

A Utilização de Inteligência Artificial para Otimização da Assertividade em Contratações no Futebol

Davi Carvalho dos Santos

*Departamento de Ciência da Computação
Universidade Federal de Minas Gerais (UFMG)
Belo Horizonte, Brasil
E-mail: davicsantos010@ufmg.br*

Orientador: Adriano César Machado Pereira

*Departamento de Ciência da Computação
Universidade Federal de Minas Gerais (UFMG)
Belo Horizonte, Brasil
E-mail: adrianoc@dcc.ufmg.br*

Resumo—Contratações no futebol envolvem incerteza técnica, contextual e financeira, o que torna relevante o desenvolvimento de instrumentos reproduzíveis de apoio ao scouting. Este trabalho implementa e avalia retrospectivamente um framework híbrido que integra eventos do StatsBomb Open Data, métricas por jogador, identificação de perfis táticos com K-Means, valores históricos do Transfermarkt, um score técnico-econômico interpretável, regras de red flags e um experimento supervisionado complementar com Random Forest. A engenharia de dados produziu 4.406 registros jogador-competição e 3.551 atletas únicos, 2.369 jogadores com ao menos 300 minutos foram utilizados na análise de clusters, e um subconjunto recente de 617 atletas foi ranqueado. A validação utilizou 63 transferências realizadas em 2023 e um desfecho retrospectivo predominantemente econômico, baseado na variação do valor de mercado e na relação com a taxa paga. O score completo apresentou MAE de 16,18, acurácia balanceada de 0,382, macro-F1 de 0,346 e correlação de Spearman de 0,196. A coerência da política foi 82,54%, mas um baseline que sempre não prioriza atingiu 84,13%, evidenciando o viés conservador dessa métrica. Em bootstrap pareado, o ganho de macro-F1 do score completo sobre esse baseline foi 0,113, com IC de 95% de 0,018 a 0,201 e superioridade em 98,9% das reamostragens. O Random Forest obteve MAE fora da amostra de 15,95, pior que o baseline de média por partição (15,27), e não demonstrou superioridade preditiva. A solução com seis clusters foi altamente estável entre inicializações (ARI médio de 0,982), embora três e quatro clusters tenham apresentado métricas internas superiores. Os resultados sustentam o uso do framework como instrumento explicável de triagem e organização de evidências, e não como mecanismo autônomo de decisão ou previsão integral do sucesso esportivo.

Abstract—Football transfers involve technical, contextual and financial uncertainty, making reproducible decision-support tools relevant for scouting departments. This work implements and retrospectively evaluates a hybrid framework that integrates StatsBomb Open Data events, player-level metrics, tactical profile identification with K-Means, historical Transfermarkt market values, an interpretable technical-economic score, rule-based red flags and a complementary

supervised Random Forest experiment. The data engineering stage produced 4,406 player-competition records and 3,551 unique players, 2,369 players with at least 300 minutes were used for cluster analysis, and a recent subset of 617 players was ranked. The retrospective validation used 63 transfers from 2023 and an economic outcome proxy based on market value variation and transfer fee efficiency. The complete score achieved MAE of 16.18, balanced accuracy of 0.382, macro-F1 of 0.346 and Spearman correlation of 0.196. The results support the framework as an explainable screening and evidence-organization tool rather than an autonomous mechanism for transfer success prediction.

Index Terms—Inteligência artificial no futebol, scouting baseado em dados, K-Means, Random Forest, IA explicável, validação retrospectiva, apoio à decisão.

1. Introdução

O mercado de transferências concentra decisões de grande impacto esportivo e financeiro. Uma contratação inadequada pode gerar custo de aquisição, salários, perda de espaço no elenco e redução do valor de revenda, enquanto uma contratação bem-sucedida pode produzir desempenho competitivo e valorização patrimonial. Apesar desse impacto, a avaliação de atletas ainda combina dados incompletos, observação humana e pressões de mercado, estando sujeita a vieses e a comparações pouco padronizadas entre ligas e funções.

A primeira etapa desta pesquisa estruturou conceitualmente um processo híbrido, inspirado na lógica de identificação de ineficiências de mercado popularizada por *Moneyball* [2] e adaptada às particularidades do futebol [1]. A proposta não buscava substituir olheiros e dirigentes, mas organizar evidências quantitativas para ampliar a triagem e tornar recomendações auditáveis. A segunda etapa implementou esse fluxo com dados públicos e avaliou suas limitações em transferências reais.

A questão central investigada neste trabalho consiste em compreender em que medida um framework híbrido, baseado em desempenho técnico, perfil tático e informação

econômica, pode apoiar a triagem retrospectiva de contratações no futebol. As questões de pesquisa que orientam a análise são:

- **RQ1** O score completo apresenta resultados superiores a baselines simples?
- **RQ2** Os perfis obtidos por clusterização são estáveis e interpretáveis?
- **RQ3** A decomposição das recomendações permite compreender acertos e divergências?

