Aplicação de algoritmos de aprendizado de máquina para predizer o desempenho de clubes de futebol europeus com base no comportamento no mercado e desempenho nas temporadas anteriores

Abel Gonçalves Chinaglia (ORCID: 0000-0002-6955-7187)[1,2] and Rafael Luiz Martins Monteiro (ORCID: 0000-0002-3208-6369)[1,2]

[1]Faculdade de Medicina de Ribeirão Preto, Universidade de São Paulo, Brasil
[2]Escola de Educação Física e Esportes de Ribeirão Preto, Universidade de São Paulo, Brasil

Resumo

O impacto financeiro das transferências de jogadores no futebol profissional tem crescido significativamente, reforçando a necessidade de investimentos mais embasados em dados por parte dos clubes. Este trabalho aplicou algoritmos de aprendizado de máquina supervisionado para predizer o desempenho futuro de clubes de futebol com base em suas atividades no mercado de transferências e no desempenho em temporadas anteriores. Foram utilizados dados das cinco principais ligas europeias — Bundesliga, Premier League, La Liga, Serie A e Ligue 1 — no período de 2009/2010 a 2019/2020. Três bases de dados foram construídas: uma baseada apenas em dados de transferências, outra contendo métricas de desempenho histórico e uma terceira combinando ambas. Sete algoritmos de classificação supervisionada foram empregados — Random Forest, Regressão Logística, K-Nearest Neighbors, Support Vector Machine, Gaussian Naive Bayes, AdaBoost e XGBoost — para classificar os clubes em três faixas de desempenho na temporada seguinte: parte superior (1º ao 6º lugar), intermediária (7º ao 14º) e inferior (15º ao 20º). A avaliação foi realizada com validação cruzada aninhada e otimização de hiperparâmetros via Grid Search, utilizando métricas como acurácia, acurácia balanceada, precisão, revocação e F1-score. Análises estatísticas, incluindo testes de Friedman e pós-hoc de Nemenyi, foram conduzidas para comparar os modelos e os conjuntos de dados. Os resultados revelaram que o ensemble dos melhores modelos apresentou o desempenho mais consistente. Não foram observadas diferenças significativas entre os cortes temporais de 1, 3 e 5 anos, exceto na Ligue 1, onde o corte de 5 anos superou o de 1 ano. A Premier League destacou-se com os melhores resultados de predição, possivelmente devido à maior constância no desempenho dos clubes. Modelos baseados apenas em dados de transferências tiveram desempenho inferior, enquanto aqueles baseados em desempenho histórico foram mais eficazes; a combinação de ambos não trouxe melhorias significativas. Este estudo demonstra a viabilidade de técnicas de aprendizado de máquina para prever o desempenho de clubes de futebol, oferecendo ferramentas analíticas valiosas para decisões estratégicas no futebol profissional.

Keywords: aprendizado de máquina, futebol europeu, transferências de jogadores, previsão de desempenho, análise de dados esportivos

1 Introdução

As transferências de jogadores de futebol profissional podem ser determinantes para o destino do time na temporada e tem envolvido cada vez maiores valores financeiros [1]. O mercado de transferências movimenta quantias substanciais de capital, com taxas crescentes que impactam significativamente a saúde financeira dos clubes [2]. Em 2023, por exemplo, foram registradas 3.279 transferências com taxas, totalizando £7.6 bilhões, um aumento considerável em relação ao ano anterior [3].

Com o aumento do dinheiro sendo investido e o impacto no time, torna-se crucial gastar dinheiro da melhor forma possível. A inteligência artificial tem sido cada vez mais aplicada na análise do desempenho esportivo, para predição de desempenho e auxiliar em tomadas de decisão [4]. Estudos já foram realizados envolvendo o valor dos atletas no mercado, porém com uma abordagem de predição do valor dos jogadores com base em estatísticas dos mesmos [5]. Até o momento não foram encontrados artigos que buscaram

predizer o desempenho do time em temporadas subsequentes com base nos dados acerca das ações daquele time no mercado, tais investigações podem ajudar a entender quais ações resultaram em melhora de desempenho e auxiliar os clubes nas tomadas de decisão ainda mais envolvendo quantidades cada vez maiores de dinheiro investido.

2 Objetivos

2.1 Geral

Aplicar algoritmos de aprendizado de máquina para predizer o desempenho futuro de times de futebol por meio da análise das transferências de jogadores e desempenho nas temporadas anteriores nas cinco principais ligas europeias (Bundesliga - Alemanha, Premier League - Inglaterra, La Liga - Espanha, Serie A - Itália e Ligue 1 - França) ao longo das temporadas de 2009/2010 a 2019/2020 utilizando diferentes filtros temporais e variáveis preditoras.

