED* com tolerância $K = 10$ minutos anterior ao evento. A abordagem é integralmente não supervisionada, dispensando rótulos em tempo real e produzindo resultados desde a primeira partida analisada. A qualidade dos alarmes é avaliada tanto em protocolo *in-sample* quanto em divisão temporal com 190 partidas de treino e 190 de teste. Até onde foi possível identificar na literatura indexada pela Scopus em abril de 2026, nenhum trabalho anterior aplicou técnicas de detecção de *concept drift* para a identificação de eventos esportivos em fluxos de dados de partidas de futebol, o que posiciona esta dissertação como uma contribuição inicial nessa direção.

II.1.2 Algoritmos de Detecção de *Drift*

Page-Hinkley

O teste de Page-Hinkley [Page, 1954] é um método de inspeção sequencial paramétrico que detecta mudanças na média de uma série temporal por meio da acumulação de desvios em relação a um valor de referência. A ideia central é monitorar se a soma cumulativa dos desvios ultrapassa um limiar que indicaria uma mudança sistemática na distribuição.

Para o modo de detecção de aumento (*up*), a estatística acumulada é definida como:

$$PH_t = \sum_{i=1}^{t} \left(x_i - \bar{x}_t - \delta \right), \tag{II.2}$$

onde x_i é a observação no instante i, $\bar{x}_t$ é a média observada até t e $\delta > 0$ é o desvio mínimo detectável, que filtra flutuações pequenas. O algoritmo rastreia o mínimo acumulado $\overline{PH}_t = \min_{i \leq t} PH_i$ e sinaliza *drift* quando:

$$PH_t - \overline{PH}_t > \lambda, \tag{II.3}$$

sendo λ o limiar de detecção. O parâmetro δ controla a sensibilidade a pequenas variações: valores menores tornam o detector mais sensível a pequenas mudanças, aumentando o risco de falsos positivos; valores maiores exigem desvios mais expressivos para acionar um alarme.

O detector opera em dois modos complementares: o modo *up* detecta aumentos na média e o modo *down* detecta quedas, invertendo o sinal dos desvios acumulados. A combinação de ambos os modos permite capturar variações bidirecionais na série de passes, sem pressupor a direção da mudança.

No contexto deste trabalho, o Page-Hinkley é aplicado sobre a média móvel da frequência de passes como detector de *drift* virtual. Por ser determinístico — dado o mesmo fluxo de entrada e os

mesmos parâmetros, produz sempre o mesmo resultado —, o algoritmo não introduz variabilidade estocástica nos experimentos, o que contribui para a interpretação e reprodutibilidade dos resultados. Adicionalmente, sua baixa complexidade computacional dispensa dados históricos rotulados para treinamento, permitindo que a detecção se inicie desde a primeira observação de cada partida.

KSWIN

O KSWIN (*Kolmogorov-Smirnov Windowing*) é um detector de *concept drift* não paramétrico que utiliza o teste estatístico de *Kolmogorov-Smirnov* (KS) para identificar mudanças na distribuição de fluxos de dados [Raab et al., 2020]. Por não assumir nenhuma forma paramétrica para a distribuição dos dados, o algoritmo é aplicável a cenários em que a distribuição subjacente é desconhecida ou não gaussiana.

O funcionamento baseia-se em uma janela deslizante Ψ de tamanho n, dividida em duas subjanelas de tamanho r: a janela recente R, composta pelos r pontos mais recentes do fluxo, e a janela de referência S, composta por r amostras selecionadas aleatoriamente da porção mais antiga de Ψ. A cada nova observação, o teste KS compara as funções de distribuição acumulada empíricas de R e S. Se a distância máxima entre essas distribuições exceder um limiar determinado pelo nível de significância α e pelo tamanho das subjanelas, o algoritmo sinaliza a ocorrência de *drift*.

O parâmetro α controla a sensibilidade do detector: valores menores tornam o teste mais conservador, reduzindo falsos positivos; valores maiores aumentam a sensibilidade à mudança. O tamanho r das subjanelas influencia a resolução temporal da detecção: janelas menores respondem mais rapidamente a mudanças, mas com maior variância estatística.

No contexto deste trabalho, o KSWIN é empregado como detector de *drift* virtual sobre a média móvel de passes, sem necessidade de rótulos em tempo real. A principal vantagem em relação a detectores paramétricos é que o teste KS detecta qualquer diferença distribucional entre R e S, o que o torna sensível a alterações na forma ou na variância da distribuição de passes ao longo da partida. Como o teste KS avalia a distribuição completa (e não apenas a média), o KSWIN é capaz de detectar mudanças de nível, variância ou forma na série de passes, tornando-o mais sensível que detectores baseados exclusivamente em comparação de médias, embora com maior propensão a falsos positivos.

ADWIN

O ADWIN (*Adaptive Windowing*) é um algoritmo que mantém uma janela de comprimento variável de amostras recentes de dados. Em vez de fixar um tamanho de janela a priori, o algoritmo recalcula o tamanho conforme a taxa de mudança observada nos próprios dados, comparando as distribuições de duas subjanelas dentro da janela principal [Bifet and Gavaldà, 2007].

O funcionamento central baseia-se na comparação estatística de subjanelas. Dada uma janela J com os dados mais recentes, o algoritmo avalia todas as possíveis partições de J em duas subjanelas J_0 (mais antiga) e J_1 (mais recente). Sempre que as médias dessas subjanelas forem suficientemente distintas, isto é, quando

$$|\bar{\mu}_{J_0} - \bar{\mu}_{J_1}| \geq \epsilon_{\text{cut}}, \tag{II.4}$$

o algoritmo conclui que houve mudança e descarta J_0. O limiar ϵ_{cut} é calculado a partir do parâmetro de confiança δ, definido pelo usuário, e do comprimento das subjanelas, utilizando desigualdades de concentração (como Hoeffding ou Bernstein) que fornecem garantias teóricas sobre as taxas de falsos positivos e falsos negativos [Bifet and Gavaldà, 2007].

O parâmetro δ controla diretamente a sensibilidade do detector: valores menores tornam o algoritmo mais conservador, reduzindo o número de alarmes; valores maiores aumentam a sensibilidade, com risco de mais falsos positivos. A janela J cresce enquanto os dados são estacionários, aumentando a precisão estatística, e encolhe ao detectar mudanças, descartando dados obsoletos automaticamente.

No contexto deste trabalho, o ADWIN é empregado como detector de *drift* virtual, monitorando mudanças na distribuição da média móvel de passes sem necessidade de rótulos em tempo real. Por comparar subjanelas em ambas as direções, o algoritmo é capaz de capturar tanto aumentos quanto reduções no volume de passes, sem pressupor a direção da mudança. Adicionalmente, a natureza adaptativa da janela J dispensa a definição prévia de quando esperar a mudança, o que é adequado ao contexto intra-partida, onde o *drift* pode ocorrer em diferentes momentos em partidas diferentes. Ou seja, o próprio algoritmo decide quanto da série histórica guardar: se a partida demorar mais para ter *drift*, a janela cresce; se o *drift* for rápido, a janela encolhe.

II.2 Avaliação de Detectores de Drift

II.2.1 Métricas Clássicas

A detecção de gols é um problema de classes desbalanceadas: em uma partida de 90 minutos com intervalos de 1 minuto, o número de minutos sem gol supera amplamente o número de minutos com gol. Nesse contexto, métricas baseadas em acurácia são inadequadas, pois um detector que nunca dispara alarmes obtém acurácia próxima de 100%, mascarando sua completa ausência de poder preditivo [Sujon et al., 2025]. As métricas adotadas neste trabalho foram selecionadas por sua robustez ao desbalanceamento e por capturarem diferentes dimensões do desempenho do detector.

A partir de uma matriz de confusão com verdadeiros positivos (*True Positive* (TP)), falsos positivos (*False Positive* (FP)), falsos negativos (*False Negative* (FN)) e verdadeiros negativos (*True*

Negative (TN)), definem-se:

$$\text{Precisão} = \frac{TP}{TP + FP} \tag{II.5}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{II.6}$$

A *precisão* mede a proporção de alarmes corretos entre todos os alarmes emitidos; o *recall* mede a proporção de gols detectados entre todos os gols ocorridos. O *F1-score* é a média harmônica entre ambos:

$$F_1 = \frac{2 \cdot \text{Precisão} \cdot \text{Recall}}{\text{Precisão} + \text{Recall}} \tag{II.7}$$

O F_1 penaliza igualmente falsos positivos e falsos negativos. Quando o custo de alarmes desnecessários é maior do que o custo de gols não detectados — como no cenário deste trabalho, em que um falso alarme consome o período de *cooldown* impedindo novas detecções —, é preferível utilizar o F_β com $\beta < 1$ [van Rijsbergen, 1979]. Especificamente, adota-se o $F_{0,5}$, que atribui o dobro de peso à *precisão* em relação ao *recall*:

$$F_{0,5} = \frac{(1 + 0{,}5^2) \cdot \text{Precisão} \cdot \text{Recall}}{0{,}5^2 \cdot \text{Precisão} + \text{Recall}} \tag{II.8}$$

Por fim, o *Matthew's Correlation Coefficient* (MCC) incorpora todos os elementos da matriz de confusão, sendo reconhecidamente robusto a distribuições de classes assimétricas [Sujon et al., 2025]:

$$\text{MCC} = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{II.9}$$

O MCC varia entre -1 (classificação invertida) e $+1$ (classificação perfeita), com 0 indicando desempenho equivalente ao acaso. Sujon et al. [2025] demonstram empiricamente que F_1 e MCC são as métricas mais estáveis e informativas em cenários desbalanceados, onde acurácia e *precisão* isoladas tendem a superestimar o desempenho real.

Neste trabalho, o MCC é adotado como métrica primária de comparação entre detectores, por incorporar todos os elementos da matriz de confusão e ser robusto ao desbalanceamento severo entre minutos com e sem gol. O F_1 e o $F_{0,5}$ são reportados como métricas complementares, permitindo analisar o equilíbrio entre precisão e *recall* e o custo relativo dos falsos positivos. Adicionalmente, o MCC permite aproximar os resultados aos de Lang et al. [2025], que adotam a mesma métrica para avaliar detectores de eventos em partidas de futebol, embora a comparação direta seja parcial em razão das diferenças na janela de tolerância temporal adotada em cada trabalho.

II.2.2 SoftED Evaluation

As métricas clássicas de classificação adotam uma lógica binária: um alarme é verdadeiro positivo somente se coincidir exatamente com o instante do evento; qualquer desvio temporal, mesmo de um único período, é penalizado da mesma forma que uma detecção completamente equivocada. Salles et al. [2024] argumentam que essa rigidez é inadequada para detecção de eventos em séries temporais, pois alarmes próximos ao evento são frequentemente valiosos na prática, pois permitem respostas antecipadas e refletem efeitos precursores que o detector efetivamente capturou.

Para incorporar tolerância temporal, Salles et al. [2024] propõem as métricas *SoftED*, inspiradas em conjuntos *fuzzy* aplicados à dimensão temporal. A cada detecção d_i é atribuída uma pontuação contínua $\mu_{e_j}(t_{d_i}) \in [0, 1]$ que mede o grau de proximidade temporal ao evento e_j:

$$\mu_{e_j}(t_{d_i}) = \max\left(\min\left(\frac{t_{d_i} - (t_{e_j} - k)}{k},\ \frac{(t_{e_j} + k) - t_{d_i}}{k}\right), 0\right) \tag{II.10}$$

A função define uma janela simétrica $[t_{e_j} - k,\ t_{e_j} + k]$ ao redor do evento t_{e_j}, com pontuação máxima igual a 1,0 no instante exato do evento e decrescimento linear até 0 nas extremidades da janela. Alarmes fora da janela recebem pontuação 0 e são tratados como falsos positivos. O parâmetro k é definido pelo domínio da aplicação e controla a tolerância admitida.

A partir das pontuações individuais $ds(d_i)$, as versões *soft* das métricas clássicas são definidas como:

$$\mathrm{TP}_s = \sum_{i=1}^{n} ds(d_i), \qquad \mathrm{FN}_s = m - \mathrm{TP}_s, \qquad \mathrm{FP}_s = \sum_{i=1}^{n} (1 - ds(d_i)) \tag{II.11}$$

onde m é o número total de eventos na série. Essas métricas preservam a mesma escala e interpretação das métricas rígidas, possibilitando o cômputo de versões *soft* de *precisão*, *recall* e F_1. A Figura II.3 explica a diferença da avaliação *soft*. Na série temporal, o diamante azul representa o evento de referência e o círculo verde o alarme emitido. Na avaliação *hard*, apenas alarmes que coincidem exatamente com o evento recebem pontuação 1,0. Na avaliação *soft*, alarmes próximos recebem crédito parcial proporcional à distância ao evento.

No contexto de predição de gols, um alarme só é útil se emitido *antes* do evento: avisos após o gol não permitem nenhuma ação relevante. Por isso, este trabalho adota uma variante assimétrica da função de pertinência original: em vez de avaliar o alarme em relação ao instante exato do gol t, o gol é deslocado K minutos para a trás (passado), tornando $t - K$ o novo ponto de referência da avaliação. A janela de tolerância cobre o intervalo $[t - K, t]$, com pontuação máxima (1,0) atribuída a alarmes emitidos exatamente em $t - K$ — isto é, com K minutos de antecedência — e decrescendo linearmente até 0 no instante t do gol, penalizando alarmes tardios e ignorando completamente detecções posteriores ao evento. A construção completa do pipeline de avaliação,

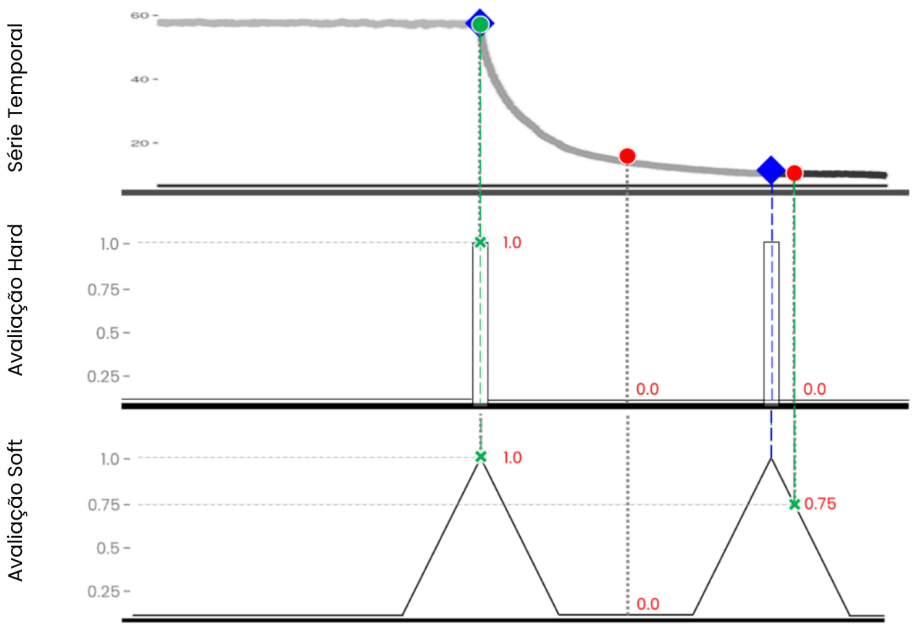

Figura II.3: Comparação entre avaliação *hard* e *soft*. Adaptado de Salles et al. [2024].

incluindo os critérios de TP, FP e FN sob essa métrica, é detalhada na Seção IV.3. Formalmente, o *score* atribuído a um alarme emitido no instante a é dado por:

$$\text{score}(a) = \begin{cases} 1 - \dfrac{t-a}{K} & \text{se } t - K \leq a \leq t \\ 0 & \text{caso contrário} \end{cases} \tag{II.12}$$

onde a denota o instante do alarme. Essa escolha reflete o objetivo do pipeline: detectar mudanças de padrão com antecedência suficiente para que o alarme seja operacionalmente relevante.

II.3 Variáveis do Futebol

Existem três principais fontes de dados relacionadas ao futebol: (i) os *logs* de eventos, que descrevem as ações ocorridas durante uma partida e são coletados por meio de *softwares* de marcação; (ii) os dados de rastreamento por vídeo, que registram as trajetórias dos jogadores a partir da análise de gravações; e (iii) os dados de *Global Positioning System* (GPS), obtidos por dispositivos utilizados pelos atletas para monitorar seus movimentos [Pappalardo et al., 2019]. Nos últimos anos, diversas empresas especializadas em análise esportiva (Opta[1], Wyscout[2] e StatsBomb[3]) passaram a disponibilizar dados de futebol, com maior ênfase em eventos, e, progressivamente, também dados de *tracking*, enquanto os dados de GPS permanecem, em geral, restritos aos clubes [Goka et al.,

[1]Disponível em: `https://www.statsperform.com/pt-br/opta-football/`
[2]Disponível em: `https://www.hudl.com/products/wyscout`
[3]Disponível em: `https://www.hudl.com/en_gb/products/statsbomb`

2024].

Dados de eventos fornecem informações espaço-temporais detalhadas, incluindo o instante do evento (*timestamp*), suas coordenadas espaciais, a classificação categórica da ação (passe, chute, substituição), a identificação dos jogadores e os resultados da ação (sucesso ou fracasso) [Goka et al., 2024]. Dentre os principais eventos táticos em uma partida estão passes, finalizações, dribles, cruzamentos, pressões e jogadas de bola parada, que refletem tanto a capacidade ofensiva de criar oportunidades de gol quanto a organização defensiva de impedir a progressão adversária [Pappalardo et al., 2019].

Esses eventos servem de base para o cálculo de métricas táticas, como o número de finalizações, ataques perigosos e o *Expected Goals* (xG), que quantifica a probabilidade de conversão de cada chute em gol. Métricas de volume, em particular o número de passes por minuto, também capturam mudanças no ritmo de jogo, sendo relevantes tanto para a perspectiva ofensiva (aumento de atividade antes de marcar) quanto defensiva (queda de atividade antes de sofrer o gol). A análise dessas métricas permite avaliar a dinâmica das equipes e sua relação potencial com a ocorrência de gols [Lago-Peñas et al., 2011].

Capítulo III Trabalhos Relacionados

O objetivo desta revisão da literatura é identificar estudos em ciência de dados relacionados à predição de gols e outros eventos no futebol. Para tanto, foram considerados apenas artigos publicados em periódicos ou conferências, redigidos em inglês, obtidos por meio de busca realizada na base Scopus em abril de 2026.

A busca sistemática na base Scopus foi realizada utilizando a seguinte *string* de pesquisa: TITLE-ABS-KEY ((*"sport"* OR *"football"* OR *"soccer"*) AND (*"event prediction"* OR *"goal prediction"* OR *"goal forecast*"* OR *"event forecast*"* OR *"concept drift"* OR *"drift detection"*)). Essa consulta retornou um total de 68 artigos, que posteriormente foram refinados seguindo o modelo *Preferred Reporting Items for Systematic Reviews* (PRISMA), garantindo a seleção sistemática e transparente dos estudos incluídos na revisão, conforme a Figura III.1.

Além dos 68 artigos identificados na busca sistemática, outros 5 foram adicionados por meio da técnica de *snowballing*. Após a leitura de títulos e resumos, 41 estudos foram excluídos, restando 32 para leitura completa. Entre os 41 excluídos, 6 abordavam outros esportes, 3 tratavam de *video games*, 3 discutiam temas distintos dentro do esporte (como lesões) e os demais 29 eram não relacionados ao tema, incluindo previsões climáticas, análise de sentimento e predição de participantes em eventos (por exemplo, quantas pessoas compareceriam a um festival). Dos 32 artigos que avançaram para a leitura completa, 16 apresentaram maior proximidade com o objetivo deste trabalho por tratarem de predição de gols e/ou eventos, utilizarem dados baseados em eventos, explorarem informações intra-partida e/ou abordarem *concept drift*. Os critérios de exclusão consideraram: artigos não escritos em inglês, estudos baseados em dados de vídeo e artigos do tipo *survey*.

Com o objetivo de contextualizar o presente estudo (detecção de *concept drift* intra-partida para determinar o momento do gol), a análise da literatura será dividida em três eixos principais: (1) Modelagem dinâmica e predição intra-jogo; (2) Mudança de desempenho; e (3) Previsão de eventos sequenciais.

III.1 Modelagem Dinâmica e Predição Intra-Jogo

A previsão em tempo real (*in-game prediction*) é o campo mais diretamente relacionado ao objetivo de determinar o momento de um gol. Esses estudos se concentram em atualizar as pro-

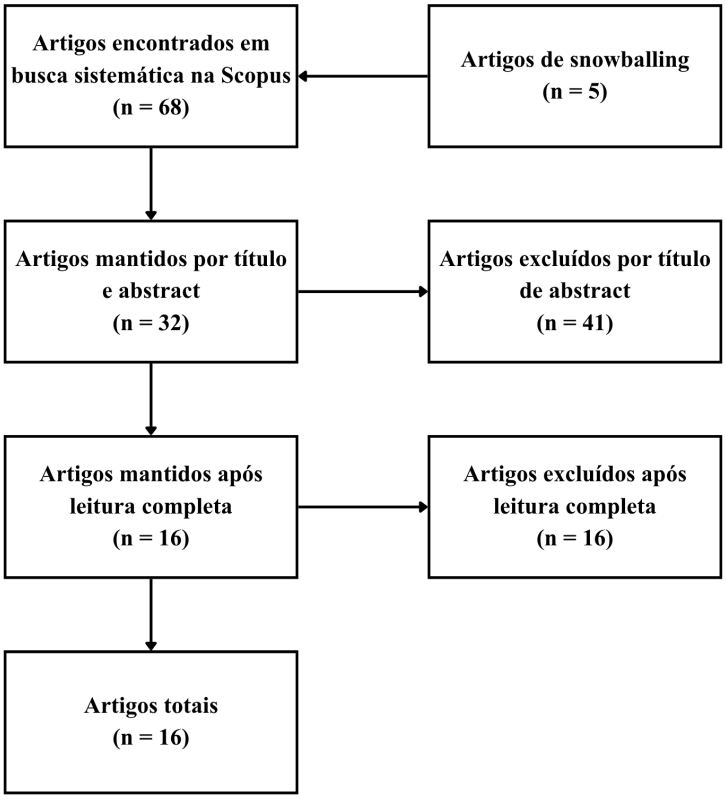

Figura III.1: Fluxograma PRISMA do processo de seleção dos artigos.

babilidades de resultado ou contagem de gols à medida que a partida avança, usando o tempo transcorrido e o estado atual da partida (placar, cartões, etc.). A ideia de estimar a probabilidade de gol a partir do estado do jogo remonta aos modelos de *expected goals* (xG), amplamente difundidos na primeira metade da década de 2010. O framework *Valuing Actions by Estimating Probabilities* (VAEP) [Decroos et al., 2019] generalizou esse conceito para qualquer ação intra-partida, atribuindo valor a cada evento com base na variação que ele causa nas probabilidades de marcar e sofrer gol nos próximos dez eventos. O VAEP tornou-se referência na área e fundamenta vários trabalhos subsequentes desta revisão.

O trabalho de Arntzen and Hvattum [2021] emprega um modelo de Risco Concorrente (baseado em análise de sobrevivência) para gerar previsões intra-jogo. Este modelo estima as taxas de pontuação (*scoring rates*) e pode ser estendido para incluir a taxa de ocorrência de cartões vermelhos, juntamente com covariáveis que refletem o estado atual do jogo, como a diferença de gols (placar) e se um time tem menos jogadores (cartão vermelho). A perda média preditiva desse modelo, como esperado, diminui à medida que a partida se aproxima do final. Também utilizando análise de sobrevivência, Dutta et al. [2024] focam em prever o tempo até o primeiro gol ocorrer no segundo tempo. O estudo revelou que as estatísticas de desempenho do primeiro tempo (como passes completados e defesas do goleiro) influenciam significativamente o *timing* do gol na etapa seguinte. Essa

abordagem é relevante, pois trata a ocorrência do gol como um evento temporal, em vez de apenas uma probabilidade binária.

Dobson and Goddard [2017] utilizam um modelo de simulação calibrado por funções de risco estimadas (*estimated hazard functions*) para a chegada de gols e expulsões, obtendo probabilidades intra-jogo condicionadas ao estado da partida. Robberechts et al. [2019] propuseram um modelo Bayesiano de probabilidade de vitória intra-jogo que lida com a natureza de baixa pontuação do futebol, modelando o número de gols futuros de cada equipe como um processo estocástico temporal. São utilizadas características contextuais que refletem o desempenho intra-jogo, cuja importância varia ao longo do tempo (não linearmente), como passes de ataque bem-sucedidos e força em duelos. O método *In-Game Outcome Prediction* (IGSOP) proposto por Yao et al. [2022] divide a partida em pequenos quadros de tempo (200 quadros) e modela a probabilidade de um gol ser marcado por uma Distribuição de Bernoulli dentro de cada quadro. Este método explicitamente visa prever o resultado em tempo real e busca capturar as mudanças de momentum que ocorrem após um gol. Notavelmente, o IGSOP demonstrou desempenho superior em comparação com o modelo de Distribuição de Poisson nos momentos finais do jogo.