As contribuições deste trabalho são:

- Implementação de um pipeline reproduzível para transformar eventos de jogo em métricas por atleta;
- Identificação de perfis táticos por aprendizado não supervisionado;
- Integração entre dados de desempenho e informações econômicas;
- Definição de um score técnico-econômico interpretável;
- Validação retrospectiva em transferências reais;
- Comparação com baselines, ablações, bootstrap pareado e Random Forest complementar.

O restante do artigo é organizado da seguinte forma. A Seção II apresenta o referencial teórico, a Seção III descreve a metodologia, a Seção IV detalha a implementação, a Seção V apresenta os resultados, a Seção VI discute os achados, a Seção VII reúne as limitações e a Seção VIII apresenta as conclusões e os trabalhos futuros.

2. Referencial Teórico

Esta seção apresenta os conceitos que fundamentam o framework proposto. São discutidos o scouting orientado por dados, as métricas técnicas utilizadas para representar o desempenho dos atletas, o uso de modelos supervisionados e informações econômicas na avaliação de jogadores, além da importância de mecanismos de explicabilidade em sistemas de apoio à decisão no futebol.

2.1. Scouting orientado por dados

A análise de dados aplicada ao esporte ganhou visibilidade com a lógica do *Moneyball*, que mostrou como estatísticas podem revelar ineficiências de mercado e apoiar decisões de recrutamento [2]. No futebol, a aplicação dessa lógica é mais complexa devido à natureza contínua, coletiva e contextual do jogo. Ainda assim, a literatura argumenta que muitos padrões relevantes são mensuráveis e que vieses de interpretação podem afetar a avaliação de jogadores [1]. Assim, a análise quantitativa deve ser entendida como complemento ao conhecimento especializado, não como substituição da decisão humana.

2.2. Métricas técnicas e perfis táticos

Estatísticas tradicionais, como gols e assistências, capturam apenas parte da contribuição de um atleta. Métricas como *expected goals* (xG) ajudam a avaliar a qualidade das chances de finalização, reduzindo a dependência

do resultado final de cada jogada [3]. Além disso, ações como passes progressivos, conduções, dribles, pressões e ações defensivas permitem observar jogadores que contribuem na construção, recuperação e progressão da equipe. A partir desses vetores de métricas, técnicas de aprendizado não supervisionado podem identificar perfis funcionais que transcendem posições nominais, agrupando jogadores por semelhança de estilo [5].

2.3. Modelagem supervisionada e valoração econômica

Modelos supervisionados podem ser usados para prever desempenho, valor de mercado ou probabilidade de valorização futura, desde que exista uma base histórica suficiente e um alvo confiável. Métodos de ensemble, como Random Forest, são úteis por lidar com relações não lineares e interações entre variáveis técnicas, econômicas e contextuais. Porém, sua capacidade de generalização depende do volume e da representatividade dos casos disponíveis. Em mercados esportivos, a valoração financeira também é afetada por idade, contrato, reputação, liga de origem, narrativa midiática e contexto institucional, o que dificulta a definição de um único alvo de sucesso.

2.4. IA explicável e decisão híbrida

A adoção de modelos de IA em ambientes decisórios exige interpretabilidade. Em departamentos de futebol, recomendações opacas tendem a gerar resistência porque treinadores, scouts e dirigentes precisam entender os motivos da priorização. A área de IA explicável busca tornar modelos e recomendações compreensíveis para o usuário final [6]. Nesse trabalho, a explicabilidade é tratada principalmente por meio da decomposição do score em componentes técnicos, econômicos e de transferência, além da análise individual de casos e da importância de variáveis no experimento supervisionado.

3. Metodologia

Esta seção descreve o desenho metodológico adotado para transformar dados públicos em recomendações retrospectivas de contratação. O processo abrange a construção das métricas por atleta, a identificação de perfis táticos por K-Means, a integração com informações econômicas, o cálculo do score híbrido, o tratamento de red flags, o experimento supervisionado complementar e o protocolo de validação retrospectiva.

3.1. Visão geral do pipeline

O pipeline implementado segue uma arquitetura modular, na qual cada etapa transforma os dados de entrada em artefatos intermediários utilizados pela etapa seguinte. Primeiro, dados de eventos do StatsBomb Open Data são extraídos e convertidos em métricas agregadas por jogador.

Em seguida, essas métricas são normalizadas por tempo de participação e padronizadas para permitir comparações entre atletas com diferentes volumes de jogo e diferentes escalas de atributos. A partir desses vetores de desempenho, aplica-se o K-Means para identificação de perfis táticos, permitindo que jogadores sejam comparados dentro de grupos funcionais mais próximos, e não apenas por posição nominal.