2.2 Específicos

1. Comparar a capacidade de predição dos algoritmos com diferentes filtros temporais nas transferências e desempenho anterior dos times;
2. Comparar a acurácia das predições utilizando diferentes variáveis preditoras, sendo elas: 1) dados de transferências de jogadores; 2) dados relativos ao desempenho do time nas temporadas anteriores; 3) dados de transferências de jogadores e relativos ao desempenho dos times nas temporadas anteriores;
3. Comparar a acurácia das predições nas diferentes ligas europeias;

3 Materiais e Métodos

3.1 Aquisição e Estruturação do Conjunto de Dados

Os dados utilizados neste trabalho compreendem informações de transferências de atletas e desempenho de clubes em cinco principais ligas europeias (Bundesliga, La Liga, Serie A, Premier League e Ligue 1), cobrindo as temporadas de 1992/1993 a 2021/2022. As transferências de jogadores foram extraídas do repositório `github.com/ewenme/transfers` no GitHub [6]. Já os dados de classificação dos clubes foram obtidos manualmente, temporada a temporada, por meio do site `fbref.com` [7], respeitando os termos de uso e evitando raspagem automatizada.

O dataset final foi consolidado em arquivos CSV padronizados, contendo as seguintes colunas:

- Identificação: nome do clube, liga, ano e temporada;
- Informações de transferência: nome do jogador, idade, posição, clube de origem/destino, tipo (compra, empréstimo e outros) e valor da negociação (incluindo valores ajustados);
- Desempenho do clube: posição final, partidas disputadas, vitórias, empates, derrotas, gols pró, gols contra, saldo de gols, pontos, pontos por partida e público médio;
- Estatísticas avançadas: artilheiro, goleiro (clean sheets), notas médias, gols esperados (xG), gols esperados contra (xGA), diferença de xG e xG por 90 minutos.

Cada arquivo CSV foi convertido em um *DataFrame* do `pandas` para posterior processamento, totalizando diferentes quantidades de observações por liga (por exemplo, 13,446 linhas para a Bundesliga e 27,147 para a Serie A).

3.2 Pré-processamento e Seleção Temporal

Para focar em um período representativo e isento de vieses introduzidos pela pandemia de COVID-19, realizou-se um recorte temporal, selecionando apenas dados das temporadas de 2009/2010 a 2019/2020. Em seguida, foram aplicados filtros para manter apenas as transferências que envolveram pagamento ao clube de origem, seja por meio de cláusulas de rescisão contratual ou acordos de transferência. Foram excluídas da amostra contratações sem compensação financeira, como empréstimos gratuitos, retornos de empréstimo e transferências ao fim do contrato.

Após o pré-processamento e a exclusão de transferências que não envolveram compensações financeiras, as variáveis de interesse foram extraídas e organizadas em três conjuntos experimentais distintos:

1. **Transferências de jogadores**: idade média, valores agregados de compras e vendas e contagem de atletas contratados e vendidos, desagregados por categoria de posição (goleiros, defesa, meio-campo e ataque) e sentido da transferência (entrada/saída).
2. **Desempenho histórico do clube**: posição na liga, número de vitórias, empates, derrotas, gols marcados, gols sofridos, saldo de gols e pontos por partida em janelas de 1, 3 e 5 anos anteriores à previsão.
3. **Conjunto combinado**: integração das variáveis de transferência e de desempenho histórico.

Um desafio adicional no pré-processamento foi o tratamento de clubes rebaixados, que não apresentam dados nas temporadas subsequentes à sua queda. Para lidar com essa limitação, foram implementadas três estratégias distintas:

- Exclusão completa dos clubes rebaixados da base de dados sempre que não fosse possível obter informações completas para os anos seguintes;
- Preenchimento com valores padrão (zero para métricas numéricas e posição final igual a 20°);
- Repetição dos dados da última temporada disponível para os anos subsequentes, utilizando informações históricas reais sempre que possível.

Cada uma dessas abordagens foi aplicada de forma controlada e resultou em variantes específicas dos conjuntos de dados, organizadas em subdiretórios separados. Após a análise comparativa entre essas versões, os resultados apresentados neste trabalho baseiam-se no conjunto de dados que utilizou a repetição da última temporada disponível. Essa escolha se justifica pelo melhor equilíbrio entre representatividade das classes e desempenho dos modelos preditivos, conforme discutido na seção de resultados.

A documentação detalhada sobre os scripts utilizados e a lógica de implementação pode ser consultada no repositório associado a este trabalho.

A variável resposta foi definida como a classificação final da temporada subsequente, categorizada em três grupos: topo (1^{o}–6^{o}), meio (7^{o}–14^{o}) e parte inferior (15^{o}–20^{o}). As entradas numéricas foram padronizadas via `StandardScaler` para normalizar magnitudes e melhorar a convergência dos algoritmos.