Por fim, Capobianco et al. [2019] busca analisar o jogo em tempo real para prever resultados e a contagem de gols, permitindo que o treinador, por exemplo, possa mudar a tática entre o primeiro e o segundo tempo. O método utiliza algoritmos de *machine learning* supervisionado, como o *Random Forest*, para construir dois modelos: o primeiro prevê a vitória ou derrota, e o segundo modela o número de gols marcados pela equipe vencedora (menos de dois gols ou maior igual a dois gols).

III.2 Mudança de Desempenho

Apesar dessa seleção, nenhum dos estudos identificados utilizou o conceito de *concept drift* como um alarme preditivo para o momento do gol. O trabalho de Lühr and Lazarescu [2007] propõe um sistema de rastreamento de *concept drift* voltado à análise de variações na performance dos atletas ao longo de múltiplas partidas, utilizando o *Competing Windows Algorithm* (CWA) para detectar mudanças permanentes ou temporárias em métricas de desempenho individuais. Embora não trate de análise intra-partida nem de previsão de gols, é o único trabalho identificado na busca que aplica formalmente algoritmos de detecção de *concept drift* ao domínio esportivo.

Na direção da análise intra-partida, um tema relacionado ganha destaque: a análise de *momentum* da partida, que busca descrever as variações no desempenho e na probabilidade de sucesso das equipes ao longo do jogo. O trabalho de Lang et al. [2025] propõe uma métrica de *match momentum* definida como a diferença entre as probabilidades de gol previstas para as duas equipes. Utilizando *machine learning*, busca-se prever a ocorrência de Eventos Relacionados a Sucesso ou Pontuação (*Success/Score-Related Event* (SRE)), como chutes, escanteios e entradas na área,

em um *prediction window* (por exemplo, nos próximos 3 minutos), baseando-se em indicadores de desempenho (*Performance Indicator* (PI)) de um período passado (exemplo, últimos 15 minutos). O estudo foca na previsão de eventos além do evento imediatamente seguinte e confirma que esta abordagem é um reflexo estável do desempenho da equipe. Embora a métrica de *match momentum* capture o resultado de uma mudança de desempenho, o presente trabalho busca ir além, utilizando a detecção explícita de *concept drift* como mecanismo primário para identificar o momento em que a probabilidade de gol sofre uma alteração estatisticamente significativa.

III.3 Previsão de Eventos Sequenciais

Outra linha de pesquisa utiliza modelos sequenciais complexos para prever o próximo evento, o que serve de base para a quantificação do potencial ofensivo. Goka et al. [2024] demonstraram a eficácia da análise de séries temporais bidirecionais para prever o próximo evento de ação (Passe, Drible, Cruzamento, Chute). Dado que os jogadores escolhem suas ações considerando mudanças subsequentes na partida, o estudo ressalta a importância do contexto futuro para a análise de eventos em tempo real.

O *Large Events Model* (LEM) [Mendes-Neves et al., 2024] é um modelo generativo que prevê o próximo evento, incluindo a variável *isGoal*, cobrindo 33 tipos de eventos e visando simular partidas completas. Em trabalho subsequente, Mendes-Neves et al. [2026] propõem uma versão unificada do LEM baseada em um modelo autorregressivo tabular com mascaramento causal, substituindo a cadeia de classificadores original, que apresentava problemas de sincronização, escalabilidade limitada e janela de contexto restrita a um único evento. O novo *framework* prevê sequencialmente cada atributo do próximo evento condicionado ao contexto dos três eventos anteriores e ao estado global da partida, com modelos de 100k a 10M parâmetros; superando o LEM original na maioria das métricas. O LEM é também utilizado no *framework Open Spatio-Temporal Agent Research Lab* (OpenSTARLab) [Yeung et al., 2025], que emprega *deep learning* para a previsão de eventos e suporta métricas avançadas baseadas no potencial de ataque.

Romero et al. [2026] estendem o *framework* LEM para incorporar informações de jogador e time na previsão do tipo do próximo evento, investigando o impacto de identidade de time, identidade de jogador e papel posicional sobre o desempenho preditivo. Utilizando o dataset Wyscout (Premier League 2017/18), os autores comparam variantes do modelo *Multi-Layer Perceptron* (MLP) com *eXtreme Gradient Boosting* (XGBoost) e *baselines* heurísticos, mostrando que ambos os modelos aprendidos superam os *baselines* e que o papel posicional do jogador contribui mais do que sua identidade individual. Um resultado relevante para o presente trabalho é que o *baseline ConditionalMajority*, que prevê o evento mais frequente condicionado ao evento anterior, obtém desempenho surpreendentemente forte, sugerindo que a dependência sequencial entre ações consecutivas é por si

só um sinal preditivo robusto.

Umemoto et al. [2025] propõem o *Generalized Valuing Defense by Estimating Probabilities* (GVDEP), uma métrica de avaliação defensiva baseada na predição simultânea de quatro probabilidades de evento (marcar, sofrer gol, recuperar a bola e sofrer ataque efetivo) nos próximos eventos da partida, a partir do estado espacial do jogo. Utilizando XGBoost sobre dados StatsBomb, o modelo pondera essas probabilidades pelos valores VAEP [Decroos et al., 2019] de cada ação, quantificando a contribuição defensiva de cada jogador para a dinâmica de gol. A abordagem compartilha com o presente trabalho o uso de dados abertos StatsBomb e a premissa de que a probabilidade de gol pode ser estimada continuamente a partir de eventos intra-partida.

Outro modelo relevante na literatura é o Seq2Event [Simpson et al., 2022], que utiliza componentes Transformer/RNN para prever o próximo evento de partida e gerou a métrica Poss-Util, que integra a expectativa de ataque (chute ou cruzamento) ao longo de uma posse de bola. Por sua vez, o *Scoring COnvolution for next-Event REcognition* (SCORE) [Alves, 2025] utiliza uma abordagem convolucional para prever o próximo evento relevante (como gol ou finalização no alvo) com base apenas em dados de eventos sem rastreamento posicional. O modelo é explicitamente concebido para análise em tempo real e para detectar quando a probabilidade de um evento raro está mudando.

III.4 Síntese Comparativa

A literatura demonstra robustez na modelagem preditiva dinâmica e na análise de sequências de eventos que culminam em ataques. As principais características dos artigos revisados e suas contribuições para o tema estão sintetizados na Tabela III.1. Muitos desses trabalhos corroboram a ideia de que o desempenho e o potencial de gol variam continuamente ao longo da partida. No entanto, seus objetivos concentram-se em recalcular a probabilidade de gol, em vez de prever o *timing* do evento.

O presente trabalho, portanto, preenche uma lacuna ao aplicar formalmente os algoritmos de detecção de *concept drift* diretamente às séries temporais de passes intra-partida como *proxy* para mudanças na probabilidade de gol, fornecendo um sinal mais direto e interpretável de instabilidade preditiva associado à antecipação do gol.

III.5 Predição In-Play com Indicadores de Performance

Dentre os artigos apresentados na Tabela III.1, o trabalho de Lang et al. [2025] é um dos que mais se aproxima da dissertação. Os autores investigam em que medida eventos de sucesso relacionados ao placar (SRE), como gols, chutes e escanteios, podem ser preditos a partir de indicadores de performance coletados em uma janela de tempo passada durante a partida. Parte-se do conceito de

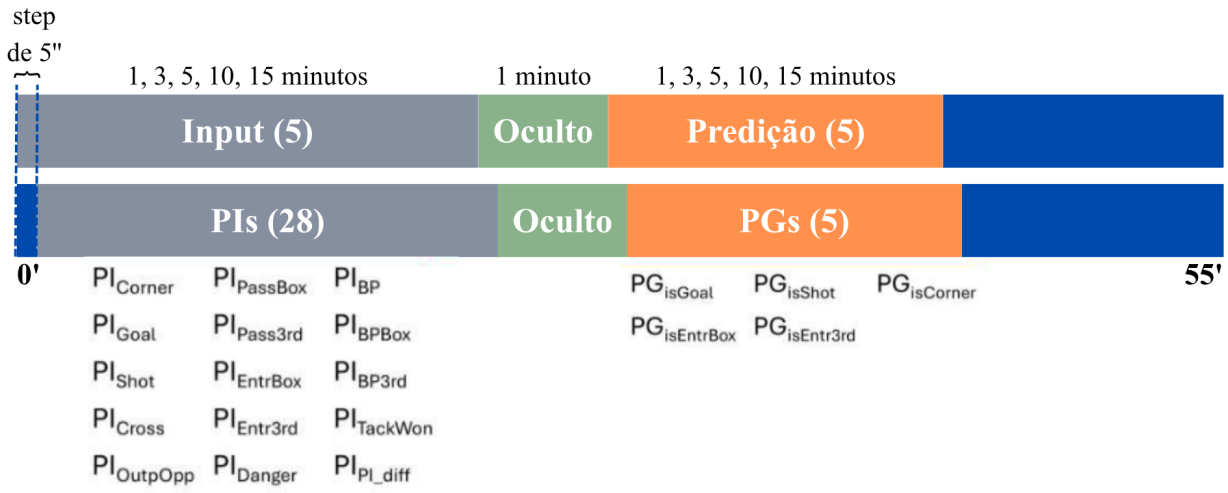

Figura III.2: Abordagem *In-Play Prediction Masking*. Adaptado de Lang et al. [2025].

match momentum, definido como a diferença de desempenho entre as equipes ao longo de curtos intervalos, e utiliza mudanças no *match momentum* para entender tendências do jogo e probabilidade de ocorrência de eventos futuros. O estudo utiliza dados de 102 partidas da Bundesliga e avalia 28 indicadores de performance através de cinco modelos distintos de *machine learning*, empregando o MCC como critério principal de ranqueamento, por sua robustez em cenários com classes desbalanceadas, característica inerente à raridade dos eventos de sucesso em futebol.

A principal contribuição metodológica do trabalho é a técnica denominada *In-Play Prediction Masking*: uma janela oculta de um minuto é inserida entre a janela de entrada e a janela de predição, impedindo que indicadores com relação causal direta ao evento, como entradas na área que precedem chutes, dominem o modelo e obscureçam a capacidade preditiva de indicadores de mais longo prazo. Os autores utilizam uma abordagem de janela deslizante (*rolling window*) com passo de cinco segundos sobre o tempo efetivo de jogo, testando janelas de entrada e de predição de 1, 3, 5, 10 e 15 minutos em todas as combinações. Os resultados indicam que a janela de entrada de 5 minutos supera sistematicamente a de 15 minutos, sugerindo que eventuais sequências de desempenho (*performance streaks*) raramente se sustentam por períodos superiores a esse intervalo. A Figura III.2 ilustra o design da abordagem *In-Play Prediction Masking*, na qual a banda superior ilustra as janelas de entrada (*Input*), oculta e de predição. A banda inferior exibe os indicadores de performance (PI) e os eventos-alvo (*Prediction Goal* (PG)) utilizados, que são traduzidos e detalhados na Tabela III.2. Para facilitar o entendimento, eles estão numerados e os PIs de 15 a 28 correspondem à diferença entre os valores individuais das duas equipes.

No que diz respeito ao ranqueamento dos indicadores, *PIPassBox* e *PIPass3rd*, variantes de passes para o terço final do campo, figuram consistentemente entre os nove indicadores mais preditivos para todos os eventos-alvo avaliados, ao lado de *PIDangerousity* e de *PIEntr3rd*. Em contrapartida,

indicadores baseados em eventos raros, como gols e escanteios isolados, demonstraram baixo poder preditivo individual. Os autores também evidenciam que utilizar a diferença entre os valores do PI das duas equipes, em vez do valor absoluto de uma equipe, eleva substancialmente as correlações em 11 dos 14 indicadores avaliados, reforçando a importância de representar o desempenho relativo entre os times. Notavelmente, o trabalho utiliza dados abertos da StatsBomb (*StatsBomb Open Data*) como fonte de eventos, a mesma base empregada no presente trabalho.

Como aplicação prática, os autores propõem uma métrica de *match momentum* baseada na diferença das predições de gol das duas equipes ao longo da partida. Essa métrica, por se atualizar continuamente e apresentar menor volatilidade do que indicadores esparsos como *Dangerousity*, oferece um sinal interpretável do desequilíbrio momentâneo entre as equipes, conceito central também na abordagem aqui adotada.

O presente trabalho guarda múltiplos paralelos com [Lang et al., 2025], ao mesmo tempo em que se diferencia em pontos fundamentais. Primeiramente, ambos utilizam janelas temporais de dados passados para antecipar gols e adotam o MCC como critério de avaliação. Em segundo lugar, a técnica de *In-Play Prediction Masking* guarda semelhança conceitual com a janela de avaliação *SoftED* aqui proposta, que delimita um intervalo $[t - K, t]$ anterior ao gol para pontuar os alarmes do detector, evitando que sinais imediatamente anteriores ao evento contaminem a avaliação. Adicionalmente, os resultados de Lang et al. [2025] corroboram empiricamente a escolha de passes como variável de entrada para o detector de *drift*: os indicadores de passes para o terço final consistentemente figuram entre os mais preditivos para eventos de gol. Por fim, a sugestão de que janelas de 5 minutos superam janelas de 15 minutos referenda a escolha conservadora de $K = 10$ minutos adotada neste trabalho. A principal diferença reside na formulação do problema: enquanto Lang et al. [2025] treinam modelos supervisionados com rótulos históricos de eventos e realizam divisão explícita em conjuntos de treino, validação e teste (60:20:20), a abordagem aqui proposta emprega detectores de *drift* estatístico de forma não supervisionada sobre a série temporal de passes, identificando mudanças no padrão de jogo sem depender de dados rotulados. A robustez dos hiperparâmetros selecionados é verificada por divisão temporal (190 partidas de treino e 190 de teste), em contraste com a divisão explícita em treino, validação e teste (60:20:20) adotada por Lang et al. [2025]. Uma diferença estrutural adicional distingue as duas abordagens: Lang et al. [2025] utilizam sistematicamente a diferença entre os valores dos indicadores das duas equipes (PIs de 15 a 28) como variáveis de entrada, capturando o desequilíbrio relativo entre os times ao longo da partida. O presente trabalho, por outro lado, monitora a variação no padrão de passes de cada time individualmente, sem referência direta ao adversário. Essa distinção é relevante: a abordagem de Lang et al. captura o *momentum* relativo entre as equipes, enquanto a detecção de *drift* aqui proposta identifica rupturas no comportamento de um time específico, independentemente do que o

adversário está fazendo. As duas perspectivas são complementares e sua combinação constitui uma direção natural para trabalhos futuros.

Tabela III.1: Comparação entre os artigos relacionados e a relevância ao tema.

Artigo	**Objetivo**	**Técnica**	**Dados**
Arntzen and Hvattum [2021]	Predição de Resultado	Elo Rating (times) + Plus-Minus (jogadores) com Competing Risk e Regressão Logística Ordenada	• Campeonatos: Champions League, Championship, League One, League Two e English League Cup (2009/2010 - 2018/2019)
Dutta et al. [2024]	Predição do *timing* do gol	Análise de sobrevivência	• Campeonatos: Indian Super League (2022/2023)
Dobson and Goddard [2017]	Probabilidade de Gols	Simulação de Monte Carlo com Hazard Functions	• Campeonatos: Premier League e Football League (2001/2002 - 2008/2009)
Robberechts et al. [2019]	Predição de Probabilidade de Vitória	Modelo Bayesiano	• Campeonatos: Premier League, La Liga, Bundesliga, Italian Serie A, Ligue 1, Eredivisie, Belgian Serie A (2014/2015 - 2016/2017)
Yao et al. [2022]	Predição de Resultado	Modelo Bernoulli por frame e Simulação de Monte Carlo	• Campeonatos: Chinese Super League (2012 - 2018)
Capobianco et al. [2019]	Predição de Resultado	Random Forest	• Campeonatos: Italian Serie A (2017/2018)
Lang et al. [2025]	Mudança de Desempenho (*match momentum*)	Random Forest, Gradient Boosting / XGBoost, SVM, kNN e Regressão Logística	• Campeonatos: Bundesliga (2017/2018)
Goka et al. [2024]	Previsão do Próximo Evento	Masked Modeling com análise temporal bidirecional	• Campeonatos: não informado
Mendes-Neves et al. [2024]	Modelo Generativo	Deep Learning para simulação e predição de eventos (Large Events Model – LEM)	• Campeonatos: Premier League, La Liga, Ligue 1, Bundesliga e Italian Serie A (2017/2018)
Mendes-Neves et al. [2026]	Modelo Generativo	Modelo autorregressivo tabular escalável (LEM unificado)	• Campeonatos: Primeira Liga, La Liga, Bundesliga, Ligue 1, Danish Superliga, Jupiler Pro League
Yeung et al. [2025]	Previsão do Próximo Evento	Deep Learning e Reinforcement Learning	• Campeonatos: La Liga (2017/2018, 2023/2024), Premier League, Ligue 1, Italian Serie A e Bundesliga (2017/2018)
Simpson et al. [2022]	Previsão do Próximo Evento	Modelo sequencial (RNN/Transformer – Seq2Event)	• Campeonatos: Premier League, La Liga, Ligue 1, Bundesliga e Italian Serie A (2010/2011 - 2016/2017)
Alves [2025]	Previsão do Próximo Evento	Rede neural convolucional (CNN)	• Campeonatos: Copa do Mundo (2018), Euro (2016), Premier League, La Liga, Ligue 1, Bundesliga, Italian Serie A (2017/2018)
Umemoto et al. [2025]	Avaliação Defensiva por Predição de Múltiplos Eventos (GVDEP)	XGBoost + ponderação de probabilidades previstas por VAEP	• Campeonatos: Euro (2020)
Romero et al. [2026]	Previsão do Próximo Evento (tipo de ação)	MLP com embeddings de jogadores e times; XGBoost	• Campeonatos: Premier League (2017/2018)
Lühr and Lazarescu [2007]	Rastreamento de *concept drift* em performance de jogadores	Competing Windows Algorithm (CWA) + ensemble C4.5	• Campeonatos: não informado

Tabela III.2: Indicadores de performance (PIs) e eventos-alvo (PGs). Traduzido de Lang et al. [2025].

Nº	**Abreviação**	**Definição**
a) Indicadores de Performance (PI)		
1	PI_{Corner}	Número de escanteios
2	$PI_{EntrBox}$	Número de entradas com posse de bola na área adversária
3	$PI_{Entr3rd}$	Número de entradas com posse de bola no terço ofensivo
4	PI_{Goal}	Número de gols marcados
5	PI_{Shot}	Número de tentativas de chute
6	PI_{Cross}	Número de cruzamentos
7	$PI_{TackWon}$	Número de desarmes vencidos
8	$PI_{PassBox}$	Número de passes bem-sucedidos dentro ou para a área adversária
9	$PI_{Pass3rd}$	Número de passes bem-sucedidos dentro ou para o terço ofensivo
10	PI_{BP}	Tempo de posse de bola
11	PI_{BPBox}	Tempo de posse de bola na área adversária
12	PI_{BP3rd}	Tempo de posse de bola no terço ofensivo
13	$PI_{OutpOpp}$	Número de adversários superados por passes bem-sucedidos
14	PI_{Danger}	Probabilidade de gol a cada momento
15–28	PI_{PI_diff}	Diferença dos valores do PI entre as equipes (Time − Adversário)
b) Eventos-alvo (PG)		
1	PG_{isGoal}	Ocorrência de um gol pela equipe
2	PG_{isShot}	Ocorrência de um chute pela equipe
3	$PG_{isCorner}$	Ocorrência de um escanteio pela equipe
4	$PG_{isEntrBox}$	Ocorrência de uma entrada na área adversária pela equipe
5	$PG_{isEntr3rd}$	Ocorrência de uma entrada no terço ofensivo pela equipe

Capítulo IV Metodologia

A metodologia desta dissertação compreende quatro etapas: (1) coleta e pré-processamento de dados, (2) análise exploratória, (3) pipeline de detecção de *concept drift* e (4) protocolo de avaliação, descritas respectivamente nas Seções IV.1, IV.2, IV.3, IV.4 e exemplificadas na Figura IV.1.

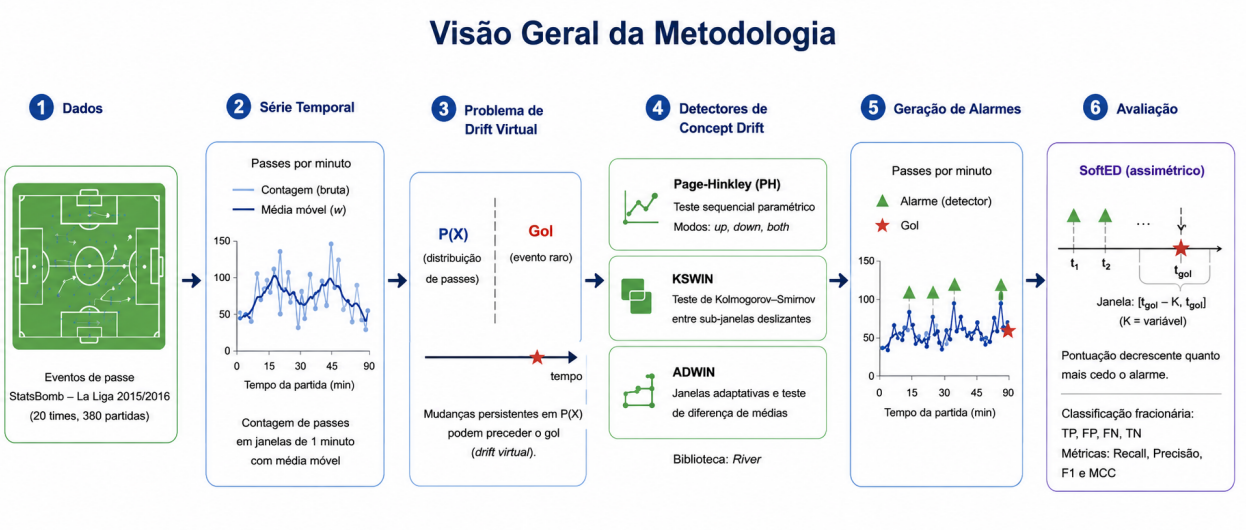

Figura IV.1: Visão geral da metodologia adotada.

IV.1 Coleta e Pré-processamento de Dados

Para a pesquisa, foram coletados dados de todas as partidas da temporada 2015/2016 do campeonato profissional de futebol espanhol La Liga, envolvendo os 20 clubes da primeira divisão ao longo de 38 rodadas, totalizando 380 jogos e aproximadamente 1,3 milhão de registros de eventos. Os dados foram obtidos a partir do conjunto público de análise de futebol disponibilizado pela StatsBomb[1], que registra eventos ocorridos durante as partidas com granularidade temporal por ação. Essa competição e temporada foram escolhidas pela cobertura completa da temporada, nível de detalhe dos eventos e disponibilidade pública dos dados. As escolhas metodológicas descritas a seguir (seleção de *features*, arquitetura do *pipeline* e protocolo de avaliação) são orientadas diretamente pela pergunta de pesquisa: avaliar se detectores não supervisionados conseguem antecipar gols a partir de mudanças no padrão de passes.

[1]Disponível em: `https://github.com/statsbomb/open-data`

Para criar a base a ser utilizada na seção IV.2, as partidas foram agregadas em janelas de 1 minuto por time, produzindo contagens dos eventos táticos. Cada linha contém 92 atributos, incluindo identificação da partida (*id* da partida, time mandante, time visitante), marcadores de tempo (data da partida, minutagem, primeiro ou segundo tempo), eventos táticos (passes, chutes, dribles, defesas) e qualificadores dos eventos (resultado do chute, por exemplo, que pode ser salvo pelo goleiro, na trave, pra fora etc).

Após a agregação inicial, foram aplicadas transformações metodológicas para viabilizar a análise e a implementação dos detectores de *concept drift*. Gols de bola parada (faltas e escanteios) e pênaltis foram excluídos por duas razões complementares. Primeiro, essas situações interrompem o fluxo natural do jogo: o árbitro para a partida, as equipes se reposicionam e o ritmo de passes é artificialmente suspenso, o que introduz uma ruptura na série temporal não relacionada ao padrão tático que o detector monitora. Segundo, a conversão dessas jogadas em gol depende de fatores específicos (posicionamento em bola parada, especialização do cobrador e distância da falta), independentes da dinâmica de passes dos minutos anteriores. Incluir esses eventos como alvos de detecção contaminaria o sinal com situações em que o detector não teria poder preditivo por construção. Essa separação é consistente com a prática da literatura: trabalhos como Mendes-Neves et al. [2024], Simpson et al. [2022] e Yao et al. [2022] tratam pênaltis, faltas e escanteios como categorias de eventos distintas das ações em jogo aberto, reconhecendo que seus desfechos seguem dinâmicas próprias.