Depois da etapa técnica, os jogadores são associados a dados econômicos do Transfermarkt, incluindo valores históricos de mercado e informações de transferências. Essa integração é necessária porque a decisão de contratação não depende apenas do desempenho em campo, mas também da atratividade econômica do atleta e da relação entre custo, valor de mercado e potencial de retorno. A combinação dessas dimensões resulta em um score híbrido de recomendação, composto por um componente técnico, um componente econômico e um componente relacionado à eficiência da transferência.

A avaliação foi organizada separadamente da geração do ranking. Essa separação é importante porque o ranking final representa a saída prática do sistema, enquanto a validação retrospectiva mede a coerência do método em transferências que já ocorreram. Para evitar vazamento de informação, a lógica retrospectiva utiliza dados disponíveis antes ou no momento simulado da decisão e compara a recomendação com um desfecho posterior, definido por uma proxy econômica. Assim, o pipeline não é estruturado como um mecanismo de decisão automática, mas como um fluxo reprodutível de construção de evidências, ranqueamento, comparação com baselines e análise crítica dos resultados.

A arquitetura também foi pensada para favorecer rastreabilidade. Como cada módulo gera saídas tabulares específicas, torna-se possível auditar separadamente a engenharia de métricas, a formação dos perfis táticos, o pareamento com dados econômicos, o cálculo do score e a avaliação retrospectiva. Essa organização facilita a identificação de erros, a reprodução dos experimentos e a análise individual de casos nos quais a recomendação do sistema diverge do desfecho observado.

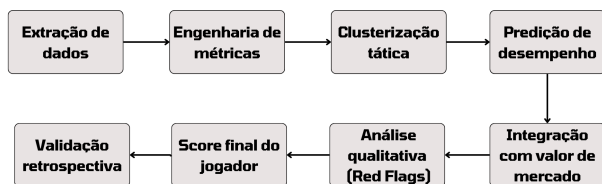


Figura 1. Pipeline metodológico implementado no protótipo.

3.2. Fontes de dados

Foram utilizadas duas fontes públicas principais. A primeira foi o StatsBomb Open Data, empregado para obter

eventos de jogo e construir métricas técnicas. A base oficial combinou oito recortes competitivos, incluindo a Copa do Mundo de 2018, a Euro 2020, a Copa do Mundo de 2022 e as cinco grandes ligas europeias da temporada 2015/2016. Essa combinação gerou 4.406 registros jogador-competição e 3.551 jogadores únicos antes dos filtros finais. Como a cobertura aberta é heterogênea, essa base não deve ser interpretada como painel longitudinal completo, mas como base reprodutível para implementação e avaliação do protótipo.

A segunda fonte foi o Transfermarkt, usado para obter valores históricos de mercado e informações de transferências. Como StatsBomb e Transfermarkt não compartilham identificadores únicos, foi necessário implementar um processo de pareamento por normalização de nomes, similaridade textual e auditoria de correspondências. O objetivo foi reduzir falsos pareamentos e associar cada atleta ao valor de mercado disponível antes da data de referência.

3.3. Engenharia de métricas

Eventos brutos foram convertidos em métricas por jogador. Foram consideradas variáveis ofensivas, de criação, progressão, posse e defesa, incluindo finalizações por 90 minutos, xG por 90 minutos, gols por 90 minutos, passes por 90 minutos, taxa de acerto de passe, passes progressivos, conduções progressivas, assistências para finalização, dribles, dribles completos, ações defensivas e pressões. A normalização por 90 minutos reduz distorções causadas por diferenças de tempo em campo, tornando comparáveis jogadores com volumes de participação distintos.

Além da normalização por tempo, as variáveis usadas em clusterização foram padronizadas com *z-score*. Essa etapa é necessária porque o K-Means utiliza distância entre vetores, ou seja, sem padronização, variáveis com escala naturalmente maior, como passes por 90, poderiam dominar variáveis com escala menor, como xG por 90. A padronização transforma cada métrica para uma escala centrada na média e medida em desvios-padrão, permitindo que o agrupamento reflita diferenças relativas, e não apenas magnitudes absolutas.

3.4. Identificação de perfis táticos com K-Means

A clusterização foi aplicada sobre os vetores de métricas de jogadores com pelo menos 300 minutos observados. O K-Means foi escolhido por ser simples, reprodutível, interpretável e adequado para descobrir agrupamentos sem rótulos prévios. O algoritmo aprende centroides que representam padrões médios de estilo e atribui cada jogador ao grupo mais próximo. A nomeação dos perfis finais foi realizada a partir das médias dos clusters, das posições predominantes e dos jogadores representantes. Os grupos resultantes foram laterais construtores e apoiadores, meio-campistas de pressão e recuperação, defensores de contenção, pontas e meias criativos, goleiros e centroavantes e finalizadores.