3.3 Modelos de Aprendizado de Máquina e Otimização de Hiperparâmetros

Nesta seção, detalhamos os modelos de aprendizado de máquina utilizados, a configuração dos hiperparâmetros, a estratégia de validação cruzada aninhada (*Nested Cross Validation*), a otimização de hiperparâmetros via `GridSearchCV`, e a construção de um ensemble de votação.

3.3.1 Modelos Utilizados

Foram empregados sete algoritmos de classificação supervisionada, implementados por meio das bibliotecas `scikit-learn` e `XGBoost`:

- Floresta Aleatória (`Random Forest`)
- Regressão Logística (`Logistic Regression`)
- K-Vizinhos Mais Próximos (`K-NN`)
- Máquina de Vetores de Suporte (`SVM`)
- Naive Bayes Gaussiano (`Gaussian NB`)
- AdaBoost com árvore de decisão base (`AdaBoost_DT`)
- XGBoost (`XGB`)

3.3.2 Hiperparâmetros e Otimização

Para cada modelo, foi definida uma grade de hiperparâmetros a ser explorada via `Grid Search`, uma técnica de busca exaustiva que avalia todas as combinações possíveis dentro de um espaço pré-definido. A métrica de avaliação utilizada foi a acurácia balanceada.

Na tabela 1 à seguir, apresentamos os hiperparâmetros otimizados para cada modelo:

Tabela 1: Hiperparâmetros dos Modelos de Aprendizado de Máquina

Algoritmo	**Hiperparâmetro**	**Valores**	**Kernel**
Random Forest	`n_estimators`	50, 100, 1000	
	`max_depth`	None, 5	
	`min_samples_split`	2, 5	
Logistic Regression	`C`	0.01, 0.1, 1, 10	
	`penalty`	'l2'	
	`solver`	'lbfgs'	
K-NN	`n_neighbors`	3, 5, 7, 9	
	`weights`	'uniform', 'distance'	
SVM	`C`	0.01, 0.1, 1, 10	linear
	`C`	0.1, 1, 10	rbf
	`gamma`	'scale', 'auto', 0.01, 0.1	rbf
	`C`	0.1, 1	poly
	`degree`	2, 3	poly
	`gamma`	'scale', 0.01	poly
	`coef0`	0.0, 0.5	poly
	`C`	0.1, 1	sigmoid
	`gamma`	'scale', 0.01	sigmoid
	`coef0`	0.0, 0.5	sigmoid
AdaBoost_DT	`n_estimators`	50, 100	
	`learning_rate`	0.1, 1.0	
XGB	`n_estimators`	10, 100, 500	
	`max_depth`	3, 5	
	`learning_rate`	0.01, 0.1	

Para o modelo Naive Bayes Gaussiano, não foi aplicado hiperparâmetros a serem otimizados.

3.3.3 Validação Cruzada Aninhada

Para garantir uma avaliação robusta, foi adotada uma estratégia de validação cruzada com 5 *folds* externos e 5 *folds* internos, ambos estratificados (`StratifiedKFold`). No loop interno, o `Grid Search` realiza a busca pelos melhores hiperparâmetros usando os *folds* internos, enquanto no loop externo, o modelo é treinado com esses hiperparâmetros e avaliado no *fold* de teste correspondente.

Esse processo é repetido para cada um dos sete modelos, gerando métricas de desempenho para cada *fold* externo.

3.3.4 Treinamento e Avaliação dos Modelos

Para cada modelo e cada *fold* externo, o seguinte procedimento é executado:

1. **Pré-processamento**: Os dados de treinamento e teste são padronizados via `Standard Scaler` para normalizar as *features*.

2. **Otimização de Hiperparâmetros**: No conjunto de treinamento, o `Grid Search` é utilizado com validação cruzada interna (5 *folds*) para selecionar a melhor combinação de hiperparâmetros.

3. **Treinamento**: O modelo é treinado com os melhores hiperparâmetros encontrados no conjunto de treinamento completo.

4. **Avaliação**: O modelo é avaliado no conjunto de teste do *fold* externo, calculando métricas como acurácia, acurácia balanceada, precisão, revocação, F1-score e, quando aplicável, AUC-ROC.

As métricas são armazenadas para cada *fold* e cada modelo, permitindo a análise estatística subsequente.

3.3.5 Construção e Avaliação do Ensemble

Com base nas médias das métricas de F1-score obtidas na validação cruzada externa, os três melhores modelos individuais são selecionados para compor um ensemble de votação (`VotingClassifier`). Cada um desses modelos é re-treinado utilizando todo o conjunto de dados, com os respectivos hiperparâmetros previamente otimizados.