Completando o tratamento de dados, os minutos da partida foram tratados como relativos a cada período do jogo em vez de tempo corrido (ou seja, o primeiro tempo vai de 0 a 45+ minutos e o segundo reinicia do zero, em lugar de continuar de 45 a 90+), o que permite comparação direta entre os dois tempos e evita que o segundo período herde a média acumulada do primeiro ao alimentar os detectores (para evitar vazamento temporal dos dados). Cada observação também foi rotulada com o contexto do time (*mandante* ou *visitante*) e com a perspectiva da jogada (*ataque* ou *defesa*), possibilitando análises segmentadas.

O resultado desse pré-processamento é uma tabela no formato largo (*wide*) com aproximadamente 36 mil linhas, em que cada linha representa um minuto de uma partida, contendo as contagens de eventos dos dois times em colunas separadas. Essa estrutura serviu como entrada tanto para a análise exploratória quanto para o *pipeline* de detecção de *concept drift*. Para a detecção de *drift*, foram selecionadas três variáveis de volume de passes: total de passes, passes certos e passes errados, por concentrarem o maior volume de eventos táticos da partida, conforme detalhado na seção IV.2. Todos os códigos estão disponibilizados no Github[2].

[2]Disponível em: https://github.com/anavibiga/mestrado-drift-detection

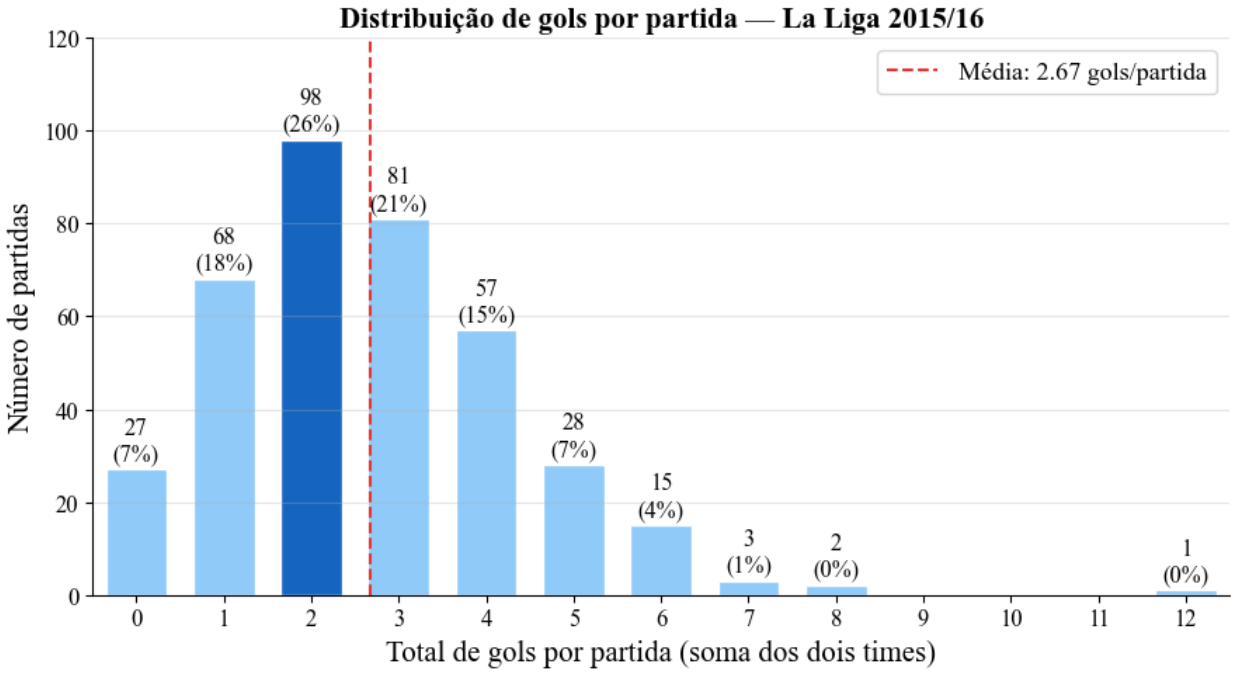

Figura IV.2: Histograma de gols por partida (soma dos dois times) na temporada 2015/16 da La Liga.

IV.2 Análise Exploratória de Dados

A análise exploratória dos dados teve como motivação: (1) verificar se existe sinal nos dados que justifique a abordagem de *concept drift* e (2) identificar quais *features* fazem sentido monitorar. Para tanto, foram feitas análises iniciais com volumetria dos dados e, posteriormente, análises com contextos como ataque e defesa, times mandantes ou visitantes, e avaliação das métricas até afunilar para as mais relevantes para os detectores. Ao todo, o dataset apresenta 27 tipos distintos de eventos; a análise de volume de ocorrências orientou a seleção daqueles com maior presença e relevância tática na partida.

As primeiras análises sobre volumetria dos dados trouxeram que a temporada contou com 1.014 gols ao longo das 380 partidas, resultando em uma média de 2,67 gols por jogo. A distribuição entre os períodos mostrou leve predominância do segundo tempo, responsável por 54,6% dos gols marcados. Em relação ao volume por partida, 57,6% dos jogos tiveram ao menos 2 gols, enquanto 17,9% tiveram apenas 1 gol e 7,1% terminaram sem gols, conforme indicado na Figura IV.2.

Além da distribuição de volume, analisou-se o impacto do primeiro gol sobre o resultado final da partida. Das 353 partidas com ao menos um gol, o time que marcou primeiro venceu em 70,5% dos casos ($n = 249$), resultado estatisticamente significativo tanto pelo teste binomial unilateral ($p < 0{,}001$, *Intervalo de Confiança* (IC) 95%: [0,66; 1,00]) quanto pelo qui-quadrado de independência entre quem marcou primeiro e o resultado final ($\chi^2 = 14{,}41$, $gl = 2$, $p < 0{,}001$), conforme Figura IV.3. Esse resultado reforça a relevância de detectar antecipadamente mudanças no padrão

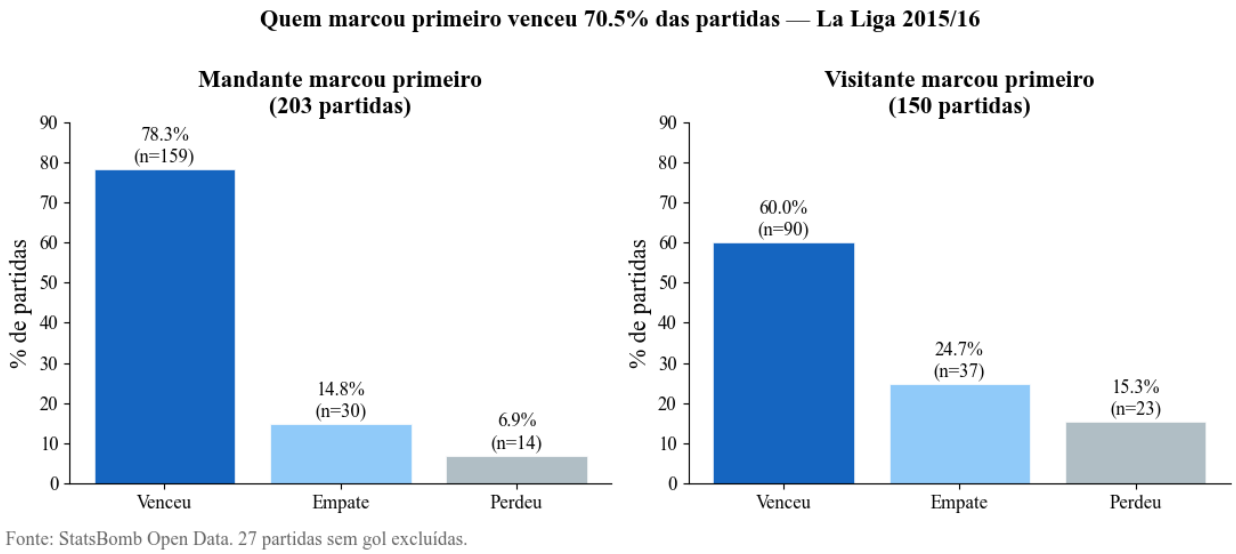

Figura IV.3: Proporção de vitórias, empates e derrotas conforme quem marcou primeiro (La Liga 2015/16, $n = 353$ partidas com gol).

de jogo que precedem o gol: além de ser um evento raro, o primeiro gol condiciona fortemente o desfecho da partida. Em média, o primeiro gol de cada período ocorreu aos 18 minutos (mediana de 16 minutos, desvio padrão de 13 minutos), o que indica que, na maioria das partidas, há janela temporal suficiente para o detector completar o aquecimento e operar na fase ativa antes da ocorrência do primeiro gol.

Considerando o contexto das partidas, 59,0% dos gols foram marcados pelos times mandantes, o que motivou a separação por contexto de jogo (mandante ou visitante) nas análises subsequentes. Em relação ao tipo de gol, 90,3% foram classificados como *open play*. Conforme descrito na seção IV.1, as análises de janelas temporais e os detectores de *drift* foram aplicados exclusivamente sobre essa categoria de gol.

Estabelecida a relevância do gol para o resultado da partida, a análise volta-se para os tipos de evento que compõem o contexto tático anterior a ele. Os 10 eventos com maior ocorrência na partida estão ilustrados na Figura IV.4. Foram excluídos eventos de controle de jogo (*Ball Receipt**, *Carry*) e marcadores de tempo. *Passe* representa qualquer passe entre jogadores do mesmo time (passes curtos, longos e cruzamentos) e corresponde a 53,6% de todos os eventos táticos da temporada, sendo o evento mais frequente por larga margem.[3] *Pressão* (16,8%) ocorre quando um jogador pressiona o adversário com a bola, tentando forçar erro ou recuperar a posse. *Recuperação de bola* (5,7%) registra recuperações de bola após disputa ou erro adversário, enquanto *Duelo* (4,6%) captura disputas físicas diretas, aéreas ou no chão, pela posse. Os demais eventos somados representam menos de 12% dos eventos táticos: *Afastamento defensivo*, *Drible*, *Bloqueio*, *Interceptação*, *Falta*

[3]Os nomes dos eventos foram traduzidos livremente do inglês com base na documentação oficial da StatsBomb: *Open Data Events v4.0.0* [StatsBomb, 2019].

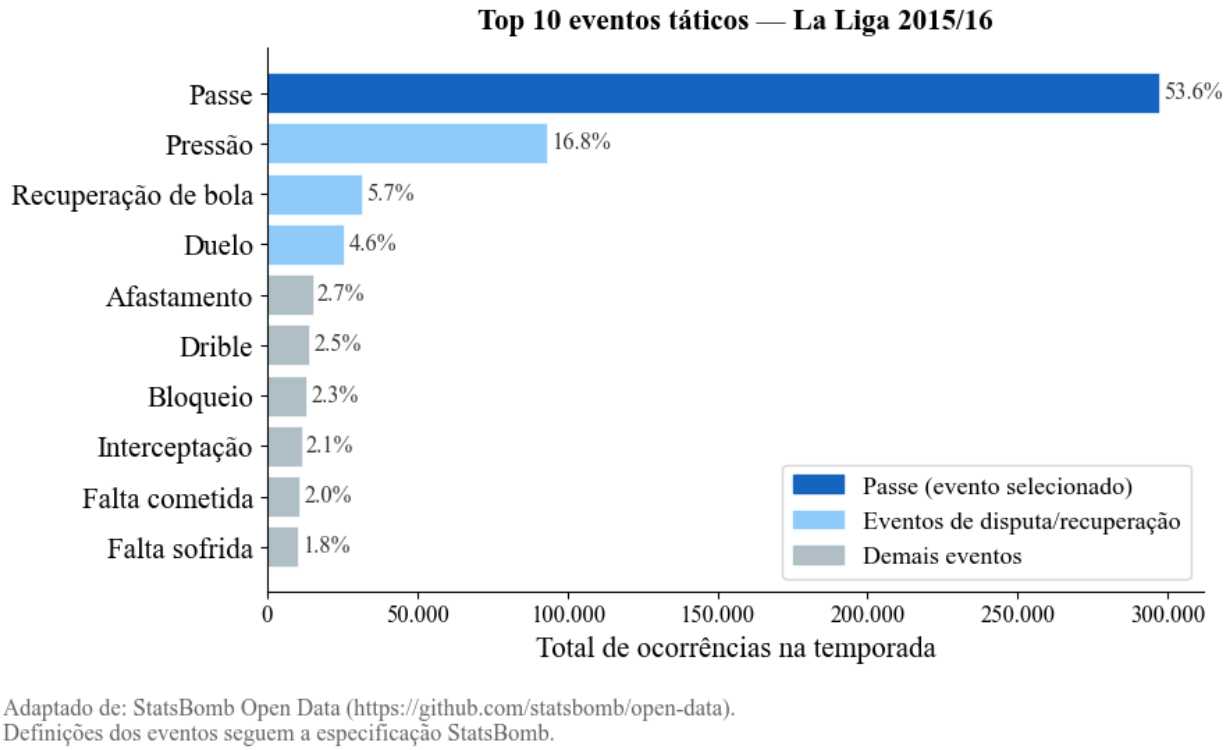

Figura IV.4: Os 10 eventos táticos com maior volume de ocorrências na temporada 2015/16 da La Liga.

cometida e *Falta sofrida*.

Para corroborar empiricamente a escolha do passe como *feature* do detector de *drift*, a Figura IV.5 apresenta os coeficientes de correlação de Pearson entre a contagem de cada tipo de evento e a ocorrência de gol nas janelas de $K = 5, 10, 15$ minutos anteriores ao evento, calculados sobre as 380 partidas da La Liga 2015/16. Para cada gol da temporada, foram contabilizados os eventos realizados pelo time marcador no intervalo $[t - K,\, t)$; janelas controle de mesmo tamanho, sem gol próximo, foram amostradas aleatoriamente para compor o grupo de comparação. Todos os tipos avaliados apresentaram correlação positiva e estatisticamente significativa ($p < 0{,}05$), indicando que partidas com maior volume de qualquer tipo de ação tendem a preceder gols com mais frequência.

Dentre os tipos avaliados, passes obtiveram o maior coeficiente de Pearson em todas as variações de janela ($r = 0{,}29$ para $K = 5$, $r = 0{,}30$ para $K = 10$, $r = 0{,}35$ para $K = 15$), resultado em linha com os achados de Lang et al. [2025], que identificaram indicadores baseados em passes entre os mais preditivos para eventos de gol na Bundesliga. A hierarquia dos demais tipos de evento, contudo, não é estável entre as janelas: para $K = 5$, pressão supera chute e drible em correlação com o gol, padrão que se inverte à medida que K aumenta, com chutes progressivamente mais correlacionados ($r = 0{,}17$ para $K = 5$, $r = 0{,}25$ para $K = 10$, $r = 0{,}27$ para $K = 15$). Isso sugere que ações de pressão refletem dinâmicas de curto prazo próximas ao gol, enquanto o acúmulo de

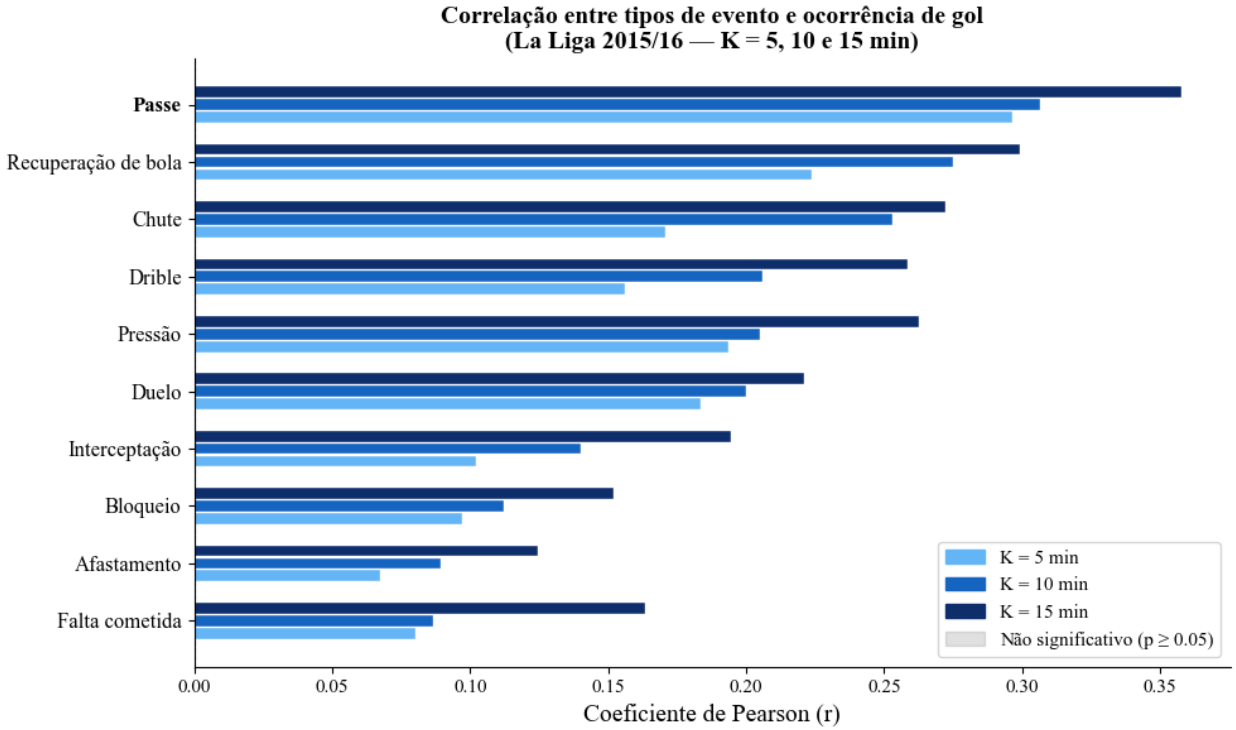

Figura IV.5: Correlação de Pearson entre tipos de evento e ocorrência de gol comparadas por janelas $K = 5$, $K = 10$ e $K = 15$ min (La Liga 2015/16, $n = 380$ partidas).

chutes se torna mais discriminativo em janelas mais amplas. Esse resultado sustenta a escolha de passes como *feature* de entrada: eventos com maior correlação com a ocorrência de gol e maior volume absoluto produzem séries temporais mais densas, favorecendo a estabilidade estatística dos detectores. Cabe ressaltar, contudo, que a correlação mensura associação entre volume médio e proximidade de gol, e não entre variação estrutural na série e ocorrência do evento. A validade do sinal de *drift* em si é avaliada empiricamente pelos resultados do Capítulo V, e a análise de variação nas janelas temporais a seguir oferece evidência exploratória complementar nessa direção.

Para verificar se existe mudança de volume de eventos nos minutos anteriores ao gol, foram construídas janelas temporais de agregação em três escalas: 5 minutos (*lag5* vs *lag5_ antes*), 10 minutos (*lag10* vs *lag10_ antes*) e 15 minutos (*lag15* vs *lag15_ antes*), em que *lag_ antes* representa a janela de igual duração imediatamente anterior, utilizada como *proxy* do ritmo normal de jogo. O filtro $t \geq 11$ foi aplicado para a janela de 5 minutos, $t \geq 21$ para a de 10 minutos e $t \geq 31$ para a de 15 minutos, garantindo que ambas as janelas (*lag* e *lag_ antes*) existam integralmente para cada gol analisado.

As Figuras IV.6, IV.7 e IV.8 apresentam a variação média de volume para as 10 métricas táticas mais frequentes nas duas escalas. Em todas, o passe é o evento com maior variação, mostrando, inclusive, em $K = 10$ uma diferença de variação de passes para ataque e defesa.

Para ilustrar o comportamento temporal dos dados, foram selecionadas partidas com critérios específicos: apenas gols ocorridos após o minuto 20 de cada período foram considerados, garantindo

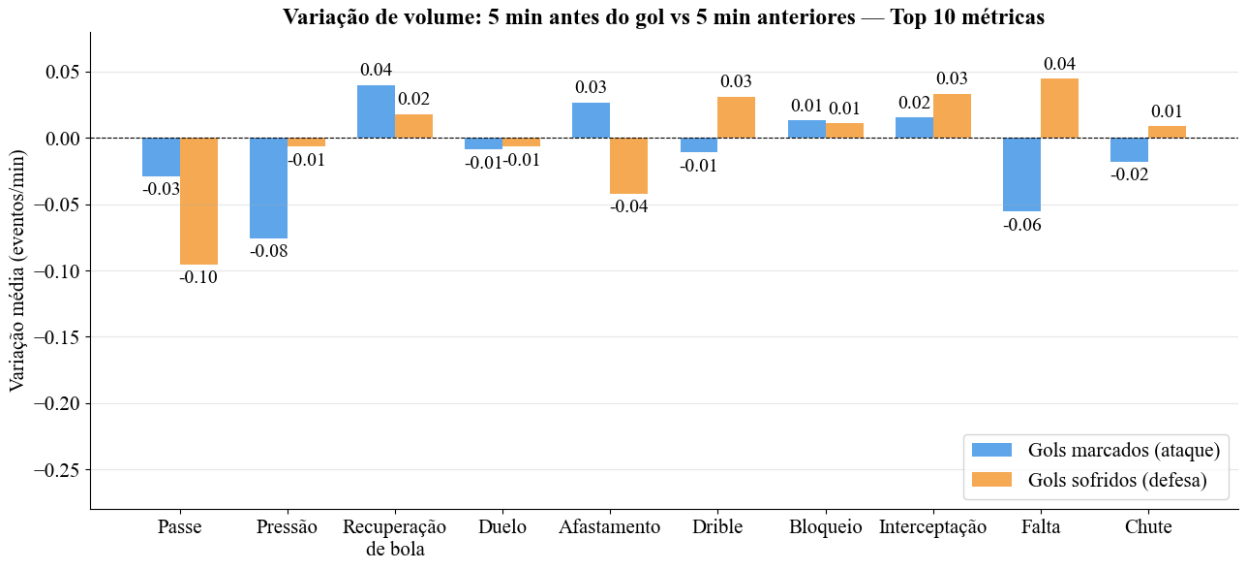

Figura IV.6: Variação de volume nos 5 min anteriores ao gol vs 5 min precedentes (La Liga 2015/16), por ataque e defesa.

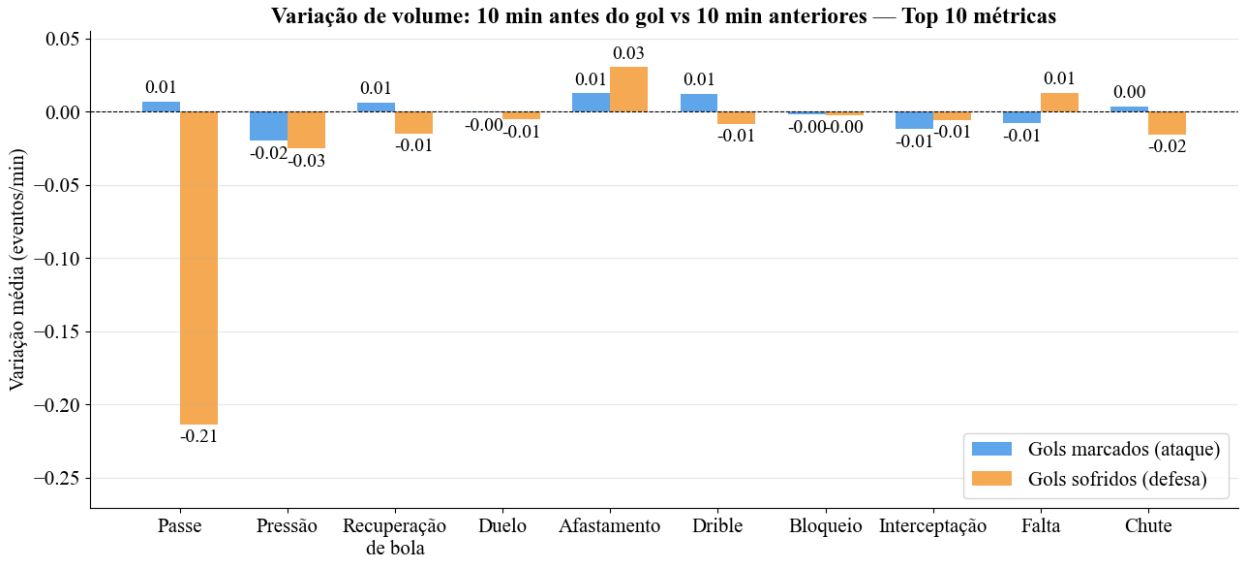

Figura IV.7: Variação de volume nos 10 min anteriores ao gol vs 10 min precedentes (La Liga 2015/16), por ataque e defesa.