A escolha de seis clusters não foi baseada apenas no melhor valor absoluto de uma métrica interna. Foram testados valores de *K* de 2 a 10, com diferentes sementes e múltiplas

inicializações. Foram avaliados silhouette, Davies-Bouldin, Calinski-Harabasz, tamanho dos grupos e estabilidade por ARI. A solução com seis clusters foi mantida por oferecer granularidade tática e alta estabilidade, embora soluções com três e quatro clusters tenham apresentado métricas internas superiores em alguns critérios.

3.5. Integração econômica e score híbrido

A integração com o Transfermarkt permitiu associar cada jogador ao valor de mercado mais recente disponível antes da data de referência e à transferência anterior registrada quando existente. O score final foi definido como combinação ponderada de três componentes:

$$S_i = 0,75P_i + 0,15E_i + 0,10T_i, \quad (1)$$

em que P_i representa o score técnico, E_i o score econômico e T_i o score de transferência do jogador i . O score técnico é calculado dentro do perfil tático, usando pesos específicos para as métricas mais relevantes de cada função. O score econômico favorece jogadores com menor valor relativo dentro do perfil, usando transformação logarítmica para reduzir o efeito de valores extremos. O score de transferência utiliza a razão entre taxa paga e valor de mercado quando há informação recente e confiável. Em casos ausentes ou antigos, o componente recebe valor neutro.

Após o cálculo do score final, os jogadores são classificados em baixa prioridade, monitorar ou bom potencial. Na validação retrospectiva, a interpretação decisória utiliza os limites 50 e 65. Valores abaixo de 50 indicam que o sistema não priorizaria o jogador, valores entre 50 e 65 indicam monitoramento, e valores a partir de 65 indicam recomendação.

3.6. Red flags qualitativas

O módulo de red flags foi implementado como uma camada complementar baseada em regras e categorias pré-definidas. As categorias consideradas incluem histórico de lesões, disciplina, instabilidade contratual, risco de adaptação, instabilidade de desempenho e risco jurídico ou reputacional. O módulo normaliza textos associados aos jogadores, identifica termos de risco, aplica pesos por categoria e por recência e produz um score qualitativo de alerta.

A escolha por regras foi motivada pela rastreabilidade e pela ausência de um corpus textual público, uniforme e temporalmente controlado para todos os atletas. Portanto, as red flags não foram usadas como evidência quantitativa principal na validação agregada. Elas devem ser interpretadas como mecanismo de alerta para revisão humana, não como avaliação abrangente de saúde, comportamento ou adaptação.

3.7. Modelo supervisionado com Random Forest

Como experimento complementar, foi treinado um Random Forest Regressor para estimar diretamente o score observado das transferências retrospectivas. O modelo utilizou

como entrada variáveis já produzidas pelo pipeline, incluindo score completo, score técnico, score econômico, score de transferência, valor de mercado de referência, minutos estimados, métricas por 90 minutos, perfil tático, posição e classe esperada. O alvo foi o *observed outcome score*, uma variável contínua de 0 a 100.

O Random Forest usou 300 árvores, profundidade máxima seis, mínimo de duas observações por folha e validação cruzada em cinco partições. As previsões finais foram out-of-fold, isto é, cada transferência foi prevista por um modelo treinado sem aquele caso. Esse experimento não compõe o motor principal de recomendação. Ele funciona como comparador supervisionado para verificar se um modelo aprendido superaria o score explicável e baselines simples.

3.8. Validação retrospectiva

A validação retrospectiva foi construída com transferências reais realizadas entre 1º de janeiro e 4 de setembro de 2023, com taxa mínima de 5 milhões de euros. Para cada contratação, o pipeline simulou uma decisão anterior à transferência e comparou a recomendação com um desfecho posterior. Como não havia dados públicos pós-transferência padronizados de desempenho esportivo para todos os casos, o desfecho observado foi definido como uma proxy econômica baseada na evolução do valor de mercado e na relação com a taxa paga.

Sejam V_0 o valor de mercado anterior à transferência, V_1 o valor posterior e F a taxa paga. O score observado O_i é dado por:

$$O_i = \text{clip}\{50 + 30 \tanh(1,5(V_1 - V_0)/V_0) + 20 \tanh((V_1 - F)/F), 0, 100\}. \quad (2)$$

Quando a taxa não está disponível, o segundo termo é omitido. Valores abaixo de 50 são classificados como fracasso econômico, entre 50 e 65, como neutros, e a partir de 65, como sucesso econômico. Essa variável não mede integralmente adaptação, minutos, contribuição tática ou sucesso esportivo.