O ensemble é então avaliado seguindo a mesma estratégia de validação cruzada externa (5 *folds*), garantindo que, para cada *fold*, o treinamento ocorra apenas sobre os dados de treino e a avaliação seja realizada nos dados de teste correspondentes. As métricas resultantes são agregadas da mesma forma que para os modelos individuais.

O tipo de votação empregado pelo ensemble é definido como "soft" caso todos os modelos selecionados sejam capazes de fornecer probabilidades de classe; caso contrário, adota-se a votação "hard" (majoritária).

3.4 Visualização e Testes Estatísticos

Para comparar distribuições de métricas entre modelos, foram gerados:

- Boxplots das métricas (balanced accuracy, precision, recall e F1-Score) com indicação de médias;
- Matrizes de confusão agregadas para cada modelo, com anotação da acurácia geral.

Testes estatísticos não-paramétricos de Friedman foram realizados para cada métrica agrupada, avaliando diferenças significativas entre os modelos no nível de $\alpha < 0{,}05$. Quando significativo, aplicou-se post-hoc de Nemenyi via `scikit-posthocs` para identificar pares de classificadores com desempenho diferenciado.

Todas as etapas de pré-processamento, análise e modelagem foram implementadas em Python 3, utilizando bibliotecas `pandas`, `numpy`, `scikit-learn`, `xgboost`, `matplotlib`, `seaborn` e `scikit-posthocs`. Dados de entrada e scripts de treinamento e avaliação estão disponíveis no repositório associado.

3.5 Análises iniciais e escolha da estratégia de pré-processamento

Ao analisar as métricas de performance dos modelos treinados com diferentes estratégias de pré- processamento para lidar com os clubes rebaixados, optou-se pelo uso do dataset com repetição dos dados da última temporada disponível. Essa escolha se deu pela análise das matrizes de confusão e demais métricas por meio de boxplots. As figuras 1, 2 e 3 exemplificam a visualização dos resultados que nos auxiliaram na tomada de decisão. Todas as matrizes de confusão e boxplots dos modelos treinados estão disponíveis no repositório disponibilizado ao final deste documento.

Ao observar os resultados obtidos com o conjunto de dados colocando zero no lugar das métricas dos times nos anos em que eles foram rebaixados, notou-se que foi criado um grande viés que facilitou a classificação da classe 2, relativa aos times da parte debaixo da tabela (Figura 1). Já quando foi observado os resultados excluindo os dados de todos os times que foram rebaixados a classe 2 ficou sub-representada, o que a tornou muito difícil de classificar (Figura 2). Já quando utilizamos os dados repetidos dos anos anteriores observou-se maior equilibrio das métricas e menor viés (Figura 3). Para análise final foram escolhidos somente os modelos com maior acurácia de cada liga em cada dataset e corte temporal.

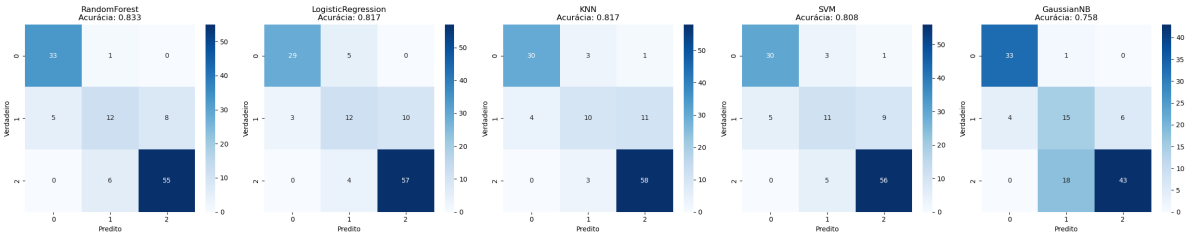

Figura 1: Matriz de confusão do *data_2* considerando o corte temporal de 5 anos, com os valores das variáveis dos times rebaixados zerados.

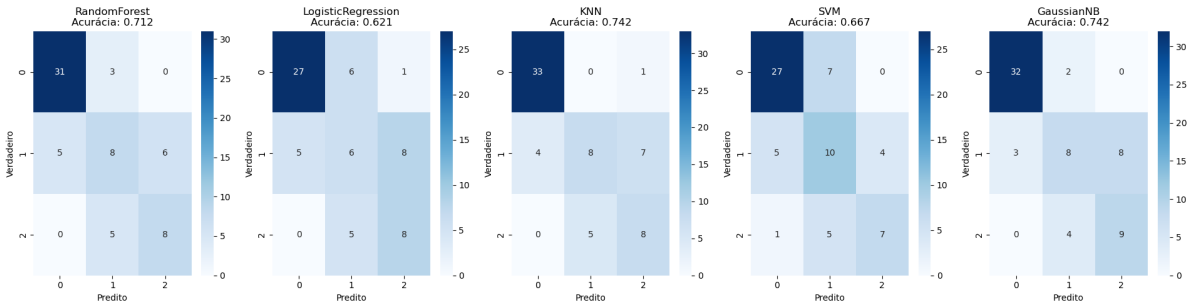

Figura 2: Matriz de confusão do *data_2* considerando o corte temporal de 5 anos, com os valores das variáveis dos times excluídas.