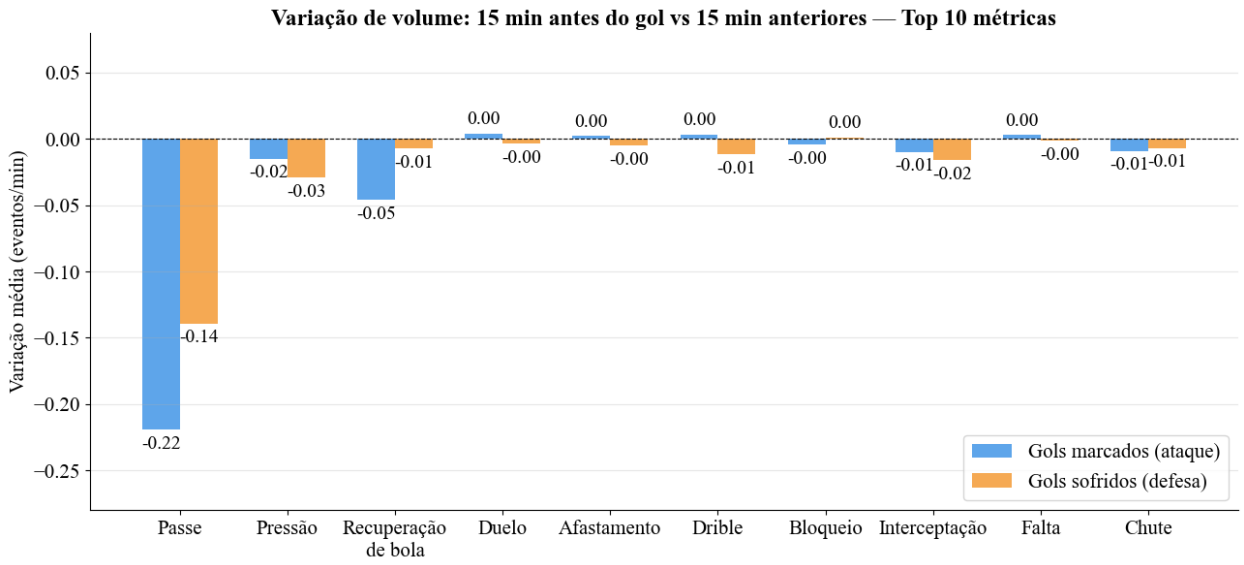

Figura IV.8: Variação de volume nos 15 min anteriores ao gol vs 15 min precedentes (La Liga 2015/16), por ataque e defesa.

que as duas janelas de análise estivessem completas; partidas com gols consecutivos em intervalo inferior a 20 minutos foram excluídas para evitar sobreposição de janelas. Nas figuras a seguir, a região azul representa *lag10* (10 minutos imediatamente anteriores ao gol), a região laranja representa *lag10_ antes* (10–20 minutos antes do gol), utilizada como referência do ritmo normal de jogo, a linha tracejada indica gol do time monitorado e a pontilhada gol do adversário; os valores anotados nas janelas correspondem à média de passes brutos por minuto calculada diretamente sobre o período sombreado.

Com esses critérios, foram selecionadas duas partidas com perfis contrastantes. Na partida entre Atlético de Madrid e Getafe (Figura IV.9), o Getafe registrou aumento no volume de passes nos 10 minutos anteriores a cada gol em relação à janela de referência. Já na partida entre Las Palmas e Espanyol (Figura IV.10), foi o Las Palmas quem apresentou esse aumento, com diferença especialmente pronunciada no segundo tempo (4,0 passes/min na janela de referência contra 10,7 passes/min na janela pré-gol). Em ambos os casos, a série suavizada por média móvel de 10 minutos evidencia variação no padrão de atividade nos minutos que antecedem o gol, motivando o uso de detectores de *drift* para identificar essas rupturas de forma sistemática.

As Figuras IV.11 e IV.10 ilustram as duas perspectivas complementares dos dados. A visualização com valores brutos por minuto (Figura IV.11) evidencia a natureza esparsa da série: mesmo passes, evento responsável por mais de 50% de todos os eventos táticos, apresentam minutos com contagem zero, reflexo de interrupções naturais do jogo ou falta de posse de bola do time. As barras representam a contagem de passes em cada minuto e as linhas horizontais a média por janela. Essa esparsidade motivou o uso de média móvel de janela W como pré-processamento antes de alimentar os detectores(Figura IV.10), reduzindo ruído pontual e tornando o sinal mais estável para a detecção

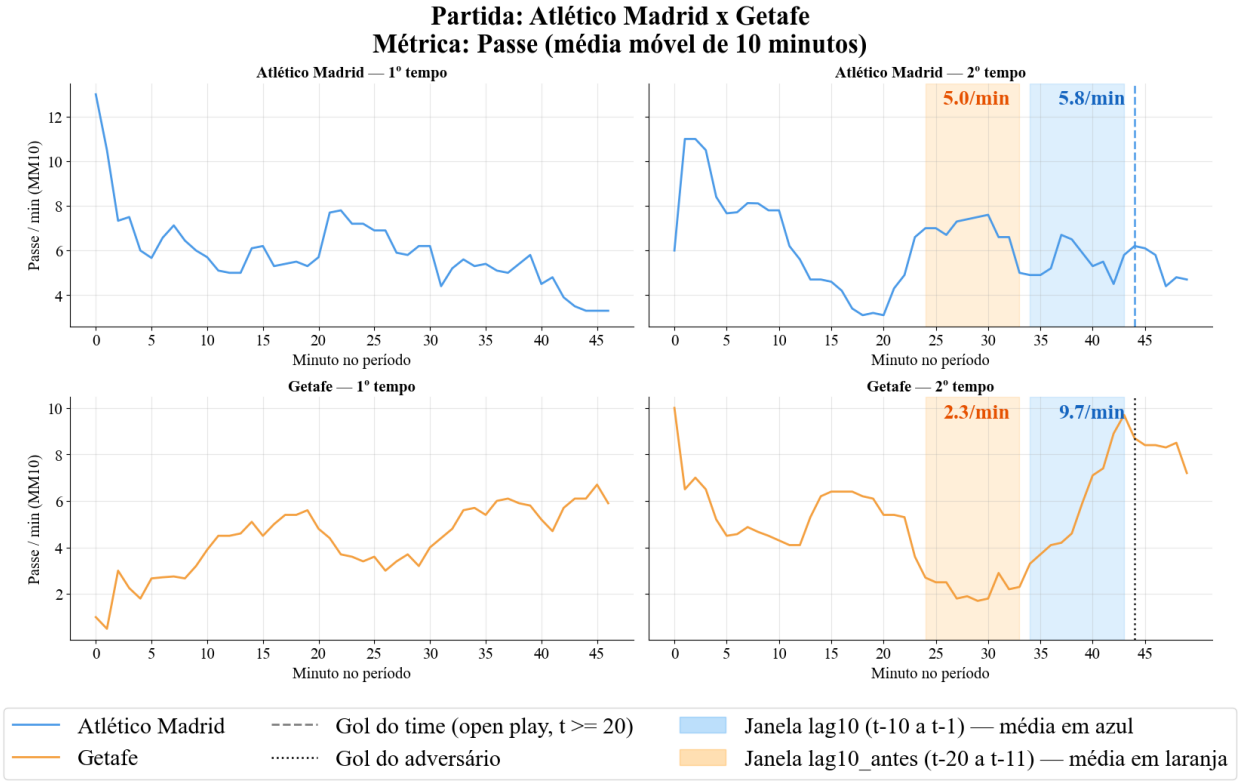

Figura IV.9: Passes/min (média móvel 10 min) — Atlético de Madrid 2×0 Getafe (La Liga 2015/16).

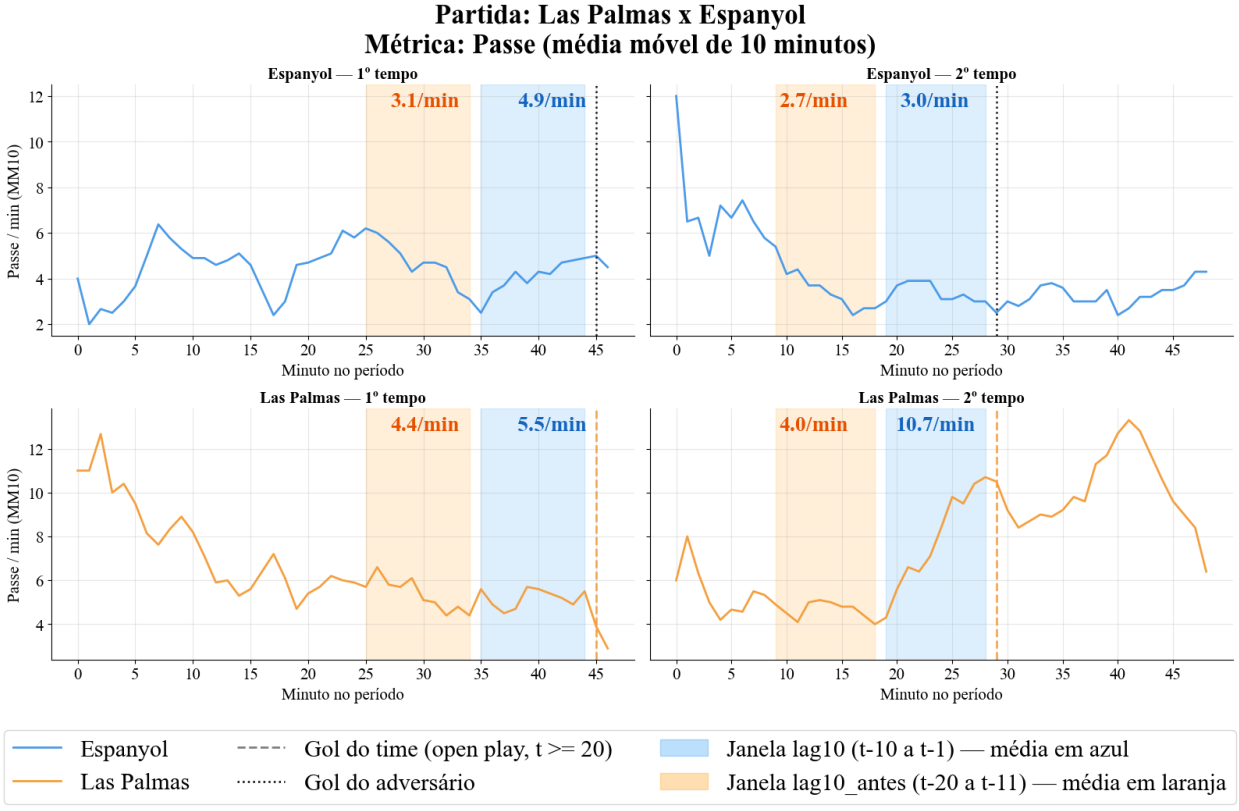

Figura IV.10: Passes/min (média móvel 10 min) — Las Palmas 4×0 Espanyol (La Liga 2015/16).

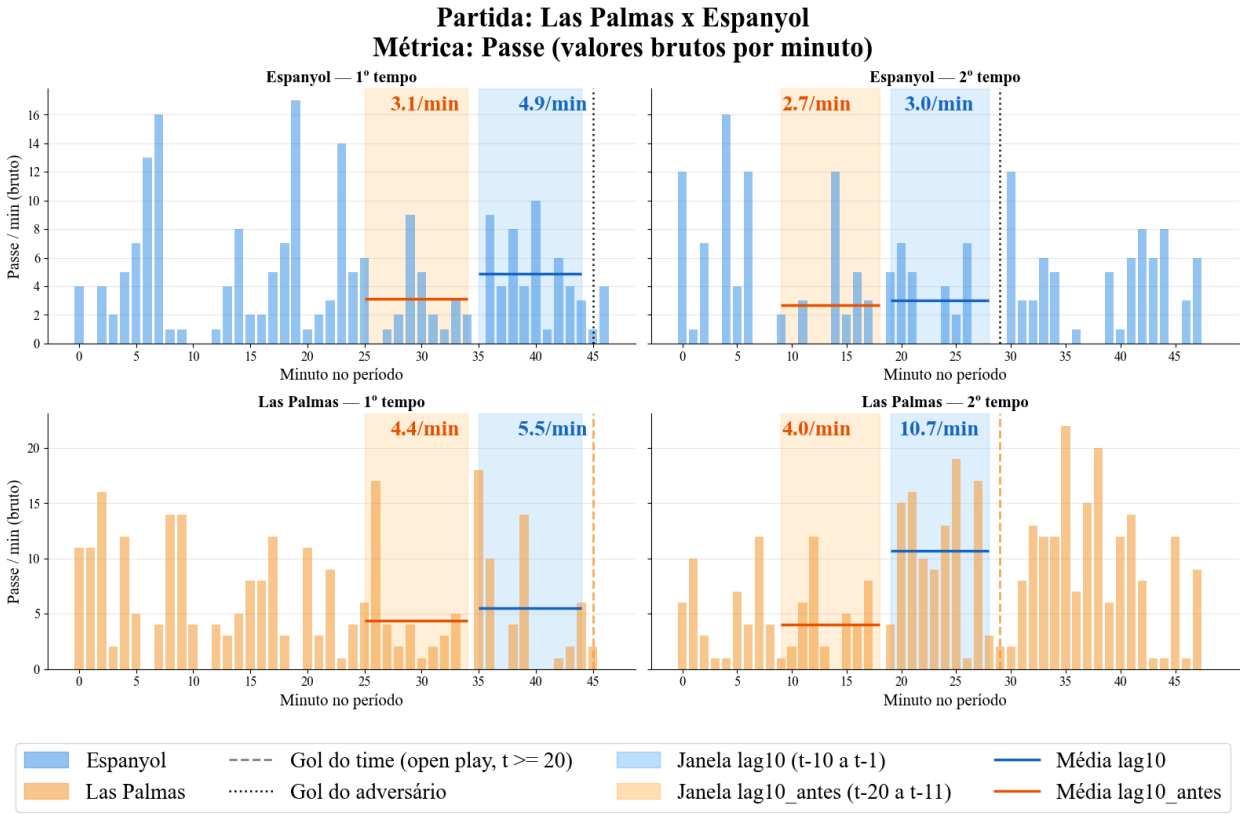

Figura IV.11: Passes brutos/min — Las Palmas 4×0 Espanyol (La Liga 2015/16).

de mudanças estruturais.

Os padrões observados nas análises exploratórias, variação no volume de passes antes dos gols e ruptura na série temporal, motivam a aplicação formal de detectores de *concept drift*. Na seção IV.3, descreve-se o *pipeline* de detecção implementado, que avalia três algoritmos distintos: ADWIN, Page-Hinkley e KSWIN. Para contextualizar os resultados, cada detector é comparado a dois *baselines*: um *baseline* de alarme fixo, que sinaliza *drift* em intervalos regulares predefinidos, e um *baseline* aleatório, que serve como limite inferior de desempenho. A combinação entre detectores e *baselines* permite avaliar se os algoritmos capturam rupturas estatisticamente relevantes além do que seria esperado por acaso ou por um critério puramente temporal.

IV.3 Pipeline de Detecção de Concept Drift

A detecção de *drift* é enquadrada como um problema de monitoramento de séries temporais: para cada partida, a sequência de contagens de passes por minuto de cada time é tratada como um fluxo de dados (*data stream*). Nesse contexto, *drift* corresponde a uma mudança estatística no volume de passes ao longo do tempo, ou seja, *drift* virtual. O objetivo do pipeline é emitir um alarme quando essa ruptura é detectada, sinalizando uma potencial mudança de ritmo de jogo que pode preceder um gol.

Antes de alimentar o detector, a série bruta de passes por minuto é suavizada por uma média

móvel de janela W. O propósito é reduzir o ruído minuto-a-minuto decorrente da esparsidade natural dos dados, tornando o sinal mais estável para a detecção de mudanças estruturais. Os primeiros W minutos de cada período constituem o período de aquecimento (*warmup*): nesse intervalo, a média móvel ainda não dispõe de W observações completas para produzir um valor estável, de modo que o detector não é alimentado para evitar alarmes artificiais causados pela inicialização da série e esses minutos não são utilizados para avaliação. Assim, W desempenha função dupla: define simultaneamente o tamanho da janela de suavização e a duração do aquecimento, garantindo que o detector receba apenas valores a partir do momento que a média móvel está integralmente calculada.

Após o período de *warmup*, os detectores podem começar a avaliar se existe ou não variação na série temporal e, caso seja detectado *drift*, um alarme é disparado. Após cada alarme emitido, a emissão de alarmes é suprimida por C minutos, período denominado *cooldown*. Essa decisão visa evitar rajadas de alarmes consecutivos para o mesmo evento: uma única mudança de ritmo pode gerar múltiplos disparos em minutos seguidos, o que dificultaria a avaliação e superestimaria a sensibilidade do detector. Com o *cooldown*, cada alarme é tratado como um evento isolado, tornando a análise mais interpretável.

Para a avaliação dos alarmes, utiliza-se uma variante do método *SoftED* [Salles et al., 2024] com janela de tolerância de K minutos. Esse método reconhece que exigir coincidência exata entre o alarme e o gol seria uma métrica excessivamente rígida: um alarme emitido poucos minutos antes do gol carrega valor preditivo relevante, enquanto um alarme distante no tempo não.

No *SoftED* original, a janela de avaliação é simétrica em torno do instante do evento t, cobrindo o intervalo $[t - K,\, t + K]$ com pontuação máxima no instante exato t. Essa formulação atribui crédito tanto a alarmes antecipados quanto a alarmes tardios.

No contexto de predição de gols, contudo, apenas alarmes emitidos *antes* do evento têm valor operacional, dado que um aviso após o gol não permite nenhuma intervenção tática. Por isso, adota-se uma variante assimétrica: o gol é deslocado K minutos para trás, tornando $t - K$ o novo ponto de referência, e a janela de avaliação é restrita ao intervalo $[t - K,\, t]$. A pontuação máxima (1,0) é atribuída a alarmes emitidos em $t - K$, isto é, com K minutos de antecedência, e decresce linearmente até zero no instante t, penalizando alarmes tardios e ignorando completamente detecções posteriores ao gol. Formalmente, o *score* atribuído a um alarme emitido no instante a é:

$$\text{score}(a,\, t) = 1 - \frac{a - (t - K)}{K}, \quad a \in [t - K,\, t] \tag{IV.1}$$

Uma consequência direta da Equação IV.1 é que o peso atribuído a um alarme depende não apenas de sua antecedência em relação ao gol $\Delta a = t - a$, como também do tamanho da janela

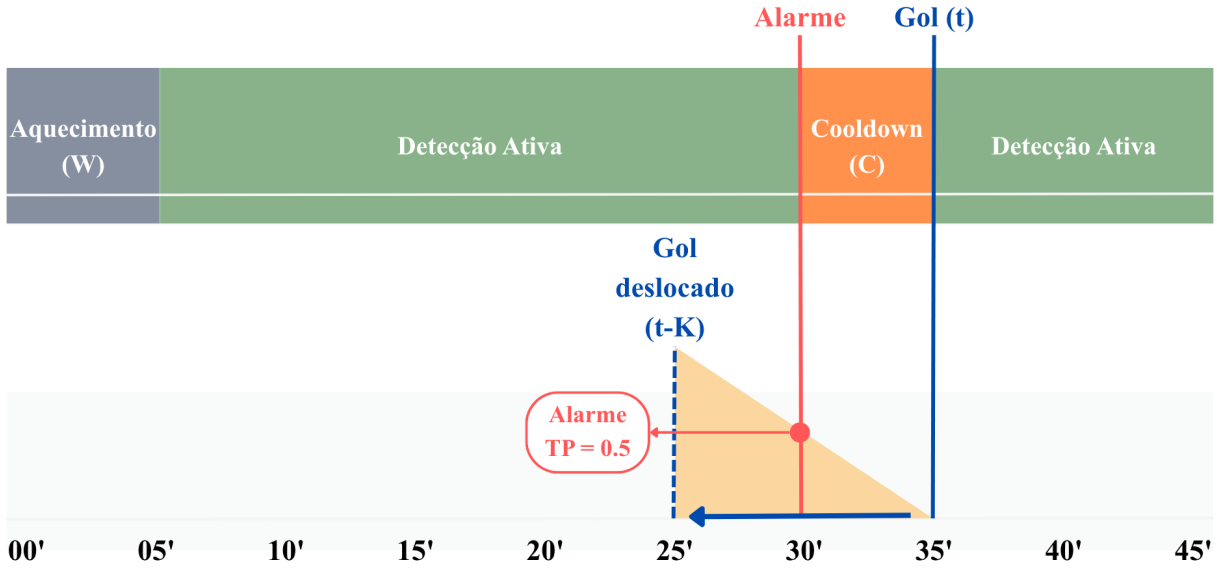

Figura IV.12: Pipeline de detecção de *drift*.

de avaliação K. Janelas maiores penalizam proporcionalmente mais os alarmes emitidos distantes do instante $t - K$. Por exemplo, em uma janela de avaliação $K = 5$, um alarme com $\Delta a = 4$ minutos de antecedência recebe pontuação 0,80. Já em uma janela de $K = 10$, 4 minutos de antecedência representam 0,40 e, para $K = 15$, representam 0,27. Assim, comparações de *recall* entre configurações com valores distintos de K devem ser interpretadas com cautela: pontuações mais baixas não indicam necessariamente pior desempenho do detector, mas refletem a penalização mais severa imposta por janelas mais amplas.

A Figura IV.12 ilustra o funcionamento completo do pipeline de detecção de *drift* adotado neste trabalho. A estrutura temporal está organizada em três fases: os primeiros W minutos de cada período correspondem ao aquecimento (*warmup*), durante o qual o detector recebe as observações iniciais sem emitir alarmes; em seguida, o detector entra em fase ativa, monitorando continuamente o fluxo de dados; após cada alarme emitido, inicia-se um período de *cooldown* de C minutos, durante o qual novos alarmes são suprimidos. A figura também detalha a pontuação atribuída pela variante assimétrica do *SoftED*: dado um gol no instante t, a janela de avaliação cobre o intervalo $[t - K, t]$, com pontuação máxima (1,0) atribuída a alarmes emitidos exatamente em $t - K$ e decrescendo linearmente até zero em t. Alarmes anteriores a $t - K$ ou posteriores a t são contabilizados como falsos positivos e não recebem crédito.

Para a definição dos parâmetros W, C e K, foram adotados os seguintes critérios:

Definição de K (janela de avaliação):

$$K_{\min} \leq \bar{t}_{gol} - \sigma_{t_{gol}} \tag{IV.2}$$

$$K_{\max} \leq \bar{t}_{gol} - W \tag{IV.3}$$

onde $\bar{t}_{gol}$ é a média do tempo até o primeiro gol e $\sigma_{t_{gol}}$ é o respectivo desvio padrão. O limite inferior garante que K não ultrapasse o intervalo típico de ocorrência do primeiro gol; o limite superior garante que a janela de avaliação não se sobreponha ao período de aquecimento.

Definição de W (janela de média móvel): W representa o tamanho da janela de suavização e deve ser suficientemente longo para identificar tendências, filtrar variações aleatórias e flutuações de curto prazo, mas suficientemente curto para não mascarar mudanças estruturais. Como referência, considerou-se que cada período da partida tem duração aproximada de 45 minutos; valores de W correspondentes a aproximadamente 10% ($W = 5$ min) e 5% ($W = 3$ min) desse intervalo foram testados inicialmente.

Definição de C (janela de inativação dos alarmes): C representa o tamanho da janela após alarme que não são contabilizados os alarmes. Para evitar que períodos importantes da partida deixem de ser analisados,

$$C \leq K \tag{IV.4}$$

Além disso, caso haja um gol enquanto estiver no cooldown, o cooldown é zerado para que possa voltar a emitir alarmes.

Cálculo de TP, FP, FN e TN: Dado um conjunto de alarmes $\mathcal{A} = \{a_1, \ldots, a_n\}$ e de gols $\mathcal{G} = \{g_1, \ldots, g_m\}$, define-se a função de pontuação meia-pirâmide como:

$$\text{score}(a,\, g) = \begin{cases} 1 - \dfrac{a - (g - K)}{K} & \text{se } g - K \leq a \leq g \\ 0 & \text{caso contrário} \end{cases}$$

Após o pareamento ótimo via algoritmo húngaro, cada alarme $a \in \mathcal{A}$ recebe uma pontuação

$\hat{s}(a) \in [0, 1]$, valendo 0 quando não associado a nenhum gol. As métricas são então calculadas como:

$$\mathrm{TP} = \sum_{a \in \mathcal{A}} \hat{s}(a) \tag{IV.5}$$

$$\mathrm{FP} = \sum_{a \in \mathcal{A}} \big(1 - \hat{s}(a)\big) \tag{IV.6}$$

$$\mathrm{FN} = |\mathcal{G}| - \mathrm{TP} \tag{IV.7}$$

$$\mathrm{TN} = n_{\mathrm{ef}} - |\mathcal{G}| - \mathrm{FP} \tag{IV.8}$$

onde $n_{\mathrm{ef}} = n_{\mathrm{total}} - W$ são os minutos líquidos da partida, excluindo o período de aquecimento.