3.9. Protocolo de avaliação

As variantes analisadas incluíram sempre não priorizar, média por partição, somente técnico, somente econômico, score completo e Random Forest. Para avaliar essas variantes, foram calculados MAE, RMSE, erro mediano, R^2 , Spearman, acurácia exata, acurácia balanceada, macro-F1, kappa, Precision@k e NDCG@k. A coerência da política considera compatíveis as combinações em que o sistema recomenda um jogador que posteriormente apresenta sucesso, monitora um jogador que apresenta sucesso ou resultado neutro, ou não prioriza um jogador que apresenta resultado neutro ou fracasso. Essa métrica é diferente da acurácia exata e tende a favorecer decisões conservadoras.

Foram executadas 1.000 reamostragens bootstrap para intervalos de confiança de 95% e comparações pareadas.

Nas comparações, uma melhoria positiva favorece o score completo, invertendo o sinal de métricas em que valores menores são melhores, como MAE.

4. Implementação e Reprodutibilidade

O protótipo foi desenvolvido em Python 3.10 e consolidado em uma interface de linha de comando chamada `football-ai`. O fluxo oficial inclui etapas para auditoria, engenharia de atributos, clusterização, integração econômica, avaliação, geração de relatórios, avaliação fortalecida, validação de clusters e encerramento científico. A configuração central foi mantida em arquivo YAML, com sementes, caminhos, limiares, número de clusters e parâmetros de avaliação definidos de forma explícita.

A implementação possui 24 testes automatizados, cobrindo módulos como métricas de jogadores, validação retrospectiva, score final, red flags, clusterização, integração econômica e avaliação fortalecida. O pipeline também gera artefatos padronizados em diferentes etapas, incluindo tabelas intermediárias, resultados de avaliação, diagnósticos de clusterização e relatórios finais. Essa organização permite acompanhar a origem dos resultados apresentados e verificar a consistência das saídas produzidas pelo protótipo.

Essas decisões de engenharia são relevantes porque caracterizam o trabalho como pesquisa tecnológica. Além da análise conceitual, foi produzido um protótipo reprodutível, auditável e orientado por artefatos, no qual as principais escolhas metodológicas permanecem explicitadas na configuração e nos resultados gerados. Dessa forma, a implementação não atua apenas como suporte operacional, mas como parte da contribuição do trabalho, ao demonstrar como o framework pode ser executado, avaliado e inspecionado de maneira sistemática.

5. Resultados

Esta seção apresenta os resultados obtidos com a execução do pipeline e com a validação retrospectiva. Primeiro, são descritas a base final e a distribuição das recomendações. Em seguida, são analisados os perfis táticos, a comparação entre variantes, os resultados de bootstrap, a matriz de decisão e os estudos de caso usados para interpretar acertos e divergências.

5.1. Base final e recomendações

A engenharia de dados produziu 4.406 registros jogador-competição e 3.551 atletas únicos. A análise de clusters utilizou 2.369 jogadores com pelo menos 300 minutos. O ranking final recente conteve 617 atletas. Desses, 600 possuíam valor de mercado disponível e 469 tinham transferência anterior registrada. A distribuição das classes de recomendação é apresentada na Tabela 1.

Tabela 1. DISTRIBUIÇÃO DAS RECOMENDAÇÕES NA BASE FINAL.

Classe	Jogadores	Percentual
Baixa prioridade	441	71,47%
Monitorar	166	26,90%
Bom potencial	10	1,62%

A distribuição indica uma política conservadora, na qual apenas uma fração pequena de jogadores foi classificada como bom potencial. Essa característica é coerente com a ideia de triagem e também ajuda a explicar a ocorrência de falsos negativos em jogadores que posteriormente tiveram forte valorização.

5.2. Perfis táticos

A solução operacional com seis clusters apresentou silhouete de 0,239, Davies-Bouldin de 1,342 e ARI médio de 0,982. O menor grupo teve 180 jogadores e o maior 646 na base de 2.369 atletas. Soluções com três e quatro clusters apresentaram desempenho superior em alguns critérios internos, mas K=6 foi preservado por oferecer maior granularidade tática e estabilidade entre inicializações. Essa escolha também favorece a interpretação prática dos grupos, pois separa funções com comportamentos distintos em campo, como defensores, goleiros, jogadores criativos, meio-campistas de pressão e finalizadores. Assim, embora soluções mais compactas apresentem melhor desempenho em algumas métricas internas, elas tendem a agrupar estilos diferentes em categorias amplas demais para apoiar uma análise de scouting. A solução com seis clusters foi considerada mais adequada ao objetivo do trabalho por equilibrar validade estatística, estabilidade e utilidade interpretativa.

Tabela 2. RESUMO DA SOLUÇÃO DE CLUSTERIZAÇÃO ADOTADA.