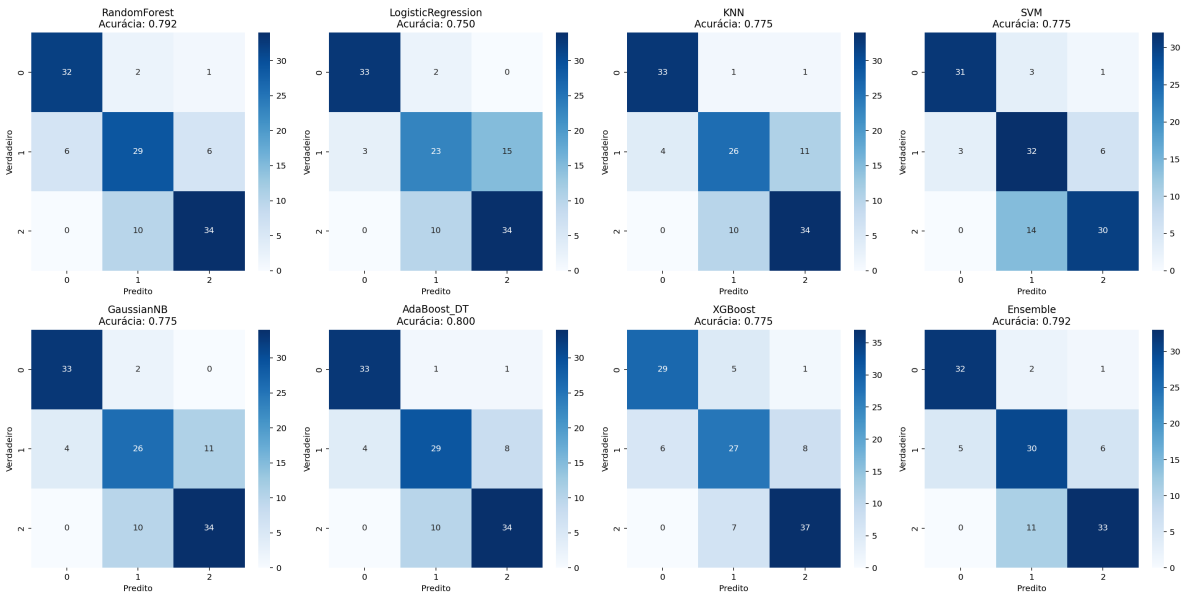

Figura 3: Matriz de confusão do *data_2* considerando o corte temporal de 5 anos, com os valores das variáveis dos times rebaixados repetidos.

4 Resultados

Para apresentação dos resultados finais da comparação entre os modelos será realizada uma divisão em subtópicos. Cada subtópio terá como objetivo responder à um dos objetivos específicos enumerados acima.

4.1 Comparação entre os filtros temporais (1, 3 e 5 anos)

As acurácias médias dos modelos podem ser observadas nas figuras 4, 5 e 6. No segundo conjunto de dados (`data_2`), o teste de Friedman para o corte de 5 anos indicou diferença global significativa entre os três `datasets` ($\chi^2 = 8{,}44$, $p = 0{,}015$) na Ligue 1 francesa. O pós-hoc de Nemenyi indicou diferença entre o corte temporal de 1 e 5 anos ($p = 0{,}0307$). Em todos os demais casos os testes de Friedman não revelaram diferenças significativas ($p > 0{,}05$) entre os cortes temporais de 1, 3 e 5 anos.