Para partidas sem alarmes ($\mathcal{A} = \emptyset$) ou sem gols ($\mathcal{G} = \emptyset$):

$$\mathrm{TP} = 0, \quad \mathrm{FP} = |\mathcal{A}|, \quad \mathrm{FN} = |\mathcal{G}|, \quad \mathrm{TN} = n_{\mathrm{ef}} - |\mathcal{G}| - |\mathcal{A}| \tag{IV.9}$$

As métricas SoftED preservam as propriedades das métricas tradicionais (*hard*) ao impor duas restrições sobre a pontuação:

1. **Unicidade por alarme:** cada alarme $a \in \mathcal{A}$ recebe uma única pontuação $\hat{s}(a) \in [0, 1]$.
2. **Limite por evento:** a pontuação total atribuída a um mesmo gol $g \in \mathcal{G}$ não pode ultrapassar 1, ou seja, $\sum_{a \in \mathcal{A}} \hat{s}(a, g) \leq 1$.

A restrição (1) implica que a contribuição de cada alarme é inteiramente alocada entre detecção correta e alarme falso:

$$\hat{s}(a) + \big(1 - \hat{s}(a)\big) = 1, \quad \forall\, a \in \mathcal{A}$$

de modo que, somando sobre todos os alarmes:

$$\mathrm{TP} + \mathrm{FP} = |\mathcal{A}|$$

A restrição (2) é garantida pelo pareamento via algoritmo húngaro, que associa no máximo um alarme a cada gol. Analogamente, como $\mathrm{FN} = |\mathcal{G}| - \mathrm{TP}$:

$$\mathrm{TP} + \mathrm{FN} = |\mathcal{G}|$$

Em conjunto, as quatro métricas particionam os n_{ef} minutos efetivos da partida:

$$\mathrm{TP} + \mathrm{FP} + \mathrm{FN} + \mathrm{TN} = n_{\mathrm{ef}}$$

Dessa forma, também fica garantido que não exista nenhum cálculo com FP, TP, FN ou TN < 0.

Três limitações do pipeline e da avaliação merecem registro. Primeiro, gols ocorridos antes do minuto W do período resultam sempre em FN, pois o detector ainda está no período de aquecimento e nenhum alarme pode ser emitido. Para gols entre o minuto W e K, o detector pode emitir alarme, mas a janela $[t - K,\, t]$ fica parcialmente truncada pelo início do período, reduzindo o score máximo alcançável. O menor W utilizado é $W = 3$; apenas 46 gols (5,0%) ocorreram antes do minuto 3 do período, afetando diretamente o *warmup*. Dos 916 gols *open play*, 86 (9,4%) ocorrem antes do minuto 5, 196 (21,4%) antes do minuto 10 e 289 (31,6%) antes do minuto 15, sendo que a severidade da limitação varia conforme os valores de K e W selecionados.

Segundo, quando dois gols consecutivos ocorrem no mesmo período com intervalo inferior a K minutos, as janelas de avaliação se sobrepõem, exigindo dois alarmes distintos nesse intervalo para que ambos os gols sejam cobertos, o que reduz a probabilidade de ambos receberem crédito parcial. Na base utilizada, identificaram-se 381 pares de gols consecutivos no mesmo período; desses, 75 (19,7%) apresentam intervalo inferior a $K = 5$ minutos, 170 (44,6%) inferior a $K = 10$ minutos e 241 (63,3%) inferior a $K = 15$ minutos.

Por fim, o parâmetro W introduz uma latência inerente ao pipeline: a série suavizada no minuto t representa a média dos últimos W minutos, de modo que uma mudança real no ritmo de jogo só se torna visível ao detector após aproximadamente $W/2$ minutos. A isso se soma o tempo de acumulação do próprio detector: o Page-Hinkley, por exemplo, só dispara quando o desvio cumulativo da média ultrapassa o limiar λ, o que pode levar mais alguns minutos dependendo da magnitude da mudança. A latência total típica do pipeline situa-se entre $W/2$ e W minutos após a mudança real. Como consequência, alarmes raramente atingem pontuação máxima (score = 1,0) na avaliação *SoftED*, pois o detector sempre introduz algum atraso em relação ao instante $t - K$ de pontuação máxima.

Para a detecção de *drift* na série temporal de passes, foram implementados e avaliados cinco algoritmos distintos: (1) Page-Hinkley, (2) KSWIN, (3) ADWIN, (4) alarme fixo e (5) alarme aleatório. Os três primeiros são detectores de *drift* propriamente ditos, utilizados como modelos principais; os dois últimos servem como *baselines* para verificar se os detectores superam o que seria obtido por estratégias *naive*.

Para avaliar a sensibilidade do pipeline aos hiperparâmetros, foram testadas combinações W, C, K, conforme regras estabelecidas em IV.2, IV.3 e IV.4. A restrição $W \leq K$ garante que o detector esteja ativo quando a janela $[t - K,\, t]$ se abre; a restrição $C \leq K$ limita o silenciamento pós-alarme à duração da janela de tolerância. Os parâmetros internos de cada detector são descritos a seguir.

Os três detectores de *drift* são avaliados em *grid search* no protocolo *in-sample*:

Tabela IV.1: Grades de hiperparâmetros por detector.

Detector	**Parâmetro**	**Valores**
Page-Hinkley	λ	$\{3,\ 5,\ 7,\ 10\}$
	δ	$\{0{,}001;\ 0{,}005;\ 0{,}01\}$
	α	$\{0{,}90;\ 0{,}95;\ 0{,}99\}$
	modo	*both*
	min_instances	$\{1,\ 3\}$
KSWIN	α	$\{0{,}001;\ 0{,}005;\ 0{,}01;\ 0{,}05\}$
	w, s	centrados em K, com restrição $w \geq 2s$
ADWIN	δ	$\{0{,}0001;\ 0{,}0005;\ 0{,}001;\ 0{,}002;\ 0{,}005;\ 0{,}01;\ 0{,}05\}$

No protocolo *in-sample*, a combinação com maior F_1 agregado é selecionada por time e tarefa sobre todos os dados disponíveis. No protocolo de divisão temporal, os parâmetros selecionados nas primeiras 190 partidas são aplicados sem re-otimização às 190 restantes, conforme descrito na Seção IV.4.

IV.3.1 Detectores

O algoritmo de Page-Hinkley [Page, 1954] é um teste paramétrico sequencial que monitora o desvio cumulativo entre os valores observados e uma média de referência. A cada novo ponto, acumula-se a diferença entre o valor atual e a média estimada; quando esse acumulado ultrapassa um limiar λ, um alarme é emitido. O parâmetro δ controla a magnitude mínima de mudança a ser detectada, e α define a taxa de esquecimento da média de referência. O detector pode operar em dois modos: `up`, que sinaliza aumentos no volume de passes (interpretados como indicador de pressão ofensiva para passes certos); e `down`, que sinaliza reduções (associadas a uma possível perda de controle do adversário para passes certos). Ambos os modos são avaliados na grade de hiperparâmetros, permitindo identificar qual direção de mudança é mais informativa para a antecipação de gols.

O algoritmo KSWIN [Raab et al., 2020] é um detector não-paramétrico que aplica o teste de Kolmogorov-Smirnov entre duas subjanelas deslizantes sobre o fluxo de dados. A janela maior, de tamanho w, representa o histórico recente; a subjanela interna, de tamanho s, representa os dados mais atuais. Um *drift* é sinalizado quando a distância estatística entre as distribuições das duas subjanelas excede o nível de significância α. Por não assumir nenhuma distribuição paramétrica, o KSWIN é mais adequado para séries em que a normalidade não é garantida, como contagens de passes por minuto.

O algoritmo ADWIN [Bifet and Gavaldà, 2007] mantém uma janela de tamanho variável sobre o fluxo de dados e detecta mudanças na média comparando sub-janelas internas. Sempre que a diferença entre as médias de duas sub-janelas consecutivas excede um limiar definido pelo parâmetro

δ, a janela mais antiga é descartada e um alarme é emitido. O ADWIN é amplamente utilizado como detector de referência em problemas de *data streams* pela sua garantia teórica de taxa de falsos positivos controlada.

Os três detectores apresentam a mesma configuração de código, alterando apenas os hiperparâmetros, já definidos anteriormente na tabela IV.1.

Algoritmo 1 Pipeline de detecção de *drift*

Require: df: dados das partidas; $teams$: lista de times; $\mathcal{W}, \mathcal{C}$: grades de janela MA e *cooldown*; Θ: grade de hiperparâmetros do detector (Tabela IV.1); K: tolerância da avaliação SoftED assimétrica
Ensure: $best$: melhor (θ, W, C) por time e tarefa, maximizando F_1

for $team \in teams$ **do**
  $results \leftarrow \emptyset$
  for $W \in \mathcal{W}$, $C \in \mathcal{C}$ **do**
    for $match \in$ get_matches(df, $team$) **do**
      for $side \in$ {casa, fora} **do**
        for $task \in$ {attack, defense} **do**
          $goals \leftarrow$ goal_series($match$, $task$, $side$)
          for $feature \in$ {passe, passe_certo, passe_errado} **do**
            for $period \in$ periods($match$) **do**
              $\tilde{x}_{period} \leftarrow$ moving_average(x_{period}, W)
            end for
            for $\theta \in \Theta$ **do**
              for $period \in$ periods($match$) **do**
                $alarms_{period} \leftarrow$ detect_drift($\tilde{x}_{period}$, θ, C, W)
                TP, FP, FN, TN += SoftED($alarms_{period}$, $goals_{period}$, K, W)
              end for
              $results \leftarrow results \cup \{task, side, feature, \theta, W, C, TP, FP, FN, TN\}$
            end for
          end for
        end for
      end for
    end for
  end for
  $agg \leftarrow$ groupby($results$, $[task, feature, \theta, W, C]$) $\cdot$ sum(TP, FP, FN, TN)
  $agg.\{precision, recall, F_1\} \leftarrow$ compute_metrics(agg)
  $best[team] \leftarrow \arg\max_{\theta,W,C} F_1(agg)$ **por** $[task]$
end for
return $best$

O *baseline* fixo possui um único hiperparâmetro: o intervalo N entre alarmes consecutivos. Foi avaliado com $N = 22$ minutos, valor obtido dividindo o tempo regular de cada período (≈ 45 minutos) ao meio e arredondando para o inteiro mais próximo. Essa escolha garante exatamente dois alarmes por período: um próximo ao meio e outro próximo ao fim, sem depender de qualquer característica do sinal observado.

Algoritmo 2 *Baseline* de alarme fixo

Require: Série temporal $\mathbf{x}$, intervalo N (minutos), janela de tolerância K
Ensure: Sequência de alarmes $\mathbf{a}$, métricas TP, FP, FN, TN

for cada time e e partida p **do**
  for cada período $h \in \{1, 2\}$ **do**
    $\mathbf{x}_h \leftarrow$ subsequência de $\mathbf{x}$ no período h
    $T \leftarrow |\mathbf{x}_h|$
    for $t = 0, 1, \ldots, T-1$ **do**
      if $t > 0$ **e** $t \bmod N = 0$ **then**
        $a_t \leftarrow 1$
      else
        $a_t \leftarrow 0$
      end if
    end for
  end for
  Avaliar $\mathbf{a}$ com SoftED assimétrico (janela K)
end for

O *baseline* aleatório emite alarmes por processo de Bernoulli com $p = 0{,}5$, ou seja, cada minuto tem probabilidade igual de disparar ou não um alarme, sem qualquer análise do sinal observado. O experimento é repetido $R = 10$ vezes com *seeds* distintas para estimar a variância do resultado.

Algoritmo 3 *Baseline* aleatório (Bernoulli)

Require: Série temporal $\mathbf{x}$, probabilidade $p = 0{,}5$, janela de tolerância K, número de repetições R
Ensure: Sequência de alarmes $\mathbf{a}$, métricas TP, FP, FN, TN

for $r = 1, \ldots, R$ **do**
  for cada time e e partida p **do**
    for cada período $h \in \{1, 2\}$ **do**
      $T \leftarrow$ duração do período h em minutos
      for $t = 0, 1, \ldots, T-1$ **do**
        $u \sim \text{Uniforme}(0, 1)$
        $a_t \leftarrow \mathbf{1}[u < p]$
      end for
    end for
    Avaliar $\mathbf{a}$ com SoftED assimétrico (janela K)
  end for
end for
Reportar média e desvio-padrão das métricas sobre as R repetições

IV.4 Protocolo de Avaliação

Foram adotados dois protocolos de avaliação complementares: o primeiro compara os detectores entre si e em relação aos *baselines*; o segundo posiciona o melhor detector em relação aos resultados reportados por Lang et al. [2025].

IV.4.1 Método de Avaliação

A avaliação dos alarmes segue a variante assimétrica do *SoftED* [Salles et al., 2024] descrita na Seção IV.3. Um alarme emitido no instante $a \in [t - K, t]$ recebe pontuação $\text{score}(a, t) = 1 - (a - (t - K)) \,/\, K$, com pico em $a = t - K$; alarmes fora dessa janela são contabilizados como FP e gols sem alarme precedente geram FN.

A janela de tolerância K foi determinada a partir da análise exploratória descrita na Seção IV.2: o tempo médio até o primeiro gol é de 18 minutos, com desvio padrão de 13 minutos. O limite inferior foi fixado em $K_{\min} = 18-13 = 5$ minutos (IV.2), e o limite superior em $K_{\max} = 18-W_{\min} = 18-3 = 15$ minutos, onde $W_{\min} = 3$ é o menor *warmup* testado (Seção IV.3). Um valor intermediário $K = 10$ foi incluído para cobrir o ponto médio do intervalo. As três janelas, $K \in \{5, 10, 15\}$, foram avaliadas para verificar se a escolha de K afeta as conclusões comparativas entre os modelos.

Os valores de C testados satisfazem a restrição $C \leq K$ (IV.4) em todos os experimentos. Para $K = 5$, foram testados $C \in \{3, 5\}$ e $W \in \{3, 5, 10, 12\}$; para $K = 10$, $C \in \{3, 5, 10\}$ e $W \in \{3, 5\}$; para $K = 15$, $C \in \{3, 5, 10, 12, 15\}$ e $W = 3$. O conjunto de candidatos de W reduz-se à medida que K aumenta, pois a restrição $K + W \leq \bar{t}_{\text{gol}} = 18$ (IV.3) garante que o período de aquecimento não antecipe o limite da janela de avaliação.

IV.4.2 Protocolo de Divisão Temporal

A robustez da seleção de hiperparâmetros é avaliada por meio de uma divisão temporal: as primeiras 190 partidas da temporada são utilizadas para seleção de uma configuração global única, e as 190 partidas restantes para avaliação *out-of-sample*, sem qualquer re-otimização. Essa análise é deliberadamente conservadora: em vez de calibração independente por time/tarefa, uma única combinação de parâmetros é aplicada simultaneamente a todos os 40 pares time/tarefa (20 times × 2 perspectivas táticas: ataque e defesa).

Os parâmetros fixados no conjunto de treino e mantidos inalterados no teste são: W e C (janela de suavização e *cooldown*); e os parâmetros internos de cada detector — δ, λ, α e modo para o Page-Hinkley; α, w e s para o KSWIN; δ para o ADWIN. Nenhum parâmetro é re-otimizado após a divisão.

IV.4.3 Métricas

Cinco métricas são reportadas para cada modelo na sua melhor configuração de hiperparâmetros:

- **Precisão**: fração dos alarmes que correspondem a TPs, penalizando modelos com excesso de FP.

- **Recall**: fração dos gols cobertos por ao menos um alarme na janela de tolerância, penalizando modelos que perdem eventos.
- $\boldsymbol{F_1}$: média harmônica entre precisão e recall, utilizada como critério de seleção de hiperparâmetros no protocolo *in-sample*.
- $\boldsymbol{F_{0,5}}$: versão de F_β com $\beta = 0{,}5$, que penaliza FP duas vezes mais que FN. Em contextos onde alarmes imprecisos têm custo operacional elevado, como alertas em tempo real durante uma partida, cobertura total é menos valiosa do que alarmes confiáveis.
- **MCC** (*Matthews Correlation Coefficient*): robusto a classes desbalanceadas [Sujon et al., 2025], característica inerente a séries com poucos gols. Utilizado como critério de seleção no protocolo de divisão temporal e como métrica de comparação direta com os resultados de Lang et al. [2025].

IV.4.4 Baselines

Dois *baselines* sem detector são incluídos como referência. O Alarme Fixo (Algoritmo 2) emite um alarme a cada $N = 22$ minutos de forma determinística, garantindo exatamente dois alarmes por período, um próximo ao meio e outro próximo ao fim. Representa o limite inferior determinístico: quantifica o desempenho esperado por uma estratégia puramente periódica, sem qualquer análise do sinal observado. O Alarme Aleatório (Algoritmo 3) emite alarmes por processo de Bernoulli com $p = 0{,}5$, ou seja, cada minuto tem probabilidade igual de disparar ou não um alarme, independentemente do sinal. O experimento é repetido $R = 10$ vezes com *seeds* distintas para estimar a variância do resultado. O objetivo é verificar se os detectores de *drift* superam ambos os *baselines*.

IV.4.5 Análises de Subgrupo

Além da comparação global entre modelos, os resultados são desagregados em duas dimensões. A primeira considera a perspectiva tática: o *pipeline* é executado separadamente para o time que marca o gol e para o time que sofre. A análise exploratória indicou que o time defensor apresenta queda mais pronunciada no volume de passes nos minutos anteriores ao gol (Seção IV.2), o que sugere que a detecção pode ter sensibilidades distintas nas duas perspectivas.

A segunda dimensão considera o contexto de jogo, separando mandantes de visitantes. Os times mandantes respondem por 59,0% dos gols da temporada, indicando assimetria entre os contextos. Essa desagregação permite verificar se os detectores capturam *drift* de forma equilibrada independentemente do contexto, ou se há viés sistemático associado ao fator local.

Em ambas as análises, a grade de hiperparâmetros é mantida fixada na configuração otimizada globalmente pelo F_1 agregado; as desagregações têm caráter exploratório e visam identificar em

quais subgrupos os detectores apresentam melhor ou pior desempenho.

Capítulo V Resultados e Discussão

V.1 Configuração do ambiente de execução dos testes

Os experimentos foram executados em um computador Apple MacBook Air com chip M1, 8GB de memória RAM e sistema operacional macOS Sonoma 14.4.1. A implementação foi realizada utilizando a linguagem de programação Python em sua versão 3.12.3, com as seguintes bibliotecas principais: *river 0.22.0* para os detectores de drift, *pandas 2.3.0* para manipulação dos dados, *numpy 1.26.4* para operações numéricas e *matplotlib 3.9.0* para geração das figuras.

V.2 Comparação entre Janelas de Avaliação

Conforme descrito na Seção IV.3, a janela de tolerância K foi variada em três valores ($K \in \{5, 10, 15\}$) derivados do tempo médio até o primeiro gol do período (18 minutos) e do menor *warmup* testado ($W_{\min} = 3$). Esta seção compara o desempenho dos detectores nessas três configurações, atravéz de três métricas complementares: (1) o MCC, que avalia o equilíbrio geral entre acertos e erros, penalizando simultaneamente FP e FN e sendo robusto a classes desbalanceadas, dado que gols são eventos raros; (2) o F_1, que pondera precisão e *recall* igualmente, sendo sensível ao volume de alarmes disparados; e (3) o *recall*, que mede a proporção de gols precedidos por ao menos um alarme dentro da janela $[t - K, t]$, capturando a cobertura dos eventos de interesse.

Para construir a Tabela V.1, cada detector foi avaliado sobre todo o corpus: 20 equipes, 380 partidas, 2 tarefas (ataque e defesa), 2 mandos de campo (mandante e visitante) e 3 *features* (passes totais, passes certos e passes errados). Para cada combinação de hiperparâmetros (W, C, δ e *feature*), os valores de TP, FP, FN e TN foram somados globalmente e o MCC calculado sobre esse total. O valor reportado na tabela corresponde à configuração que maximiza esse MCC global, trata-se, portanto, do melhor resultado *in-sample* de cada detector, não de uma média entre configurações. A avaliação com divisão temporal, mais conservadora, é apresentada na Seção V.3.

A Tabela V.1 apresenta as três métricas para cada detector nos três valores de K testados.

O Page-Hinkley obteve o maior MCC em todos os cenários, com *recall* crescente de $K = 5$ (0,45) para $K = 10$ (0,51), seguido de queda em $K = 15$ (0,46). Em $K = 5$, o KSWIN (MCC = 0,046) fica abaixo do *Baseline* Aleatório (0,058), sugerindo que a janela de cinco minutos é insuficiente para que o detector acumule evidência antes do gol; a partir de $K = 10$, o KSWIN assume consistentemente

Tabela V.1: MCC, F1 e *recall* por detector e janela de avaliação K. Melhor configuração global por detector.

	$K = 5$			$K = 10$			$K = 15$		
Detector	MCC	F1	Rec.	MCC	F1	Rec.	MCC	F1	Rec.
Page-Hinkley	0,081	0,060	0,45	**0,130**	0,092	**0,51**	**0,156**	**0,123**	0,46
KSWIN	0,046	0,055	0,15	0,090	**0,088**	0,24	0,077	0,089	0,13
ADWIN	0,014	0,009	0,01	0,019	0,012	0,01	≈0	0	0,00
Baseline Fixo	0,024	0,039	0,09	0,037	0,050	0,11	0,069	0,076	0,16
Baseline Aleatório	0,058	0,037	0,76	0,065	0,039	0,79	0,060	0,038	0,77

a segunda posição. O ADWIN permanece abaixo de ambos os baselines em todos os K testados, e em $K = 15$ seu MCC colapsa para ≈ 0, confirmando que a detecção de mudança de média não se adapta ao perfil inerentemente variável dos dados de passes.

O *Baseline* Fixo apresenta melhora aparente com o aumento de K (MCC de 0,024 para 0,069): com alarmes a cada 22 minutos, janelas maiores simplesmente ampliam a tolerância e capturam esses alarmes como TP com maior frequência. O *Baseline* Aleatório mantém *recall* elevado ($\approx$ 0,76–0,79) em todos os K por disparar em metade dos minutos, mas seu MCC permanece baixo e estável ($\approx$ 0,06), confirmando que cobertura sem precisão não se traduz em desempenho.

O ordenamento entre os detectores reais (Page-Hinkley, KSWIN, ADWIN) é estável nos três valores de K; a instabilidade ocorre apenas entre os baselines, cujo comportamento não depende do sinal. A diferença de MCC entre $K = 10$ e $K = 15$ é de +0,026 a favor de $K = 15$, não desprezível, mas deve-se considerar que janelas maiores ampliam a região de tolerância, permitindo que mais alarmes sejam contabilizados como TP e inflando artificialmente todas as métricas, conforme discutido na Seção IV.3. Nesse contexto, $K = 10$ apresenta *recall* superior (0,51 vs. 0,46) mesmo com janela mais restrita, indicando que os alarmes gerados são temporalmente mais próximos dos gols, e não apenas beneficiados pela tolerância. Como o objetivo do trabalho é detectar mudanças no padrão de passes que precedam gols com antecedência operacionalmente útil, a cobertura de eventos, medida pelo *recall*, é o critério mais relevante para a escolha de K. Portanto, as seções seguintes aprofundam a análise para $K = 10$.

V.3 Avaliação com Divisão Temporal

Com base na janela $K = 10$ selecionada na seção anterior, esta seção apresenta a comparação principal entre os detectores. A avaliação adota divisão temporal: os primeiros 190 jogos da temporada foram utilizados para seleção de uma configuração global única e os 190 jogos restantes para avaliação *out-of-sample*, sem qualquer re-otimização. Este protocolo é deliberadamente conservador em dois sentidos: (1) a configuração é selecionada uma única vez sobre o conjunto de treino e apli-

cada sem ajuste aos 190 jogos de teste; (2) em vez de calibração independente por equipe/tarefa, uma única combinação de parâmetros é aplicada simultaneamente a todos os 40 pares equipe/tarefa. A Figura V.1 apresenta os resultados.