Indicador	Valor
Número de clusters adotado	6
Jogadores avaliados	2.369
Silhouette	0,239
Davies-Bouldin	1,342
ARI médio	0,982
Menor cluster	180 jogadores
Maior cluster	646 jogadores

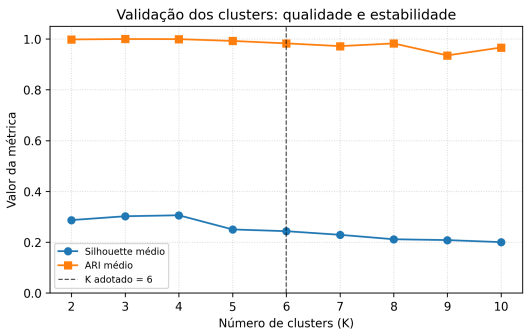


Figura 2. Diagnóstico de qualidade e estabilidade dos clusters para diferentes valores de K.

5.3. Comparação das variantes

A Tabela 3 apresenta a comparação das variantes na amostra retrospectiva de 63 transferências. O score completo teve o maior macro-F1 e a maior acurácia balanceada entre as variantes principais, indicando melhor equilíbrio entre fracasso, neutro e sucesso. Entretanto, ele não foi superior em todas as métricas. A média por partição teve menor MAE, enquanto o componente econômico isolado apresentou maior coerência e melhor desempenho em alguns critérios de ranking.

O macro-F1 é particularmente importante porque as classes observadas são desbalanceadas, com 34 fracassos, 19 neutros e 10 sucessos. Uma estratégia conservadora poderia parecer boa em acurácia ou coerência, mas não necessariamente identificaria sucessos. O macro-F1 calcula o desempenho separadamente para cada classe e depois tira uma média simples, dando peso igual a fracasso, neutro e sucesso. Por isso, essa métrica foi usada como principal referência para comparar o equilíbrio das variantes na tarefa multiclasse.

Tabela 3. COMPARAÇÃO DAS VARIANTES NA AMOSTRA RETROSPECTIVA (N = 63).

Variante	Acc.	Acc. bal.	Macro-F1	Coer.	MAE	Spearman
Sempre não priorizar	0,540	0,333	0,234	0,841	—	—
Média por partição	0,429	0,288	0,253	0,683	15,27	-0,185
Somente técnico	0,444	0,329	0,322	0,683	17,14	0,142
Somente econômico	0,571	0,368	0,302	0,873	18,20	0,098
Score completo	0,556	0,382	0,346	0,825	16,18	0,196
Random Forest	0,429	0,335	0,339	0,698	15,95	-0,065

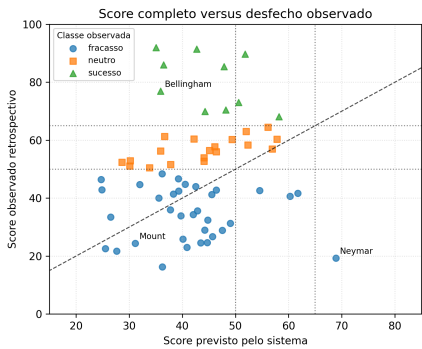


Figura 3. Relação entre o score completo previsto pelo sistema e o desfecho econômico observado. A linha tracejada representa previsão perfeita, enquanto as linhas pontilhadas indicam os limiares de 50 e 65.

5.4. Bootstrap pareado e análise de sensibilidade

O bootstrap pareado mostrou que o ganho de macro-F1 do score completo sobre o baseline “sempre não priorizar” foi 0,113, com IC de 95% de 0,018 a 0,201, e superioridade em 98,9% das reamostragens. Esse foi o resultado comparativo mais consistente a favor do score completo. Por outro lado, em MAE, o score completo não superou de forma conclusiva a média por partição.

Também foi realizada uma análise de sensibilidade dos limites de monitoramento e recomendação. A configuração

45/60 aumentaria o macro-F1 para 0,379, em comparação com 0,346 na configuração original 50/65, mas também ampliaria a quantidade de recomendações positivas e reduziria a coerência da política. Por isso, os limites originais foram mantidos para evitar ajuste retrospectivo excessivo sobre apenas 63 casos.

Tabela 4. BOOTSTRAP PAREADO: MELHORIA POSITIVA FAVORECE O SCORE COMPLETO.