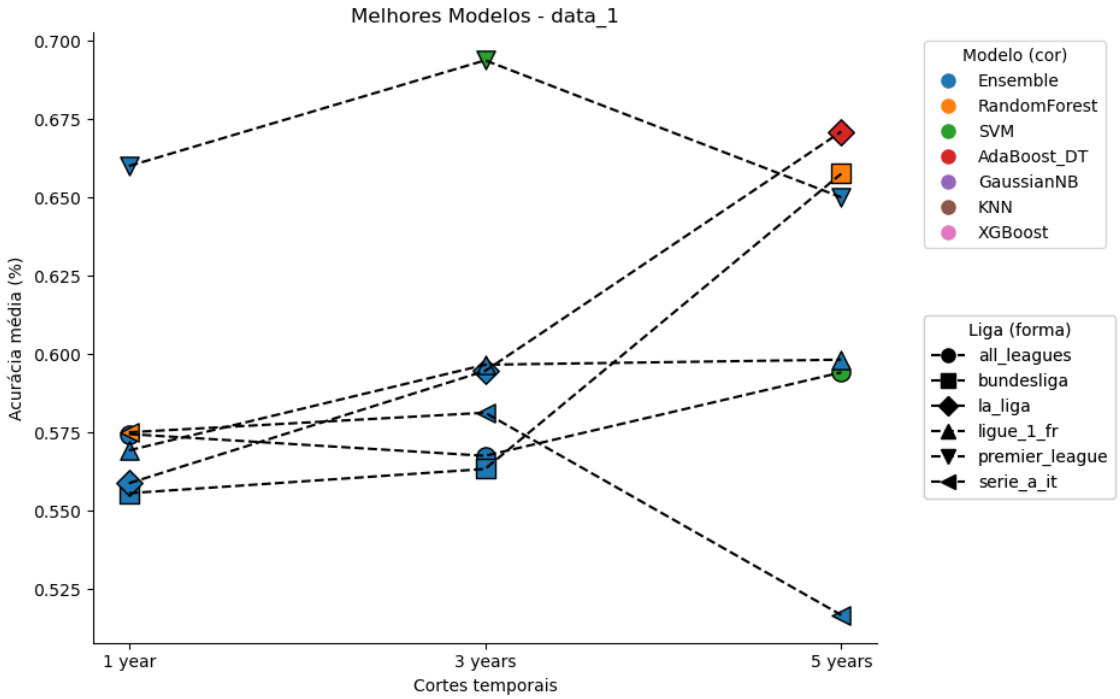

Figura 4: Acurácia média dos melhores modelos do data 1 nos diferentes cortes temporais

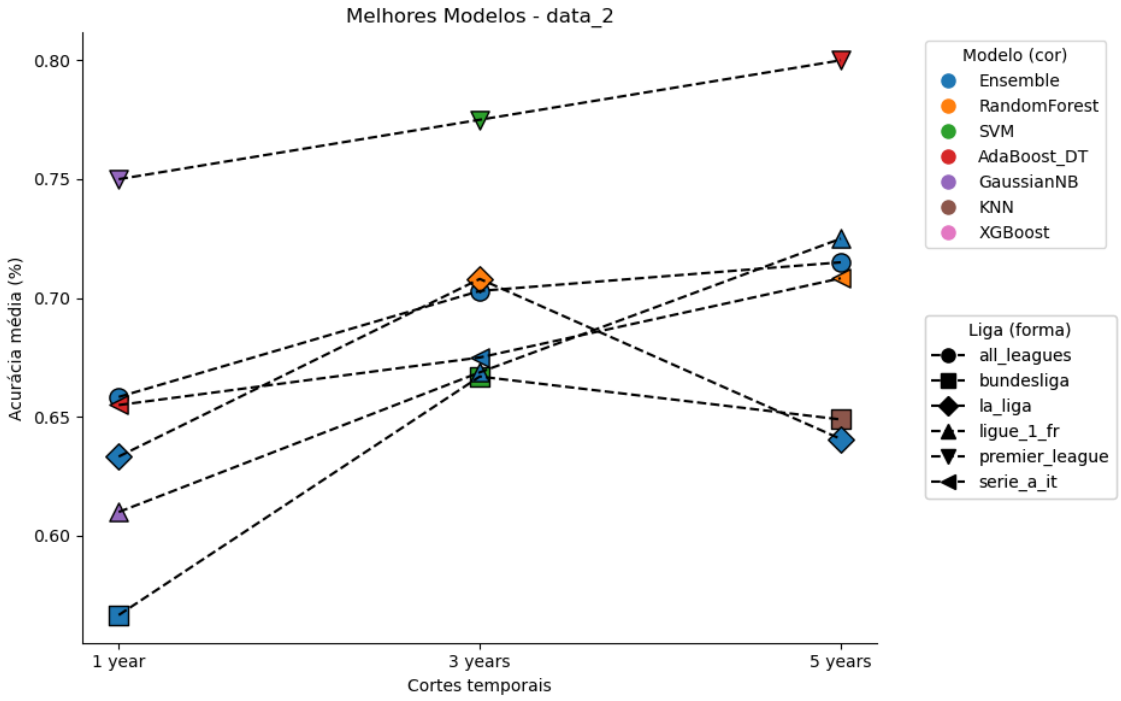

Figura 5: Acurácia média dos melhores modelos do data 2 nos diferentes cortes temporais