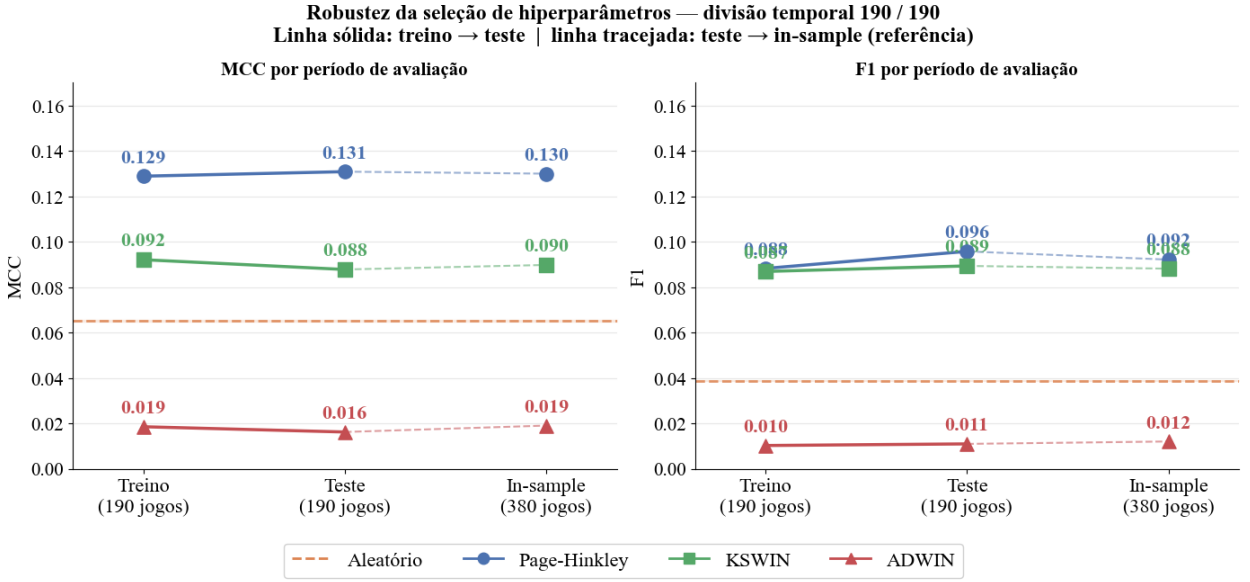

Figura V.1: Robustez da seleção de hiperparâmetros: MCC e F1 por período de avaliação (treino: partidas 1–190; teste: partidas 191–380; in-sample: todas as 380 partidas). A linha tracejada indica o desempenho do Random Walk como referência.

O Page-Hinkley mantém MCC = 0,131 no conjunto de teste, valor consistente com o *in-sample* (0,130) e superior ao Baseline Aleatório (linha tracejada, MCC ≈ 0,065), confirmando que o sinal identificado não é artefato de otimização sobre a amostra completa. O KSWIN apresenta leve queda entre treino (0,092) e teste (0,088), mas mantém-se acima do acaso em ambos os conjuntos, indicando que sua capacidade discriminativa é real, ainda que inferior à do Page-Hinkley.

O ADWIN, por sua vez, permanece sistematicamente abaixo do Baseline Aleatório em todas as partições (MCC ≈ 0,016–0,019), indicando que o detector não extrai informação útil do sinal. Dois fatores explicam esse comportamento. Primeiro, o ADWIN detecta mudanças de *média* em séries temporais, mas o volume de passes por minuto em futebol é inerentemente instável (oscila muito dentro de um período), dificultando que o detector distinga sinal de ruído mesmo com o parâmetro de sensibilidade máximo ($\delta = 0{,}05$). Segundo, o ADWIN reinicia sua janela interna ao detectar *drift* e recomeça a acumular evidência do zero; em partidas de 45 minutos, esse comportamento pode impedir a acumulação de evidência suficiente para uma detecção confiável. O Page-Hinkley, por usar soma cumulativa de desvios direcionais, adapta-se melhor a esse regime. A ampliação do grid de δ para valores acima de 0,05 configura uma direção natural de trabalho futuro para o ADWIN neste domínio.

A estabilidade entre treino, teste e *in-sample* para o Page-Hinkley e o KSWIN indica que o *overfitting* de hiperparâmetros é limitado neste contexto, resultado esperado dado o número reduzido

de parâmetros livres de cada detector. Quanto ao F_1, Page-Hinkley (0,092) e KSWIN (0,089) apresentam valores próximos no conjunto de teste, sugerindo que a vantagem do Page-Hinkley em MCC se deve a uma melhor relação entre FP e FN, não necessariamente a maior volume de acertos.

V.4 Teto de Desempenho *In-Sample*

Para estimar o teto de desempenho alcançável por cada detector, adota-se a avaliação *in-sample*: cada par (time, tarefa) seleciona independentemente a configuração de hiperparâmetros que maximiza o MCC nos dados disponíveis, sem separação temporal. O resultado global é obtido pela soma das contagens TP/FP/FN/TN de todos os pares otimizados individualmente. Trata-se, portanto, de um teto otimista que quantifica o potencial máximo de cada abordagem quando calibrada individualmente por time e tarefa.

Para tornar o volume de alarmes comparável à duração real de uma partida, normaliza-se o total de alarmes pelo número de séries temporais avaliadas: $1.520 = 380$ partidas $\times$ 2 times $\times$ 2 tarefas. A coluna "Alarmes/90 min" expressa, portanto, quantos alarmes o detector emitiria em uma única série de 90 minutos.

A Tabela V.2 apresenta os resultados para $K = 10$.

Tabela V.2: Melhor configuração por (time, tarefa), $K = 10$. Alarmes/90 min = total de alarmes /1.520.

Detector	MCC	σ_{MCC}	F1	σ_{F1}	Recall	Alarmes/90 min
Page-Hinkley	**0,147**	0,020	**0,123**	0,019	0,405	6,29
KSWIN	0,118	0,016	**0,124**	0,018	0,203	2,49
Aleatório*	0,069	0,0004†	0,040	0,0001†	0,806	47,68
Baseline Fixo	0,037	0,012	0,050	0,009	0,108	4,00
ADWIN	0,031	0,022	0,014	0,014	0,007	0,05

*Média de 10 execuções independentes. †Desvio entre runs; demais: desvio entre os 20 times.

O Page-Hinkley obtém o maior MCC ($0{,}147 \pm 0{,}020$), seguido pelo KSWIN ($0{,}118 \pm 0{,}016$). A diferença entre os dois ($\Delta = 0{,}029$) excede o desvio padrão de ambos, indicando que a vantagem do Page-Hinkley não é atribuível à variabilidade entre times. Em F_1, os dois detectores ficam estatisticamente empatados ($0{,}123 \pm 0{,}019$ vs. $0{,}124 \pm 0{,}018$), o que reforça o MCC como critério de ordenação mais discriminativo neste cenário. Ambos superam o baseline aleatório (0,069), que apesar de apresentar recall elevado (0,806) o faz às custas de um volume de alarmes proibitivo: com $p = 0{,}5$ por minuto, espera-se cerca de 47,7 alarmes em 90 min, um alarme a cada dois minutos. O Baseline Fixo (intervalo de 22 min) gera exatamente 4,0 alarmes por série, o que é verificável diretamente: $\lfloor 90/22 \rfloor = 4$. O ADWIN, com apenas 0,05 alarmes por série ($\approx$1 alarme a cada 20 partidas), fica abaixo do baseline aleatório em MCC e F_1, evidenciando que a detecção de mudança de média não se adapta ao perfil inerentemente variável dos dados de passes. Esse comportamento

é agravado pela elevada instabilidade entre times: $\sigma_{\mathrm{MCC}} = 0{,}022$ é o maior valor entre todos os detectores e representa um coeficiente de variação de 71%, com F_1 variando de 0,000 a 0,048 conforme o time. Em todos os modelos, incluindo o Page-Hinkley, a precisão permanece baixa, o que levanta a questão sobre a origem dos falsos positivos.

Uma hipótese natural seria atribuí-los às partidas sem gol, nas quais nenhum alarme pode ser verdadeiro positivo. No entanto, as 27 partidas encerradas em 0×0 (7,1% da temporada) apresentam taxa média de alarmes por série ($6{,}26 \pm 3{,}04$) praticamente idêntica à das partidas com gol ($6{,}29 \pm 2{,}78$). A imprecisão reflete, portanto, uma assimetria estrutural do problema: mudanças no padrão de passes são frequentes ao longo de uma partida, enquanto gols são eventos raros. A maioria das variações detectadas não precede um gol dentro da janela de tolerância K, independentemente do resultado final da partida.

A robustez do Page-Hinkley foi verificada por meio de uma análise de sensibilidade sobre os hiperparâmetros externos W (janela de suavização) e C (cooldown). Para cada uma das seis combinações $W \in \{3, 5\}$ e $C \in \{3, 5, 10\}$ min, os parâmetros internos do detector foram mantidos na melhor configuração por (time, tarefa) e o MCC foi calculado agregando todos os times. O intervalo de variação foi estreito: $\Delta_{\mathrm{MCC}} = 0{,}017$, de 0,121 ($W$=3, C=3) a 0,138 (W=3, C=10), conforme a Figura V.2. O resultado indica que a vantagem do Page-Hinkley não depende de uma escolha precisa de hiperparâmetros: toda a grade testada supera o baseline aleatório (0,065).

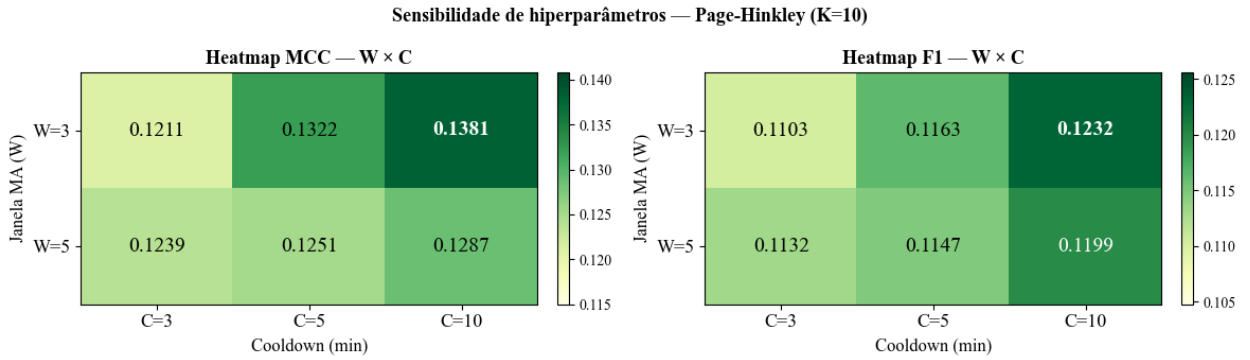

Figura V.2: Sensibilidade do Page-Hinkley aos hiperparâmetros externos janela de suavização W e cooldown C, para $K = 10$. Cada célula agrega todos os 20 times com os parâmetros internos do detector fixados na melhor configuração por (time, tarefa). O intervalo de variação é $\Delta_{\mathrm{MCC}} = 0{,}017$ (de 0,121 a 0,138), indicando baixa sensibilidade à escolha de W e C.

É importante ressaltar que estes resultados constituem um teto: a ausência de separação temporal implica que as configurações foram ajustadas nos mesmos dados em que são avaliadas. A avaliação *out-of-sample* rigorosa foi apresentada na Seção V.3.

A superioridade do Page-Hinkley sobre o KSWIN foi verificada por teste de permutação com 1.000 iterações sobre as 380 partidas da temporada, embaralhando por partida os contadores de TP, FP, FN e TN entre os dois modelos e recalculando a diferença de MCC a cada iteração. O

p-valor obtido foi $p < 0{,}001$ (bicaudal), indicando que a diferença observada de 0,040 está fora da distribuição nula e não é atribuível a variação aleatória. A Figura V.3 ilustra a distribuição das diferenças permutadas e a posição da diferença observada.

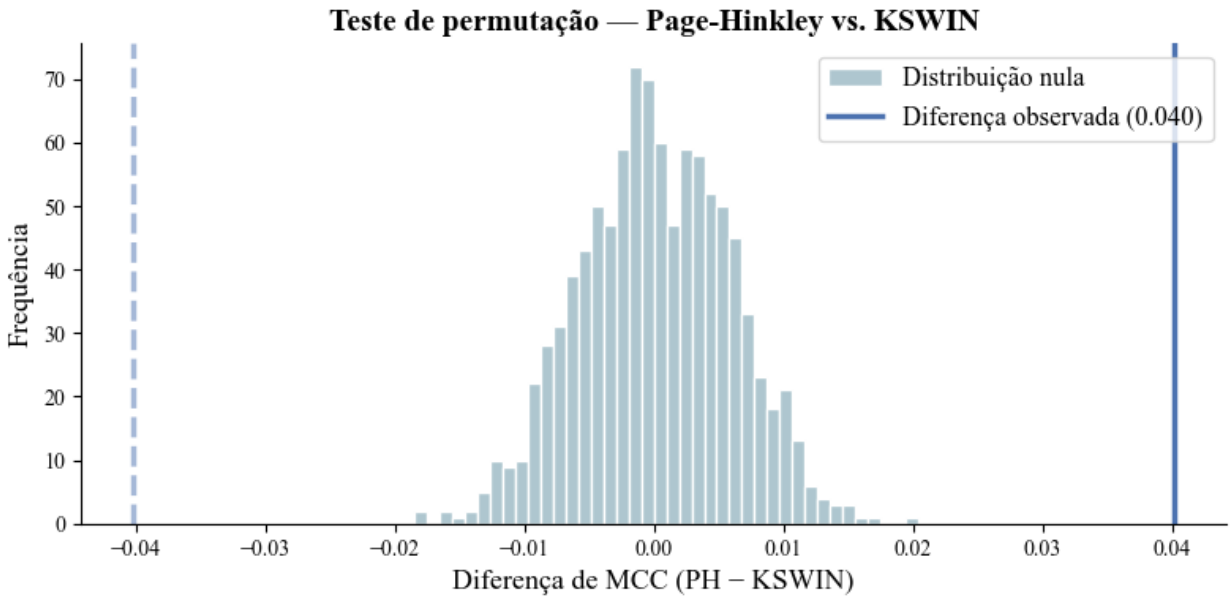

Figura V.3: Teste de permutação: distribuição nula de diferenças de MCC (Page-Hinkley − KSWIN) sob 1.000 permutações por partida. A linha sólida indica a diferença observada (0,033); a linha tracejada indica o simétrico. $p < 0{,}001$ (bicaudal).

Os valores absolutos de F_1 são baixos em todos os modelos, o que é esperado dado o desequilíbrio da classe-alvo: gols ocorrem em média duas vezes por partida, tornando a maioria dos instantes temporais negativos por definição. Em cenários assim, o MCC é a métrica mais adequada, conforme argumentado por Sujon et al. [2025], e é também a métrica adotada por Lang et al. [2025] para comparação. A Figura V.4 sintetiza os resultados de MCC por modelo.

Lang et al. [2025] aplicam aprendizado de máquina supervisionado sobre 28 indicadores de performance da Bundesliga para prever a ocorrência de um gol nos próximos 3 minutos, reportando $\text{MCC}_{\text{mean}} = 0{,}061$ para a melhor combinação de indicadores. O Page-Hinkley obtém MCC = 0,126 *out-of-sample*, valor numericamente superior, embora a comparação direta seja limitada pela diferença de janela de avaliação (janelas mais amplas tendem a inflar o MCC mecanicamente, o que favorece o detector proposto em termos absolutos).

A diferença de janela de avaliação entre os dois trabalhos não é acidental e reflete objetivos distintos. Lang et al. [2025] buscam prever se um evento ocorrerá nos próximos 3 minutos, e os próprios autores identificam como aplicações centrais do método a análise de desempenho e a otimização de *odds* em apostas ao vivo [Lang et al., 2025]. Esses são contextos em que uma janela curta é suficiente, pois não exigem que nenhum agente humano tome uma decisão e a execute dentro do período previsto. A presente abordagem, por outro lado, é orientada à intervenção tática em

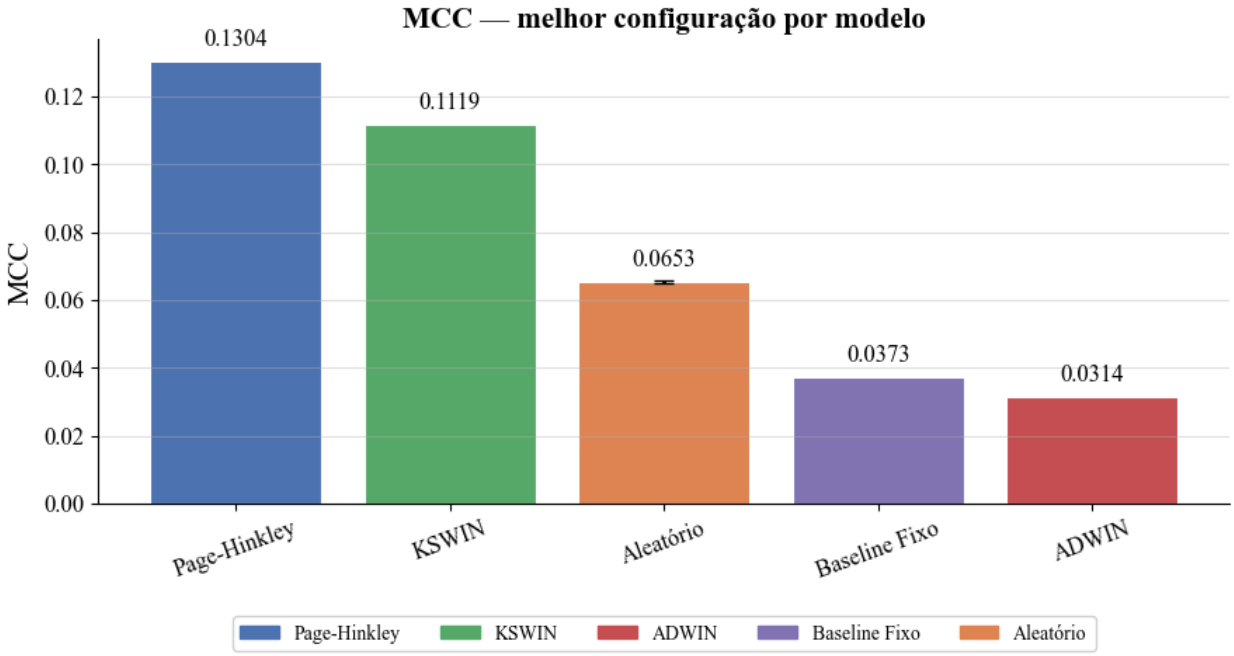

Figura V.4: MCC na melhor configuração in-sample por modelo.

tempo real: substituições e reorganizações táticas no futebol profissional exigem que o treinador perceba a mudança, delibere e comunique a instrução antes que ela seja executada em campo, o que impõe um horizonte mínimo operacionalmente relevante maior do que 3 minutos. A adoção de $K = 10$ minutos reflete esse requisito.

Reconhece-se que janelas de avaliação mais amplas tendem a favorecer detectores em termos absolutos de MCC, o que limita a comparação direta entre os dois trabalhos. Ainda assim, o resultado indica que a detecção de *drift* em séries táticas é uma abordagem viável para a caracterização de momentos de pressão ofensiva em cenários onde a antecedência tem valor operacional, mesmo com uma única *feature* não supervisionada e sem necessidade de dados rotulados.

V.5 Relação entre Volume de Alarmes e MCC

A Figura V.5 posiciona cada configuração testada no espaço MCC × volume de alarmes. Cada ponto representa uma combinação de janela de suavização W agregada sobre todos os times e tarefas.

O Page-Hinkley concentra as duas configurações testadas ($W = 3$ e $W = 5$, para $K = 10$) com alto MCC e volume moderado de alarmes (21–25 mil na temporada), acima da linha de referência do *baseline* aleatório. O KSWIN apresenta comportamento semelhante, porém com MCC inferior a ambas as configurações do Page-Hinkley: $W = 5$ gera mais alarmes e MCC maior do que $W = 3$, sugerindo que janelas de média móvel mais longas capturam melhor as mudanças no padrão de passes. O ADWIN permanece próximo à origem do eixo horizontal, com apenas 9 e 234 alarmes

para $W = 3$ e $W = 5$, respectivamente, o que explica seu MCC próximo de zero: o detector raramente emite alertas, não cobrindo os gols suficientemente.

O *Baseline* Fixo (intervalo de 22 minutos) produz cerca de 6 mil alarmes e MCC de 0,037, abaixo do *baseline* aleatório, evidenciando que alarmes periódicos sem análise do sinal não superam a aleatoriedade. Por fim, o *Baseline* Aleatório, no extremo direito do gráfico, evidencia que seu MCC moderado (0,065) é acompanhado de um volume de alarmes inviável em aplicação prática ($\approx$ 218 mil na temporada, ou cerca de 47 por série de 90 minutos).

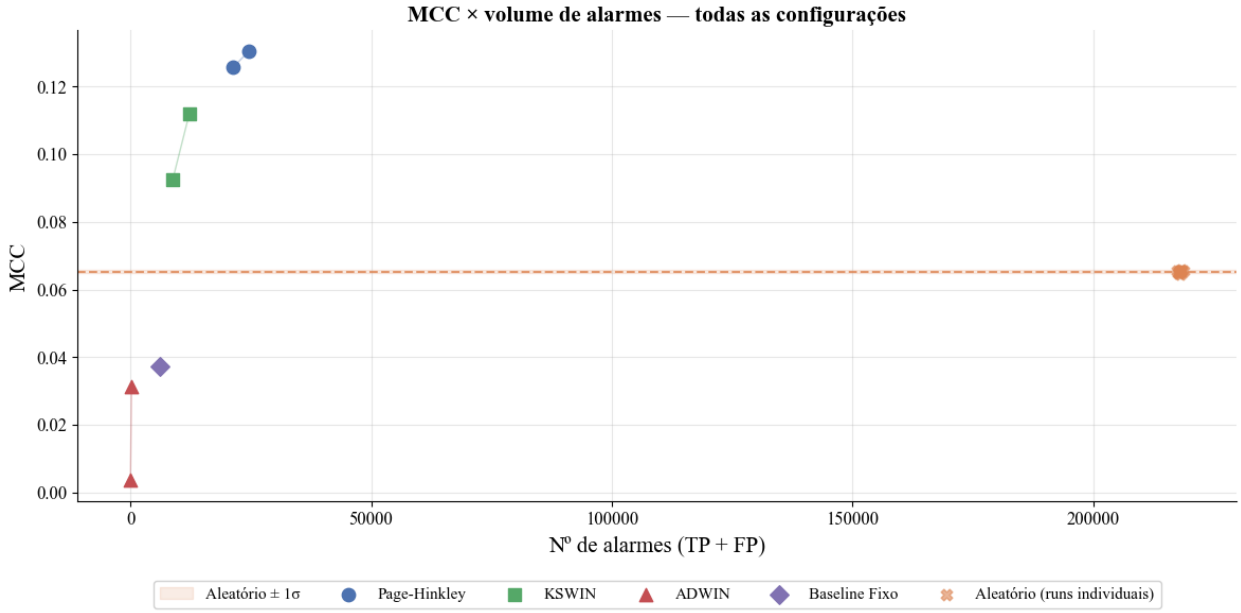

Figura V.5: MCC em função do volume de alarmes para todas as configurações avaliadas.

A Figura V.6 complementa a análise apresentando a curva Precisão × *Recall* para todas as configurações testadas.

O padrão mais evidente é a troca entre precisão e *recall* à medida que o detector emite mais alarmes. O ADWIN ocupa o extremo superior-esquerdo: precisão entre 0,10 e 0,16, mas *recall* próximo de zero ($< 0{,}01$), evidenciando que seus raros alarmes tendem a preceder gols, porém a cobertura é insuficiente. O Page-Hinkley inverte esse padrão: ambas as configurações ($W = 3$ e $W = 5$) concentram-se em torno de *recall* $\approx$ 0,31–0,33 com precisão $\approx$ 0,071, o maior *recall* entre os detectores reais testados. O KSWIN ocupa posição intermediária: *recall* entre 0,14 e 0,20 e precisão $\approx$ 0,083, com os dois valores de W produzindo precisão quase idêntica mas *recall* distintos. O *Baseline* Fixo (N= 22 min) apresenta precisão 0,033 e *recall* 0,11, sem vantagem sobre os detectores em nenhuma das dimensões. Por fim, o *Baseline* Aleatório ocupa o extremo inferior-direito (*recall* $\approx$ 0,79, precisão $\approx$ 0,020): a cobertura elevada é consequência direta do volume excessivo de alarmes ($\approx$ 48 por série de 90 min), e não de detecção genuína.