Comparador	Métrica	Melhoria	IC inf.	IC sup.	P(melhor)
Sempre não priorizar	Macro-F1	0,113	0,018	0,201	0,989
Sempre não priorizar	Acurácia bal.	0,049	-0,029	0,131	0,881
Média por partição	Spearman	0,381	0,013	0,709	0,976
Média por partição	MAE	-0,912	-3,333	1,543	0,218
Somente econômico	NDCG@10	-0,138	-0,349	0,091	0,114
Random Forest	MAE	-0,232	-2,658	2,401	0,424

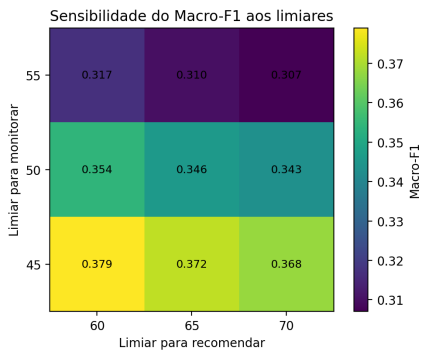


Figura 4. Sensibilidade do macro-F1 aos limites de monitoramento e recomendação. A configuração adotada no estudo foi 50 para monitorar e 65 para recomendar.

5.5. Matriz de decisão e coerência

A política original classificou 51 casos como não priorizar, 11 como monitorar e apenas um como recomendar. Entre os desfechos observados, houve 34 fracassos econômicos, 19 neutros e 10 sucessos. Foram contabilizados 52 casos coerentes e 11 divergentes, resultando em 82,54% de coerência. Contudo, sempre não priorizar produz 84,13% de coerência, pois a regra considera compatíveis tanto fracassos quanto neutros. Assim, 82,54% não deve ser descrito como acurácia de contratações.

Tabela 5. CLASSE PREVISTA VERSUS DESFECHO ECONÔMICO.

Prevista	Frac.	Neutro	Sucesso	Total
Monitorar	3	5	3	11
Não priorizar	30	14	7	51
Recomendar	1	0	0	1
Total	34	19	10	63

5.6. Estudos de caso e explicabilidade

A análise individual auxilia a identificar se um erro decorre de sinal técnico insuficiente, penalização econômica

ou ausência de variáveis contextuais. A Tabela 6 reúne casos representativos.

Tabela 6. ESTUDOS DE CASO CURADOS PARA ANÁLISE QUALITATIVA.

Jogador	Score esp.	Observado	Tipo de caso
Neymar da Silva Santos Junior	68,88	19,29	Falso positivo
Jude Bellingham	35,88	76,95	Falso negativo
Christian Pulisic	36,48	86,04	Falso negativo
Wataru Endo	48,19	70,43	Falso negativo
Kai Havertz	57,81	60,30	Coerente
Mason Mount	31,13	24,39	Coerente
Declan Rice	56,11	64,45	Coerente
Yann Sommer	61,74	41,65	Falso positivo
Jordan Henderson	36,24	16,24	Coerente
Aleksandar Mitrovic	45,55	41,27	Coerente

Neymar representa um falso positivo, pois o sistema capturou forte sinal técnico, mas o resultado econômico posterior foi baixo. Bellingham representa um falso negativo, uma vez que, por ser um jogador premium, já caro e pouco subvalorizado na lógica de custo-benefício, recebeu baixa prioridade, embora tenha apresentado sucesso econômico posterior. Mason Mount ilustra um caso coerente, no qual a baixa prioridade do sistema se alinhou ao resultado observado. Esses casos mostram que o framework organiza sinais úteis, mas não captura integralmente fatores como lesões posteriores, adaptação, contexto tático do clube comprador, minutos recebidos e dinâmica de mercado.

6. Discussão

Os resultados indicam que o score completo teve o melhor equilíbrio multiclasse entre as variantes principais, mas não demonstrou superioridade geral em todas as métricas. O principal ganho aparece no macro-F1 contra o baseline conservador, reforçado pelo bootstrap pareado. Ao mesmo tempo, a média por partição foi competitiva em erro numérico, o componente econômico apresentou forte sinal em coerência e ranking, e o Random Forest não superou o baseline de média em MAE.

Essa combinação sugere que a utilidade do framework está menos em prever exatamente o score observado e mais em estruturar uma triagem explicável. Em decisões de contratação, um sistema de apoio precisa ser interpretável, auditável e útil para reduzir o universo de análise. O score híbrido permite decompor uma recomendação em dimensões técnicas e econômicas, o que favorece a revisão humana e evita dependência exclusiva de uma saída opaca.

O resultado do Random Forest é particularmente importante. Embora um modelo supervisionado mais flexível pudesse capturar relações não lineares, a amostra retrospectiva de 63 transferências é pequena, desbalanceada e contém muitas variáveis em relação ao número de exemplos. Além disso, o desfecho observado é uma proxy econômica, não uma medida integral de sucesso esportivo. Portanto, o desempenho limitado do Random Forest reforça a decisão de tratar modelos supervisionados como complemento experimental, e não como motor principal de recomendação.