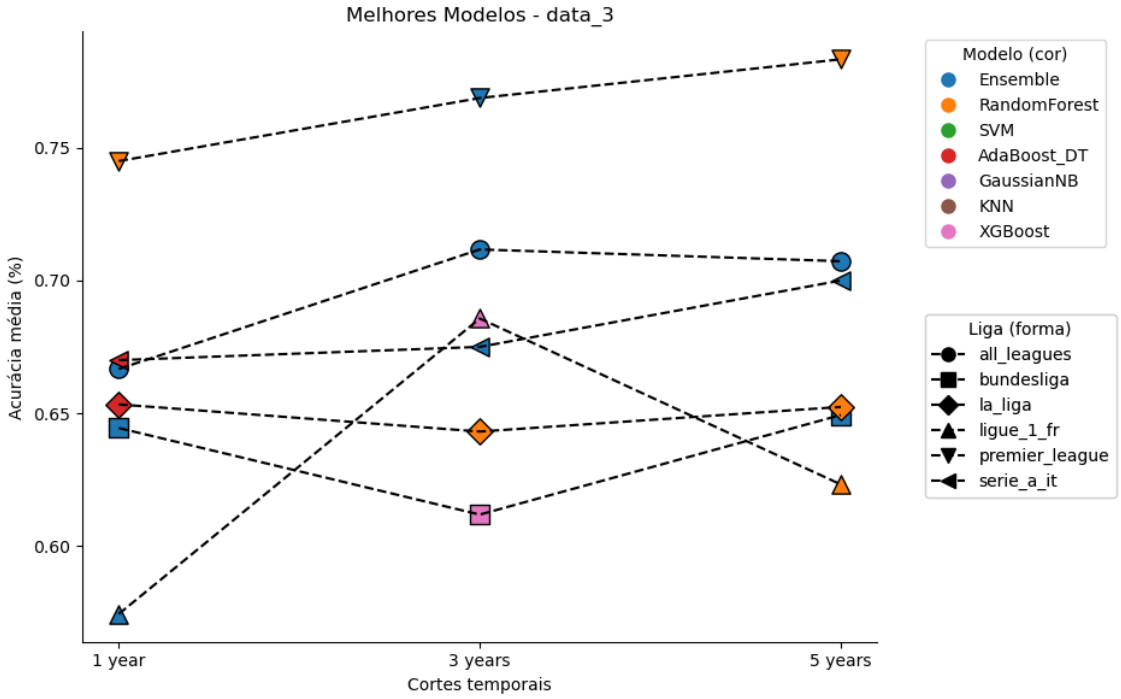

Figura 6: Acurácia média dos melhores modelos do data 3 nos diferentes cortes temporais

4.2 Comparação entre os datasets

As acurácias médias dos modelos podem ser observadas nas figuras 7, 8 e 9. Para todas as ligas combinadas (*all_leagues*), o teste de Friedman indicou diferença global significativa em 1 ano ($\chi^2 = 7{,}60$,

$p = 0{,}022$), 3 anos ($\chi^2 = 7{,}60$, $p = 0{,}022$) e 5 anos ($\chi^2 = 7{,}60$, $p = 0{,}022$). No post hoc, observou-se diferença significativa entre os *datasets* 1 e 3 em 1 ano ($p = 0{,}031$), entre 1 e 2 em 3 anos ($p = 0{,}031$) e novamente entre 1 e 3 em 5 anos ($p = 0{,}031$).

Na *Bundesliga*, houve diferença global em 3 anos ($\chi^2 = 7{,}44$, $p = 0{,}024$) e o post hoc indicou diferença significativa entre os *datasets* 1 e 2 ($p = 0{,}031$). Na *La Liga*, o teste de Friedman inidicou diferença no corte de 3 anos ($\chi^2 = 6{,}00$, $p = 0{,}050$), com diferença post hoc entre 1 e 2 ($p = 0{,}047$). Na *Premier League*, verificou-se diferença global em 5 anos ($\chi^2 = 7{,}41$, $p = 0{,}025$) e post hoc entre 1 e 2 ($p = 0{,}047$). Na *Série A Italiana*, houve diferença global em 1 ano ($\chi^2 = 8{,}32$, $p = 0{,}016$) e 5 anos ($\chi^2 = 7{,}60$, $p = 0{,}022$), com diferenças post hoc entre 1 e 3 em ambas as durações (1 ano: $p = 0{,}020$; 5 anos: $p = 0{,}031$). Nas demais combinações de liga e duração, não se observaram diferenças significativas ($p > 0{,}05$).