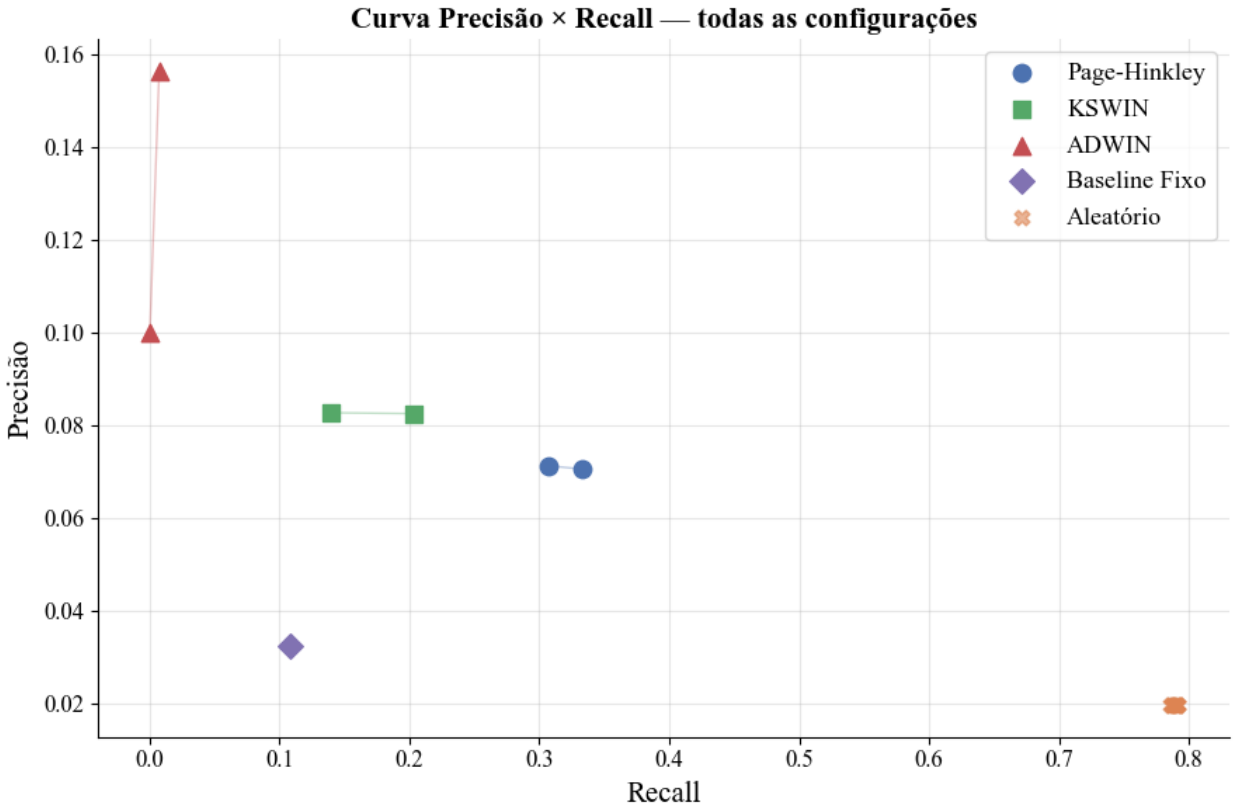

Figura V.6: Curva Precisão × Recall para todas as configurações avaliadas.

V.6 Análise por $F_{0,5}$

A Figura V.7 apresenta o F_1 e o $F_{0,5}$ de cada modelo na melhor configuração. O $F_{0,5}$ penaliza falsos positivos mais fortemente que o F_1 ($\beta < 1$), capturando melhor o custo de alarmes desnecessários em contextos de aplicação tática ao vivo.

Em F_1, o KSWIN supera marginalmente o Page-Hinkley (0,1174 contra 0,1166), inversão que não ocorre em MCC, onde o Page-Hinkley lidera com folga (0,130 contra 0,112). Essa discrepância reflete o perfil distinto dos dois detectores: o KSWIN emite menos alarmes com precisão ligeiramente superior, enquanto o Page-Hinkley cobre mais gols com volume maior de alarmes; o MCC, por incorporar os verdadeiros negativos, captura melhor essa diferença.

Sob $F_{0,5}$, o KSWIN mantém a liderança (0,094) à frente do Page-Hinkley (0,084), e a diferença entre os dois cresce, consistente com o fato de que o KSWIN gera menos falsos positivos por gol coberto. O movimento mais expressivo é o do ADWIN, que sobe acima do *Baseline* Aleatório (0,031 contra 0,025): embora o ADWIN tenha *recall* próximo de zero, sua precisão elevada ($\approx 0{,}16$) é recompensada quando a métrica penaliza falsos positivos, ao passo que o Aleatório é fortemente prejudicado por seu volume excessivo de alarmes.

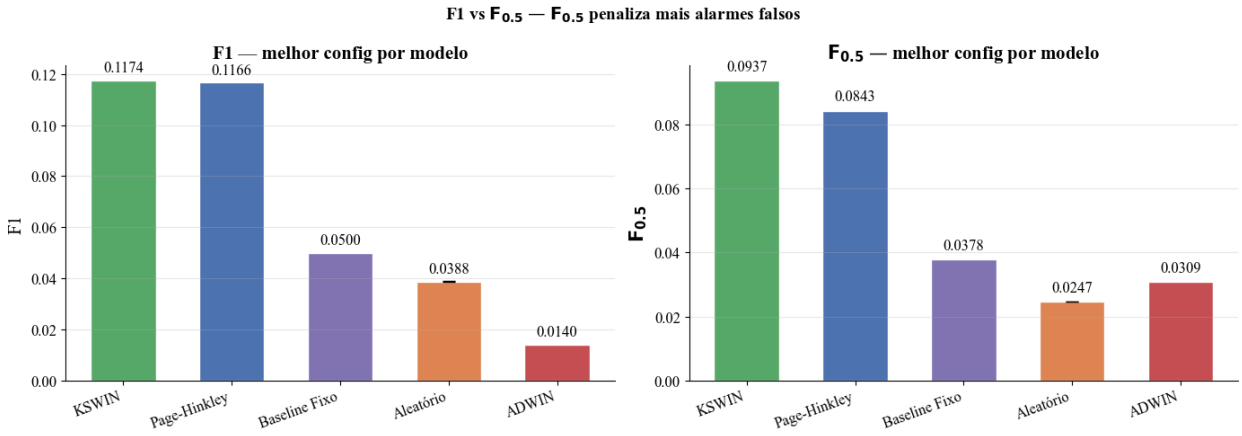

Figura V.7: Comparação entre F1 e $F_{0,5}$ na melhor configuração por modelo.

V.7 Análise por Subgrupo

V.7.1 Ataque vs. Defesa

A Figura V.8 desagrega o F_1 por perspectiva tática: alarmes gerados a partir dos passes do time que marcou o gol (*ataque*) *versus* alarmes gerados a partir dos passes do time que sofreu o gol (*defesa*).

O resultado mais expressivo é a liderança diferenciada entre os dois melhores detectores: o Page-Hinkley tem maior F_1 na perspectiva ofensiva (0,131 contra 0,127 do KSWIN), enquanto o KSWIN supera o Page-Hinkley na perspectiva defensiva (0,125 contra 0,124). Em ambos os casos as diferenças são inferiores a 0,007, e todos os detectores favorecem ligeiramente a perspectiva ofensiva sobre a defensiva, com exceção do Baseline Fixo, que apresenta desempenho idêntico entre as duas perspectivas (0,050). A ausência de um padrão assimétrico consistente sugere que a perspectiva tática não é um fator determinante no desempenho dos detectores.

Esse resultado contrasta com a hipótese exploratória levantada na Seção IV.2, que indicava queda mais pronunciada no volume de passes do time defensor nos minutos anteriores ao gol. Uma explicação possível é que a análise exploratória capturou uma diferença de volume médio entre janelas, enquanto os detectores de *drift* respondem à variação estrutural na série, fenômenos distintos que não necessariamente coincidem. Independentemente da causa, o resultado tem implicação prática positiva: o *pipeline* não requer conhecimento prévio de qual perspectiva é mais informativa, podendo operar sobre ambas simultaneamente sem perda de desempenho.

V.7.2 Mandante vs. Visitante

A Figura V.9 desagrega o F_1 pela condição de mando de campo. O KSWIN lidera em ambas as condições (0,172 como mandante e 0,170 como visitante), seguido pelo Page-Hinkley, que apresenta desempenho idêntico nas duas condições (0,167). O *Baseline* Fixo e o ADWIN são os únicos modelos

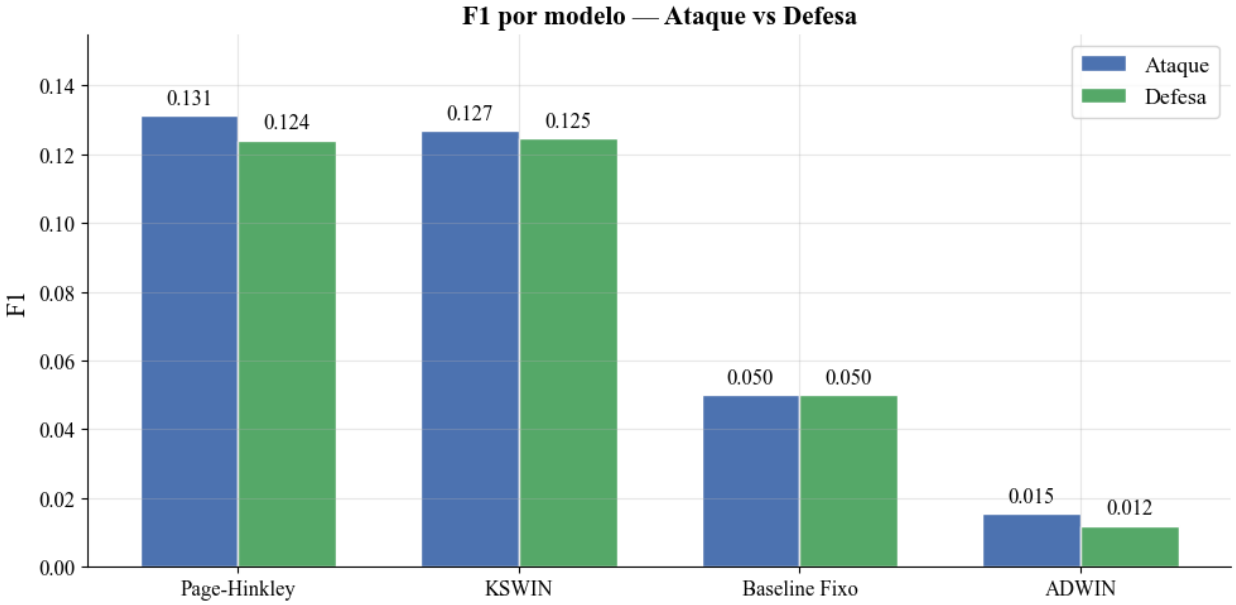

Figura V.8: F1 desagregado por perspectiva tática (ataque vs. defesa) na melhor configuração por detector.

em que o visitante supera o mandante (0,068 vs. 0,064 e 0,030 vs. 0,019, respectivamente), embora as diferenças sejam pequenas. De forma geral, a condição de mando de campo não é um fator determinante para o desempenho dos detectores.

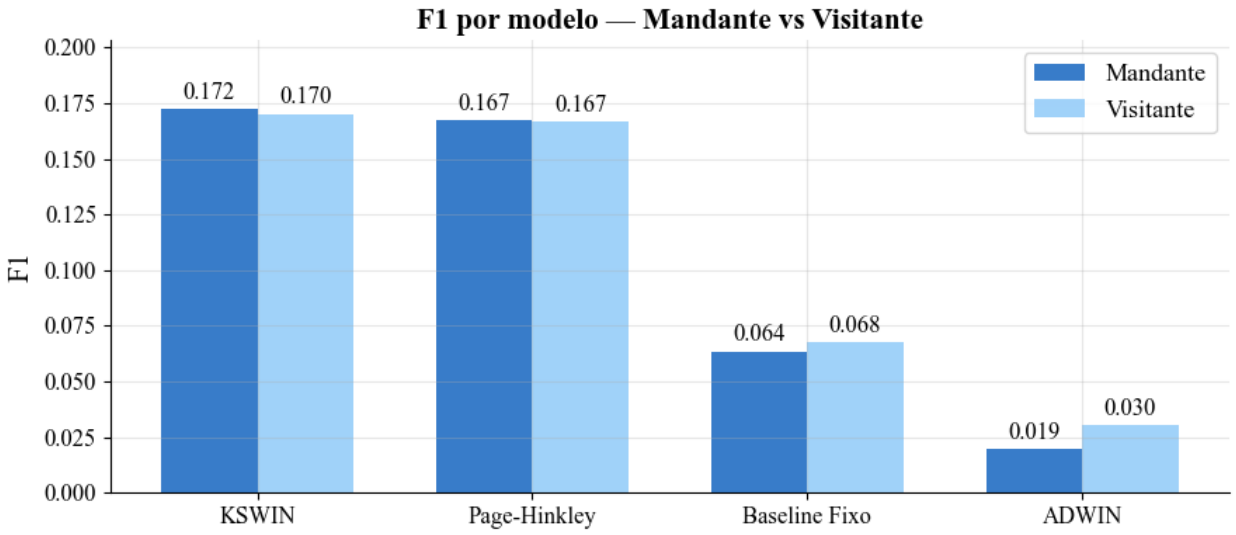

Figura V.9: F1 desagregado por condição de jogo (mandante vs. visitante) na melhor configuração por detector.

A alternância na liderança entre Page-Hinkley e KSWIN ao longo das análises desagregadas reflete, essencialmente, diferenças marginais amplificadas pela escolha de métrica. O MCC incorpora os verdadeiros negativos — minutos sem gol em que nenhum alarme foi emitido —, de modo que o Page-Hinkley, com maior volume de alarmes ($\approx$ 25 mil) e *recall* mais elevado, beneficia-se do ganho em TP $\times$ TN que supera o custo dos falsos positivos adicionais. O F_1, por ignorar os

verdadeiros negativos, é sensível à precisão: o KSWIN, mais conservador ($\approx$ 12 mil alarmes), obtém precisão ligeiramente superior e inverte a liderança por margem mínima (0,117 contra 0,117). Nas desagregações por perspectiva tática e mando de campo, as diferenças entre os dois detectores são da ordem de 0,001–0,007, aquém do limiar de estabilidade — o teste de permutação confirma significância estatística para a comparação global ($p < 0{,}001$), mas não sustenta conclusões sobre subcategorias. Em síntese, Page-Hinkley e KSWIN são equivalentes em F_1, enquanto o Page-Hinkley apresenta vantagem mais robusta em MCC; por capturar tanto a capacidade de cobertura quanto a de discriminação, o MCC constitui a métrica primária de comparação adotada neste trabalho.

V.8 Contribuições em Relação à Literatura

Os trabalhos mais próximos ao tema, incluindo Lang et al. [2025], enquadram o problema como classificação supervisionada, requerendo um conjunto de dados rotulados para treinamento, engenharia de *features offline* e retreinamento periódico. A abordagem proposta opera de forma completamente não-supervisionada e *online*: os detectores de *drift* processam a série de passes em tempo real sem acesso a rótulos, e a robustez dos resultados é verificada por divisão temporal com 190 partidas de treino e 190 de teste, sem re-otimização dos parâmetros no conjunto de teste. Essa característica é reconhecida na literatura de detecção de *drift* como especialmente valiosa em cenários de monitoramento contínuo, nos quais rótulos confiáveis raramente estão disponíveis de forma imediata [Hinder et al., 2024]. No contexto esportivo, isso torna a abordagem aplicável a competições com dados escassos ou a cenários em que os rótulos de gol não estejam disponíveis a priori.

Indicadores de performance baseados em passes já foram explorados na literatura [Lang et al., 2025], mas como valores agregados, como total de passes em uma janela ou taxa de acerto. O presente trabalho trata a variação estrutural na taxa de passes (isto é, a mudança no processo gerador da série) como sinal de *drift* associado à antecipação de gol. Essa distinção é relevante: não é o volume de passes em si que precede o gol, mas a quebra do padrão habitual que sinaliza uma mudança no equilíbrio tático da partida. A caracterização do fenômeno como *drift* (mudança persistente na distribuição geradora) e não como anomalia (desvio transiente) está alinhada com a distinção formal proposta por Li and Müller [2022], que operacionaliza a duração como critério de separação entre os dois fenômenos, e com a definição de *concept drift* adotada em levantamentos recentes da área [Hinder et al., 2024].

O protocolo de avaliação adotado, com o *SoftED* [Salles et al., 2024], é em si uma contribuição metodológica. Diferentemente de avaliações binárias com janela fixa, o *SoftED* atribui uma pontuação contínua a cada alarme proporcional à sua antecipação temporal em relação ao gol, com janela assimétrica $[t - K,\, t]$. Alarmes próximos ao instante $t - K$ recebem pontuação máxima; alarmes no instante t recebem pontuação zero e são tratados como FP. Isso captura a noção de que antecipar

um gol com margem temporal tem valor prático superior a detectá-lo no momento exato. Durante o desenvolvimento, foi avaliado também um protocolo de janela simétrica, em que o gol atrasado é o centro da janela com K minutos para cada lado. Essa formulação foi descartada porque atribuía o mesmo peso a alarmes tardios e alarmes antecipados, inflando artificialmente os resultados. Trabalhar o *SoftED* com assimetria corrige esse viés, alinhando a avaliação à evidência da literatura de que as variações táticas antecedentes a um gol ocorrem em janelas temporais curtas [Lang et al., 2025].

Por fim, até onde foi possível identificar na literatura indexada pela Scopus em abril de 2026, nenhum trabalho anterior aplicou detecção de *concept drift* não supervisionado para detecção de eventos esportivos em partidas de futebol, o que posiciona esta dissertação como uma contribuição inicial nessa direção. A literatura existente de detecção de *drift* se concentra majoritariamente em domínios como monitoramento industrial e médico [Hinder et al., 2024; Tavares et al., 2025; Kore et al., 2024]: enquanto Tavares et al. [2025] detectam *drift* em dados de poços de petróleo e Kore et al. [2024] identificam a emergência da COVID-19 via mudanças na distribuição de imagens radiológicas, nenhum desses trabalhos aborda a detecção de eventos táticos em tempo real durante uma partida. A aplicação ao futebol introduz requisitos específicos como curta duração do fenômeno, ambiente adversarial com dois agentes simultâneos, e ausência de rótulos durante o jogo, que distinguem este trabalho das aplicações existentes e abrem uma direção de pesquisa ainda pouco explorada.

Capítulo VI Considerações Finais

Esta dissertação investigou a seguinte pergunta: em que medida detectores de *concept drift* não supervisionados, aplicados a séries temporais de passes intra-partida, são capazes de antecipar a ocorrência de gols no futebol profissional? Para respondê-la, três detectores foram avaliados (Page-Hinkley, KSWIN e ADWIN) e comparados a dois *baselines* sem aprendizado (alarme fixo e alarme aleatório), sobre todas as partidas da temporada 2015/2016 da La Liga. A avaliação adotou uma variante assimétrica do *SoftED* [Salles et al., 2024], com janela de tolerância $K = 10$ minutos, e reportou precisão, *recall*, F_1, $F_{0,5}$ e MCC para cada configuração. Os resultados indicam que a resposta é parcialmente afirmativa: Page-Hinkley e KSWIN superaram o *baseline* aleatório em MCC, demonstrando que mudanças estatísticas no volume de passes carregam sinal preditivo real, enquanto o ADWIN não superou o acaso com as configurações testadas.

Em termos práticos, considerando a melhor configuração *in-sample* ($W = 3$, $C = 3$, *feature*: passe), o detector Page-Hinkley cobriu 51,3% dos 916 gols *open play* da temporada, aproximadamente 470 eventos, dentro da janela de tolerância de $K = 10$ minutos. Em outras palavras, em aproximadamente 5 de cada 10 gols o sistema teria emitido um aviso prévio ao treinador com até 10 minutos de antecedência, horizonte suficiente para intervenções táticas como substituições ou reorganizações de marcação. Os 48,7% restantes correspondem a gols não antecipados, parte dos quais é estruturalmente irrecuperável: na melhor configuração ($W = 3$), o detector entra em fase ativa a partir do terceiro minuto de cada período, e gols anteriores a esse ponto resultam em FN garantidos por ausência de aquecimento. Adicionalmente, gols ocorridos entre o minuto W e o minuto $K = 10$ têm sua janela de avaliação $[t - K, t]$ parcialmente truncada pelo início do período, reduzindo o *score* máximo alcançável. Na base utilizada, 196 dos 916 gols *open play* (21,4%) ocorrem antes do minuto $K = 10$ de cada período.

Esse resultado, obtido na melhor configuração *in-sample*, é alcançado sem dados rotulados e desde a primeira partida analisada. A robustez da abordagem é verificada pelo protocolo *out-of-sample* com divisão temporal 190/190, no qual a configuração selecionada no conjunto de treino é aplicada diretamente ao conjunto de teste sem re-otimização.

Os resultados *out-of-sample* confirmam a liderança do Page-Hinkley: MCC $= 0{,}126$ e $F_1 = 0{,}093$ no conjunto de teste, contra MCC $= 0{,}065 \pm 0{,}0004$ e $F_1 = 0{,}039 \pm 0{,}0001$ do *baseline* aleatório. O KSWIN também supera o acaso no conjunto de teste (MCC $= 0{,}088$, $F_1 = 0{,}089$), mas não mantém

ganho sobre o Page-Hinkley. O detector Page-Hinkley apresentou o melhor MCC entre os modelos avaliados, com configuração de janela curta ($W = 3$) e *cooldown* moderado ($C = 3$), o que sugere que mudanças abruptas no volume de passes são o sinal mais discriminativo disponível nesta base. Essa preferência por janelas curtas sugere que o fenômeno subjacente se aproxima do *Sudden Drift* na taxonomia de Lu et al. [2020]: a mudança na distribuição de passes ocorre de forma abrupta e em curto intervalo de tempo, em contraste com variações graduais que exigiriam janelas mais longas para serem detectadas.

Considerando o desempenho *in-sample* na melhor configuração por time/tarefa (teto otimista), o Page-Hinkley atinge MCC = 0,147 e $F_1 = 0{,}123$, mantendo a liderança sobre todos os modelos. O KSWIN também supera o *baseline* aleatório neste protocolo (MCC = 0,118 vs. 0,069) e supera o Baseline Fixo em ambas as métricas (MCC = 0,118 vs. 0,037; $F_1 = 0{,}125$ vs. 0,050). O ADWIN disparou apenas 78 alarmes em toda a temporada no *in-sample* e registrou o pior desempenho absoluto (MCC = 0,031, $F_1 = 0{,}014$), indicando que sua parametrização ótima neste contexto é excessivamente conservadora para cobrir eventos tão raros quanto gols.

Em termos de contribuições, esta dissertação cumpriu os três objetivos anunciados na Introdução. A primeira contribuição, a formulação do problema de detecção antecipada de gols como detecção de *drift* virtual em séries de passes intra-partida, foi operacionalizada no *pipeline* descrito no Capítulo IV e validada empiricamente pelos resultados do Capítulo V. A segunda contribuição, a adaptação assimétrica do *SoftED* para avaliação de alarmes antecipados, introduziu uma variante metodológica que penaliza alarmes tardios e valoriza detecções temporalmente próximas ao evento, alinhando a avaliação ao requisito operacional de antecedência tática. A terceira contribuição, a avaliação empírica de três detectores não supervisionados sobre dados reais de futebol profissional, revelou que Page-Hinkley e KSWIN são capazes de superar o acaso de forma consistente, enquanto o ADWIN apresentou limitações estruturais no contexto intra-partida. Juntas, essas contribuições posicionam a detecção de *drift* como uma abordagem viável e interpretável para a antecipação de eventos esportivos, com potencial de extensão a outros domínios onde rótulos em tempo real são escassos.

VI.1 Delimitações Metodológicas

Esta seção registra quatro delimitações metodológicas, apresentadas da mais ampla à mais específica. As duas primeiras dizem respeito ao escopo da avaliação; as duas últimas, a restrições inerentes ao método de pontuação adotado.

O *pipeline* foi avaliado exclusivamente sobre a temporada 2015/2016 da La Liga, o que restringe a generalização dos resultados. Padrões táticos, estilos de jogo e frequência de gols podem variar entre temporadas, ligas e contextos competitivos distintos, e não é possível afirmar que a relação

entre mudanças no padrão de passes e ocorrência de gols se mantém estável em outros campeonatos ou períodos.

O *pipeline* monitora exclusivamente o volume de passes como *feature* de entrada. Essa escolha é sustentada tanto pela análise exploratória — passes apresentaram a maior correlação com ocorrência de gols entre todos os tipos de evento avaliados ($r = 0{,}30$, Figura IV.5) — quanto pela adequação estatística ao método: com 53,6% de todos os eventos táticos da temporada, passes produzem séries temporais densas o suficiente para que detectores baseados em acumulação e comparação de distribuições operem de forma estável em janelas curtas. Eventos menos frequentes, como recuperações de bola (5,7%) ou dribles (2,5%), gerariam séries esparsas com médias móveis próximas de zero na maioria dos minutos, comprometendo a capacidade de detecção.