7. Limitações

As limitações podem ser agrupadas em quatro dimensões. A primeira envolve os dados utilizados, uma vez que o StatsBomb Open Data possui cobertura heterogênea, misturando torneios curtos e ligas de anos diferentes, o que limita a extrapolação para todos os contextos competitivos. A segunda dimensão está relacionada à validade externa, pois a amostra final contém 63 transferências de 2023 com taxa mínima de 5 milhões de euros, restringindo generalizações para futebol brasileiro, ligas menores, jogadores jovens, empréstimos e transferências gratuitas.

A terceira dimensão envolve aspectos metodológicos. O pareamento entre bases pode conter ambiguidades, a normalização min-max dentro dos perfis é sensível a extremos, e pesos e thresholds foram definidos por regras especializadas, não aprendidos em base independente. A quarta dimensão está relacionada aos dados textuais. O módulo de red flags foi implementado e testado, mas não havia uma coleção textual padronizada e temporalmente controlada para todos os atletas. Seu impacto quantitativo não foi isolado e não deve ser interpretado como avaliação abrangente de comportamento ou saúde.

8. Conclusão

Este trabalho transformou uma proposta conceitual em um pipeline funcional, reproduzível e auditável para apoiar a triagem de contratações no futebol. Foram integrados dados de eventos, perfis táticos, informações econômicas, regras qualitativas, explicações individuais, baselines e um experimento supervisionado. O sistema ranqueou 617 jogadores e avaliou 63 transferências reais.

O score híbrido apresentou o melhor equilíbrio multiclasse entre as variantes principais e superou o baseline “sempre não priorizar” em macro-F1 no bootstrap pareado. Contudo, a correlação com o desfecho foi fraca e incerta, o MAE não superou o baseline de média, e o Random Forest também não demonstrou vantagem. A coerência de 82,54% é uma medida permissiva de política, não uma taxa de acerto das contratações. Os seis clusters são estáveis e interpretáveis, mas não constituem a solução estatística ótima em todos os critérios.

Assim, a contribuição principal não é um preditor autônomo de sucesso, mas um framework explicável para estruturar evidências técnicas e econômicas, gerar shortlists e apoiar revisão humana.

Como trabalhos futuros, é recomendado ampliar as cortes retrospectivas, utilizar temporadas consistentes de um mesmo provedor, incorporar minutos e desempenho pós-transferência, modelar o contexto da equipe compradora, validar red flags em corpus temporal e testar o sistema em cenários do futebol brasileiro.

Referências

- [1] C. Anderson and D. Sally, *The Numbers Game: Why Everything You Know About Soccer Is Wrong*. New York: Penguin Books, 2013.

- [2] M. Lewis, *Moneyball: The Art of Winning an Unfair Game*. New York: W. W. Norton & Company, 2003.
- [3] J. H. Hewitt and O. Karakuş, "A Machine Learning Approach for Player and Position Adjusted Expected Goals in Football (Soccer)," *Franklin Open*, vol. 4, p. 100034, 2023.
- [4] R. Beal, T. J. Norman, and S. D. Ramchurn, "Artificial intelligence for team sports: A survey," *The Knowledge Engineering Review*, vol. 34, e28, 2019.
- [5] Z. Pu *et al.*, "Orientation and decision-making for soccer based on sports analytics and AI: A systematic review," *IEEE/CAA Journal of Automatica Sinica*, vol. 11, no. 1, pp. 37–57, 2024.
- [6] D. Gunning and D. Aha, "DARPA's Explainable Artificial Intelligence (XAI) Program," *AI Magazine*, vol. 40, no. 2, pp. 44–58, 2019.
- [7] D. Khalife *et al.*, "Dynamic Financial Valuation of Football Players: A Machine Learning Approach Across Career Stages," *International Journal of Financial Studies*, vol. 13, no. 2, p. 111, 2025.
- [8] C. Markopoulou, G. Papageorgiou, and C. Tjortjis, "Diverse Machine Learning for Forecasting Goal-Scoring Likelihood in Elite Football Leagues," *Machine Learning and Knowledge Extraction*, vol. 6, no. 3, pp. 1762–1781, 2024.
- [9] D. Zhou *et al.*, "Artificial intelligence in sport: A narrative review of applications, challenges and future trends," *Journal of Sports Sciences*, pp. 1–16, 2025.
- [10] D. F. Hill, J. Skinner, and A. Grosman, "A review of football player metrics and valuation methods," *Managing Sport and Leisure*, pp. 1–24, 2025.
- [11] StatsBomb, "StatsBomb Open Data," [Online]. Available: <https://github.com/statsbomb/open-data>. Accessed: 2026.
- [12] Transfermarkt, "Football transfers, rumours, market values and news," [Online]. Available: <https://www.transfermarkt.com>. Accessed: 2026.