O terceiro ponto refere-se aos gols ocorridos em janelas de avaliação incompletas, situação que afeta dois subconjuntos distintos. Gols nos primeiros W minutos de cada período são excluídos da avaliação. Como o *warmup* impede estruturalmente qualquer disparo nesse intervalo, independentemente do sinal, classificar esses eventos como FN seria injusto: a ausência de alarme não reflete falha do detector, mas uma restrição operacional do método. Por isso, esses minutos e os gols neles contidos são removidos do cálculo das métricas. Gols entre o minuto W e o minuto K estão sujeitos a uma penalização parcial: o alarme pode ser emitido, mas a janela $[t - K,\ t]$ é truncada pelo início do período, reduzindo o *score* máximo alcançável abaixo de 1,0. Na base utilizada, 196 dos 916 gols *open play* (21,4%) ocorrem antes do minuto $K = 10$ do respectivo período e, portanto, sob algum grau de truncagem, o que deprime sistematicamente o *recall* de todos os modelos.

A quarta consideração metodológica refere-se ao comportamento do *matching* em segmentos com janelas sobrepostas. Quando dois gols consecutivos ocorrem com intervalo inferior a K minutos, suas janelas $[t-K,\ t]$ se fundem em um único segmento, e é realizado um *matching* 1:1 entre alarmes e gols nesse segmento. A consequência é dupla: se um único alarme cair na região de sobreposição, ele é atribuído ao gol de maior pontuação e o outro gol torna-se FN, mesmo que o alarme estivesse temporalmente compatível com ambos; se dois alarmes cobrirem o mesmo gol, apenas o de melhor pontuação é atribuído ao evento e o segundo torna-se FP, reduzindo a precisão. Na temporada analisada, 44,6% dos 381 pares de gols consecutivos no mesmo período ocorrem com intervalo igual ou inferior a $K = 10$ minutos (mediana 11,0 min, média 13,2 min), tornando esse cenário frequente. Esse comportamento é inerente ao *SoftED* e ao requisito de não dupla contagem, mas deve ser considerado na interpretação dos valores de *recall* e MCC reportados.

VI.1.1 Sobre o Uso do Termo *Concept Drift*

Na definição formal de Gama et al. [2014], *concept drift* pressupõe a existência de um modelo preditivo cujo desempenho degrada à medida que a distribuição conjunta $P(X, y)$ evolui, o que

motiva a reestimação ou adaptação do modelo. Neste trabalho, os detectores operam exclusivamente sobre $P(X)$ sem acesso à variável alvo y, o que suscita uma discussão sobre como categorizar a abordagem.

Uma interpretação possível é enquadrá-la como um classificador dinâmico: um sistema que adapta seu comportamento em função de mudanças na distribuição dos dados ao longo do tempo. Nessa perspectiva, a seleção dinâmica de classificadores tem sido proposta diretamente como mecanismo de resposta ao *concept drift* em ambientes não-estacionários [Almeida et al., 2016; Cruz et al., 2018], explorando a competência local dos modelos em relação ao conceito corrente para adaptar as predições sem supervisão contínua.

Por outro lado, uma linha crescente na literatura estabelece o uso de algoritmos de detecção de *concept drift* diretamente como detectores de eventos em séries temporais, sem modelo preditivo subjacente [Hinder et al., 2024; Tavares et al., 2025; Kore et al., 2024]. Nesses trabalhos, a ruptura em $P(X)$ é o evento de interesse, não um sintoma de degradação de modelo, mas um sinal operacional independente. Kore et al. [2024] demonstram explicitamente que monitorar a performance do modelo não é substituto confiável para detectar drift nos dados: a detecção baseada em $P(X)$ capturou a emergência da COVID-19 em radiografias quando o monitoramento de performance do modelo falhou. O presente trabalho segue essa mesma perspectiva: os detectores são empregados como instrumentos de monitoramento de séries temporais, e a denominação *concept drift* reflete o campo de origem das técnicas e o alinhamento com essa corrente da literatura, não uma afirmação de que há modelo preditivo sendo adaptado.

VI.1.2 Sobre Implantação em Tempo Real

Uma direção relevante é a avaliação do custo computacional e da viabilidade de implantação em tempo real. O *pipeline* implementado processa cada minuto de jogo de forma incremental: os detectores Page-Hinkley, KSWIN e ADWIN operam em $O(1)$ por observação, e a média móvel de janela W mantém apenas os últimos W valores em memória. A avaliação *SoftED* tem custo $O(n \cdot m)$ por período, onde n é o número de alarmes e m o número de gols, tipicamente pequenos em uma partida.

O *grid search* completo sobre 380 partidas e três valores de $K \in \{5, 10, 15\}$ foi executado em hardware convencional (Apple MacBook Air, chip M1, 8 GB RAM) com paralelização em 7 *workers*. Os tempos acumulados por detector foram: Page-Hinkley $\approx$ 53 minutos, KSWIN $\approx$ 90 minutos (detector mais custoso pelo maior espaço de hiperparâmetros), ADWIN $\approx$ 7 minutos e *baselines* $<$ 5 minutos, totalizando aproximadamente 2,5 horas de processamento. Os dados utilizados são públicos e obtidos gratuitamente via repositório StatsBomb Open Data; em um cenário de uso profissional, o custo de obtenção de dados ao vivo de provedores como StatsBomb ou Opta representa

uma barreira de acesso relevante a considerar.

Em um cenário de tempo real, o *pipeline* precisaria ser alimentado por um *feed* de eventos ao vivo, agregando eventos minuto a minuto e disparando o detector ao final de cada intervalo, com latência máxima de 1 minuto entre o evento real e o alarme. A avaliação formal desse cenário, incluindo latência de ingestão, custo de licenciamento de dados e integração com APIs de eventos ao vivo, constitui uma direção natural para trabalhos futuros.

Por fim, vale ressaltar uma limitação inerente à aplicação prática em tempo real. Uma vez implantado, o sistema pode induzir intervenções táticas (substituições, reorganizações defensivas) sempre que um alarme é disparado. Se essas intervenções forem eficazes, o gol previsto não ocorre, tornando impossível distinguir se o alarme foi um falso positivo ou se a intervenção o impediu. Esse paradoxo de causalidade, em que o sucesso do preditor apaga a evidência de sua própria acurácia, é um desafio conhecido em sistemas de suporte à decisão em tempo real e deve ser considerado no desenho de qualquer avaliação prospectiva do *pipeline*.

VI.2 Trabalhos Futuros

A análise de sensibilidade do parâmetro K constitui a direção mais imediata de possível trabalho futuro. Como 21,4% dos gols ocorrem antes do minuto $K = 10$ de cada período, esses eventos ficam parcialmente ou totalmente fora da janela de avaliação $[t - K, t]$, deprimindo sistematicamente o *recall* de todos os modelos. Avaliar o *pipeline* com valores entre $K = 5$ e $K = 10$ permitiria quantificar o impacto dessa truncagem e orientar a escolha do horizonte em função do contexto de aplicação: intervenção tática imediata ou análise *post-hoc*.

A exploração de múltiplas resoluções temporais constitui outra direção relevante. O *pipeline* atual agrega eventos em janelas de um minuto; resoluções menores, como 30 segundos, poderiam capturar mudanças mais abruptas com maior antecedência, ao custo de maior esparsidade na série e necessidade de suavização mais intensa. A escolha da resolução temporal interage diretamente com os parâmetros W e K, e sua otimização conjunta constitui uma extensão natural do presente trabalho.

Uma direção promissora seria adotar a posse de bola como unidade de análise em substituição à janela temporal fixa. A partir dos dados disponíveis — que registram, para cada passe, o executor, o receptor e o resultado — seria possível reconstruir posses operacionalmente como sequências ininterruptas de passes bem-sucedidos de um mesmo time. Os detectores de *concept drift* seriam então aplicados sobre as variações nas características dessas sequências ao longo da partida, como número de passes por posse, taxa de sucesso e padrão de circulação. Essa abordagem poderia capturar transições táticas abruptas — como a adoção de um estilo de contra-ataque, caracterizado por posses curtas e diretas — que uma janela temporal fixa tenderia a diluir por agregar sequências

de naturezas distintas.

A adoção de protocolos de avaliação mais robustos, como validação cruzada temporal com múltiplos blocos ou *rolling window* com re-calibração partida a partida, permitiria estimar o desempenho do *pipeline* com menor variância e verificar se a configuração ótima se mantém estável ao longo da temporada, o que é relevante dado que times evoluem taticamente entre rodadas.

Uma outra direção refere-se à robustez da calibração de hiperparâmetros. Neste trabalho, a melhor configuração foi selecionada de forma global (um único conjunto de parâmetros para todos os times e tarefas) ou por time/tarefa individualmente (*in-sample*). Uma avaliação com validação cruzada temporal sobre múltiplas temporadas permitiria investigar se uma configuração global robusta, treinada em temporadas anteriores, generaliza para temporadas futuras sem re-otimização, aproximando o *pipeline* de um cenário de uso prático.

A extensão para múltiplas temporadas e ligas constitui um desdobramento relevante. Avaliar o *pipeline* sobre outras temporadas da La Liga e sobre competições com características táticas distintas, como a Bundesliga utilizada por Lang et al. [2025], permitiria verificar a robustez dos resultados além do contexto original e investigar se a relação entre mudanças no padrão de passes e ocorrência de gols é estável entre diferentes culturas táticas.

Incorporar indicadores compostos e detectores multivariados representa uma quarta linha de trabalho. *Features* como entradas no terço ofensivo ou posse de bola na área, combinadas com detectores multivariados de *drift*, podem aumentar a sensibilidade do *pipeline* a mudanças de ritmo que passes isolados não capturam, ao custo de maior complexidade computacional e interpretativa.

Uma nova contribuição diz respeito à homogeneidade temporal entre os dois períodos da partida. Caso a distribuição do tempo até o primeiro gol difira sistematicamente entre primeiro e segundo tempo, tratar ambos os períodos de forma indiferenciada pode introduzir viés na calibração dos detectores. Investigar essa assimetria e, se confirmada, adaptar a metodologia para estimar parâmetros separados por período constituiria um refinamento relevante tanto para a avaliação quanto para o uso prático do *pipeline*.

Um caminho adicional consiste na implementação de um terceiro *baseline* baseado em limiar absoluto de passes. Em vez de detectar mudanças na distribuição, esse detector dispararia um alarme sempre que o volume de passes em uma janela ultrapassasse um limiar fixo, definido a partir da média histórica da temporada. Tal *baseline* permitiria isolar a contribuição do mecanismo de detecção de *drift* em relação a uma regra simples de intensidade, tornando a comparação entre abordagens mais granular e a análise de valor agregado do *pipeline* mais rigorosa.

Outra possibilidade de trabalho futuro é investigar a causalidade nas correlações identificadas na análise exploratória, em particular a relação entre volume de eventos (como passes) e ocorrência de gols. Como times em vantagem no placar tendem a ter maior posse de bola, há risco de causa-

lidade reversa que pode inflar artificialmente essas correlações. Recomenda-se o uso de técnicas de inferência causal ou modelos que controlem o estado do placar ao longo da partida, a fim de validar as premissas preditivas adotadas neste trabalho.

Avaliar o impacto da exclusão de gols de bola parada e pênaltis sobre a calibração dos parâmetros K e W do modelo é uma ramificação possível. Embora os gols open play representem 90,3% do total, não foi verificado se sua distribuição temporal ao longo das partidas é semelhante à dos gols excluídos. Caso essas distribuições difiram sistematicamente, a calibração dos parâmetros pode estar enviesada. Recomenda-se, portanto, conduzir uma análise comparativa das distribuições temporais entre os dois subconjuntos, bem como testar variantes do modelo que incorporem ou controlem diferentes tipos de gol.

Por fim, a avaliação do impacto do mecanismo de *cooldown* sobre as métricas reportadas constitui uma direção metodológica relevante. O *cooldown* suprime alarmes nos C minutos seguintes a cada disparo, o que reduz artificialmente o volume de falsos positivos e pode inflar levemente a precisão e o $F_{0,5}$. Replicar a avaliação sem supressão pós-alarme permitiria quantificar esse efeito e separar o ganho atribuível ao detector em si do ganho atribuível ao mecanismo de silenciamento.

Referências

Almeida, C. H., Ferreira, A. P., and Volossovitch, A. (2014). Effects of match location, match status and quality of opposition on regaining possession in uefa champions league. *Journal of Human Kinetics*, 41:203–214.

Almeida, P. R. L. d., Oliveira, L. S., Britto Jr., A. d. S., and Sabourin, R. (2016). Handling concept drifts using dynamic selection of classifiers. In *2016 IEEE 28th International Conference on Tools with Artificial Intelligence*, pages 989–994.

Alves, R. (2025). Score: A convolutional approach for football event forecasting. *International Journal of Forecasting*.

Anwar, M. Q. H. K., Jamaludin, M., Razali, M. R. M., Talip, N. K. A., and Ismail, Z. (2022). Offensive and defensive team performances: relation to successful and unsuccessful participation in the uefa champions league 20/21. *Journal of Physical Education and Sport*, 22:2649–2654.

Arntzen, H. and Hvattum, L. M. (2021). Predicting match outcomes in association football using team ratings and player ratings. *Statistical Modelling*, 21:449–470.

Berrar, D., Lopes, P., and Dubitzky, W. (2019). Incorporating domain knowledge in machine learning for soccer outcome prediction. *Machine Learning*, 108:97–126.

Bifet, A. and Gavaldà, R. (2007). *Learning from Time-Changing Data with Adaptive Windowing*, pages 443–448.

Bunker, R. and Susnjak, T. (2022). The application of machine learning techniques for predicting match results in team sport: A review.

Capobianco, G., Giacomo, U. D., Mercaldo, F., Nardone, V., and Santone, A. (2019). Can machine learning predict soccer match results? In *ICAART 2019 - Proceedings of the 11th International Conference on Agents and Artificial Intelligence*, volume 2, pages 458–465. SciTePress.

Chacón-Fernández, E., Brunsó-Costal, G., Duarte, A., Sánchez-Oro, J., and Alonso-Pérez-Chao, E. (2025). Home advantage in football: Exploring its effect on individual performance. *Applied Sciences (Switzerland)*, 15.

Cruz, R. M. O., Sabourin, R., and Cavalcanti, G. D. C. (2018). Dynamic classifier selection: Recent advances and perspectives. *Information Fusion*, 41:195–216.

de Souza, D. B., Campo, R. L.-D., Blanco-Pita, H., Resta, R., and Coso, J. D. (2019). An extensive comparative analysis of successful and unsuccessful football teams in laliga. *Frontiers in Psychology*, 10.

Decroos, T., Haaren, J. V., Bransen, L., and Davis, J. (2019). Actions speak louder than goals: Valuing player actions in soccer. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1851–1861. Association for Computing Machinery.

Delgado-Bordonau, J. L., Domenech-Monforte, C., Guzmán, J. F., and Mendez-Villanueva, A. (2013). Offensive and defensive team performance: Relation to successful and unsuccessful participation in the 2010 soccer world cup. *Journal of Human Sport and Exercise*, 8:894–904.

Dobson, S. and Goddard, J. (2017). Evaluating probabilities for a football in-play betting market. In Rodríguez, P., Humphreys, B. R., and Simmons, R., editors, *The Economics of Sports Betting*, pages 52–70. Edward Elgar Publishing, Cheltenham, UK.

Dutta, A., Saikia, H., Gogoi, J., and Bhattacharjee, D. (2024). Forecasting the opening goal in second-half of a football match: Bayesian and frequentist perspectives. *Computational Statistics*.

Forcher, L., Forcher, L., Altmann, S., Jekauc, D., and Kempe, M. (2024). The keys of pressing to gain the ball–characteristics of defensive pressure in elite soccer using tracking data. *Science and Medicine in Football*, 8:161–169.

Gama, J., Zliobaite, I., Bifet, A., Pechenizkiy, M., and Bouchachia, A. (2014). A survey on concept drift adaptation.

Goka, R., Moroto, Y., Maeda, K., Ogawa, T., Shih, H.-C., and Haseyama, M. (2024). Masked modeling-based action event prediction considering bidirectional time-series in soccer. page 118. SPIE-Intl Soc Optical Eng.

Hinder, F., Vaquet, V., and Hammer, B. (2024). One or two things we know about concept drift—a survey on monitoring in evolving environments. Part A: detecting concept drift. *Frontiers in Artificial Intelligence*, 7:1330257.

Hoens, T. R., Polikar, R., and Chawla, N. V. (2012). Learning from streaming data with concept drift and imbalance: An overview.

Hubáček, O., Šourek, G., and Železný, F. (2022). Forty years of score-based soccer match outcome prediction: An experimental review.

Iwashita, A. S. and Papa, J. P. (2019). An overview on concept drift learning. *IEEE Access*, 7:1532–1547.

Kore, A., Abbasi Bavil, E., Subasri, V., Abdalla, M., Fine, B., Dolatabadi, E., and Abdalla, M. (2024). Empirical data drift detection experiments on real-world medical imaging data. *Nature Communications*, 15:1887.

Lago-Peñas, C., Lago-Ballesteros, J., and Rey, E. (2011). Differences in performance indicators between winning and losing teams in the uefa champions league. *Journal of Human Kinetics*, 27:135–146.

Lang, S., Wimmer, T., Erben, A., and Link, D. (2025). Which indicators matter? using performance indicators to predict in-game success-related events in association football. *International Journal of Computer Science in Sport*, 24:16–44.

Li, B. and Müller, E. (2022). STAD: State-Transition-Aware anomaly detection under concept drifts. In *Proceedings of the First Workshop on Online Learning from Uncertain Data Streams (OLUD 2022)*, volume 3380 of *CEUR Workshop Proceedings*.

Liu, T., de Alcaraz, A. G., Wang, H., Hu, P., and Chen, Q. (2021). Impact of scoring first on match outcome in the chinese football super league. *Frontiers in Psychology*, 12.

Lu, J., Liu, A., Dong, F., Gu, F., Gama, J., and Zhang, G. (2020). Learning under concept drift: A review.

Lühr, S. and Lazarescu, M. (2007). A visual data analysis tool for sport player performance benchmarking, comparison and change detection. In *Proceedings - International Conference on Tools with Artificial Intelligence, ICTAI*, volume 1, pages 289–296.

Mendes-Neves, T., Meireles, L., and Mendes-Moreira, J. (2024). Towards a foundation large events model for soccer. *Machine Learning*, 113:8687–8709.

Mendes-Neves, T., Meireles, L., and Mendes-Moreira, J. (2026). A scalable approach for unified large events models in soccer. volume 16022, page 354 – 371. Cited by: 0.

Page, E. S. (1954). Continuous inspection schemes. *Biometrika*, 41(1/2):100–115.

Pappalardo, L., Cintia, P., Rossi, A., Massucco, E., Ferragina, P., Pedreschi, D., and Giannotti, F. (2019). A public data set of spatio-temporal match events in soccer competitions. *Scientific Data*, 6.

Purucker, M. C. (1996). Neural network quarterbacking. *IEEE Potentials*, 15(3):9 – 15. Cited by: 36.

Raab, C., Heusinger, M., and Schleif, F.-M. (2020). Reactive soft prototype computing for concept drift streams. *Neurocomputing*, 416.

Reed, D. and O'Donoghue, P. (2005). Development and application of computer-based prediction methods. *International Journal of Performance Analysis in Sport*, 5:12–28.

Robberechts, P., Haaren, J. V., and Davis, J. (2019). Who will win it? an in-game win probability model for football. *MLSA – Machine Learning in Sports Analytics*.

Romero, R., Mashayekhi, Y., Lai, F., Van Roy, M., De Bie, T., and Davis, J. (2026). Next-event prediction in soccer: Assessing the impact of team and player information. volume 2833 CCIS, page 62 – 73. Cited by: 0.

Salles, R., Lima, J., Reis, M., Coutinho, R., Pacitti, E., Masseglia, F., Akbarinia, R., Chen, C., Garibaldi, J., Porto, F., and Ogasawara, E. (2024). SoftED: Metrics for soft evaluation of time series event detection. *Computers and Industrial Engineering*, 198.

Sarmento, H., Figueiredo, A., Lago-Peñas, C., Milanovic, Z., Barbosa, A., Tadeu, P., and Bradley, P. S. (2018). Influence of tactical and situational variables on offensive sequences during elite football matches. Technical report.

Simpson, I., Beal, R. J., Locke, D., and Norman, T. J. (2022). Seq2event: Learning the language of soccer using transformer-based match event prediction. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 3898–3908. Association for Computing Machinery.

StatsBomb (2019). *Open Data Events v4.0.0*. StatsBomb. Acesso em: abril de 2026.

Sujon, K. M., Hassan, R., Choi, K., and Samad, M. A. (2025). Accuracy, precision, recall, F1-score, or MCC? Empirical evidence from advanced statistics, ML, and XAI for evaluating business predictive models. *Journal of Big Data*, 12:268.

Tavares, L. G., Lima, J., Melo, M., Chen, C., Garibaldi, J., Scatena, G. d. S., Costa, A. H. R., Gomi, E. S., Salles, R., Pacitti, E., Santos, I., Siqueira, I. G. a., Carvalho, D., Coutinho, R., Porto, F., and Ogasawara, E. (2025). Fuzzy-based ensemble method for robust concept drift detection in multivariate time series. In *2025 International Joint Conference on Neural Networks (IJCNN)*.

Umemoto, R., Tsutsui, K., and Fujii, K. (2025). A generalized valuation method for team defense by estimating probabilities in football games. In *Proceedings of the 13th International Conference on Sport Sciences Research and Technology Support (icSPORTS 2025)*, pages 79–89.

van Rijsbergen, C. J. (1979). *Information Retrieval*. Butterworth-Heinemann, London, 2 edition.

Webb, G. I., Hyde, R., Cao, H., Nguyen, H. L., and Petitjean, F. (2016). Characterizing concept drift.

Wiechno, W., Bartosik, B., and Duch, P. (2025). Time series and deep learning approaches for predicting english premier league match outcomes. In *International Conference on Agents and Artificial Intelligence*, volume 3, pages 789–796. Science and Technology Publications, Lda.

Yao, W., Wang, Y., Zhu, M., Cao, Y., and Zeng, D. (2022). Goal or miss? a bernoulli distribution for in-game outcome prediction in soccer. *Entropy*, 24.

Yeung, C., Ide, K., Someya, T., and Fujii, K. (2025). Openstarlab: open approach for spatio-temporal agent data analysis in soccer. *Complex and Intelligent Systems*, 11.

TERMO DE AUTORIZAÇÃO PARA PUBLICAÇÃO/DIVULGAÇÃO DE DOCUMENTO ELETRÔNICO NO CATÁLOGO DO SISTEMA DE BIBLIOTECAS DO CEFET/RJ E NO REPOSITÓRIO INSTITUCIONAL

Na qualidade de titular dos direitos de autor da publicação abaixo citada, de acordo com a Lei nº 9610/98, autorizo, a partir da presente data, o CEFET/RJ, a disponibilizar gratuitamente, sem ressarcimento de direitos autorais, conforme permissões assinadas abaixo, no Catálogo Online do Sistema de Bibliotecas e no Repositório Institucional do Cefet/RJ, ou qualquer outro formato de disseminação da informação que a instituição implemente, no formato PDF, para fins de leitura, impressão e/ou download pela internet, a título de divulgação da produção científica gerada por esta Instituição.

1- Tipo de trabalho: [ ] Projeto Final (Graduação) [ ] Especialização
[X] Dissertação [ ] Tese

2- Identificação do trabalho/autor(es):

Curso: Mestrado em Ciência da Computação (PPCIC) _____
Título: Aplicação de Métodos Baseados em Concept Drift para Previsão de Gols no Futebol Profissional _____
Autor(es): Ana Gabriela Viana de Araújo _____
Orientador: Jorge de Abreu Soares _____
Co-orientador:_____

3- Informações de acesso ao trabalho:

Este trabalho é confidencial? [ ] Sim [X] Não
Pode ser liberado para publicação? [X] Total [ ] Parcial [ ] Não
Justifique:_____
_____
Em caso de publicação parcial, assinale as permissões:
[ ] Sumário
[ ] Capítulos. Especifique: _____
[ ] Bibliografia
[ ] Outras partes do trabalho: _____

Rio de Janeiro, 18/06/2026

Documento assinado digitalmente
gov.br
ANA GABRIELA VIANA DE ARAUJO
Data: 18/06/2026 09:22:44-0300
Verifique em https://validar.iti.gov.br

_____
Assinatura do(s) autor(es)

Documento assinado digitalmente
gov.br
JORGE DE ABREU SOARES
Data: 18/06/2026 10:18:31-0300
Verifique em https://validar.iti.gov.br

_____
Assinatura do orientador