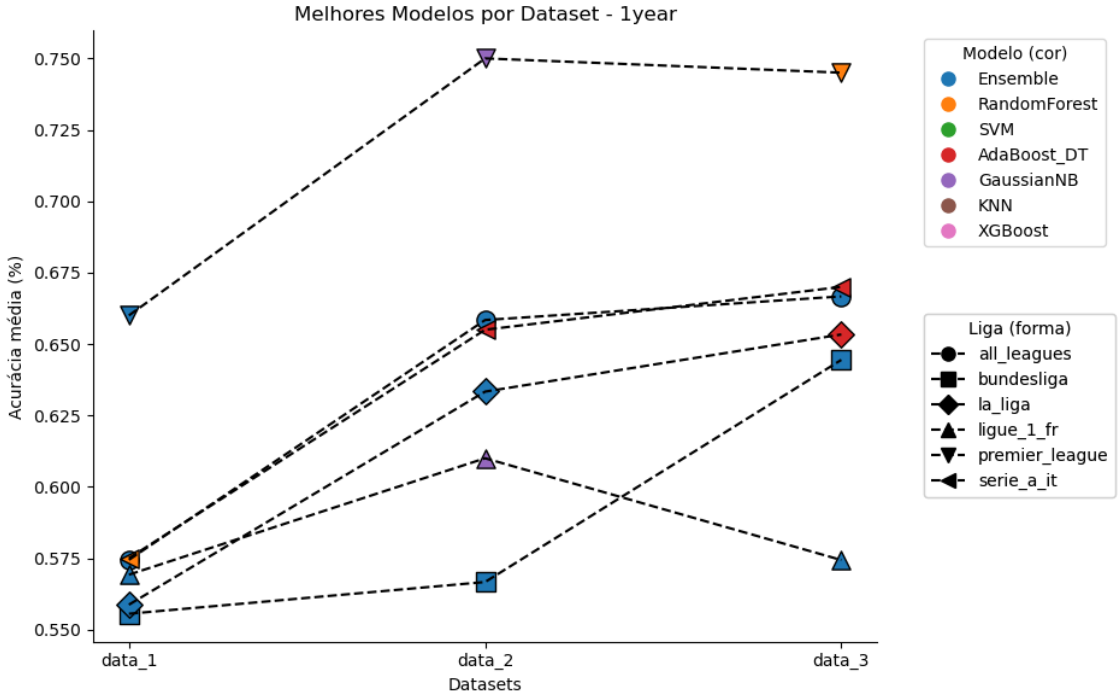

Figura 7: Acurácia média dos melhores modelos com corte temporal de 1 ano nos diferentes datasets

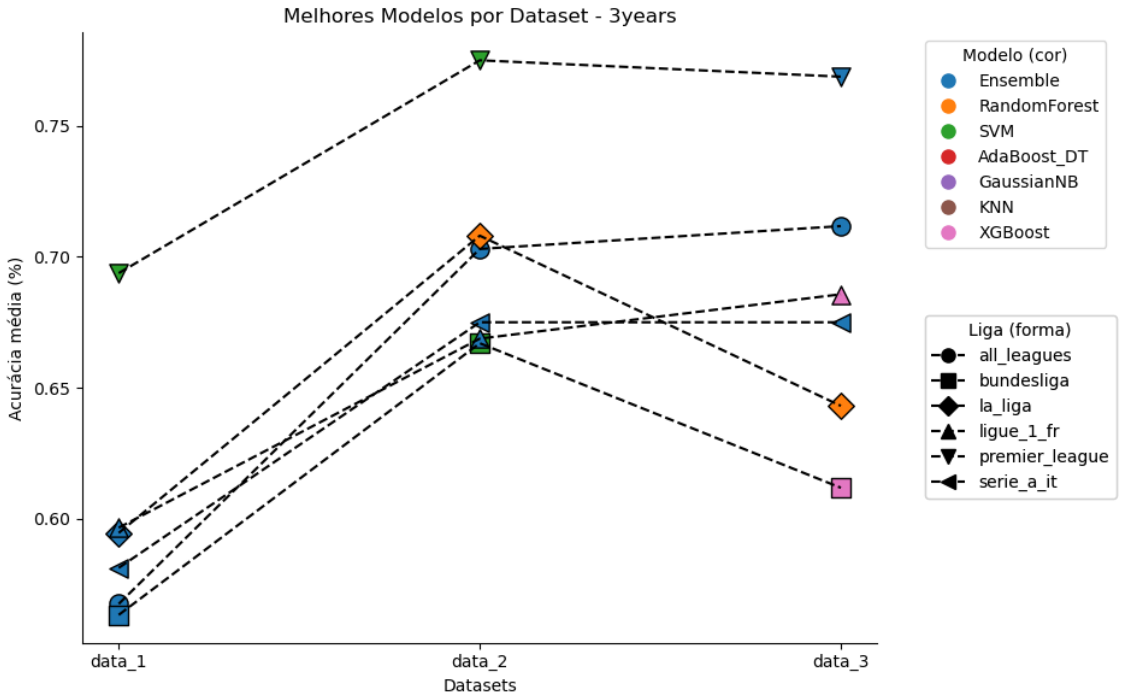

Figura 8: Acurácia média dos melhores modelos com corte temporal de 3 anos nos diferentes datasets

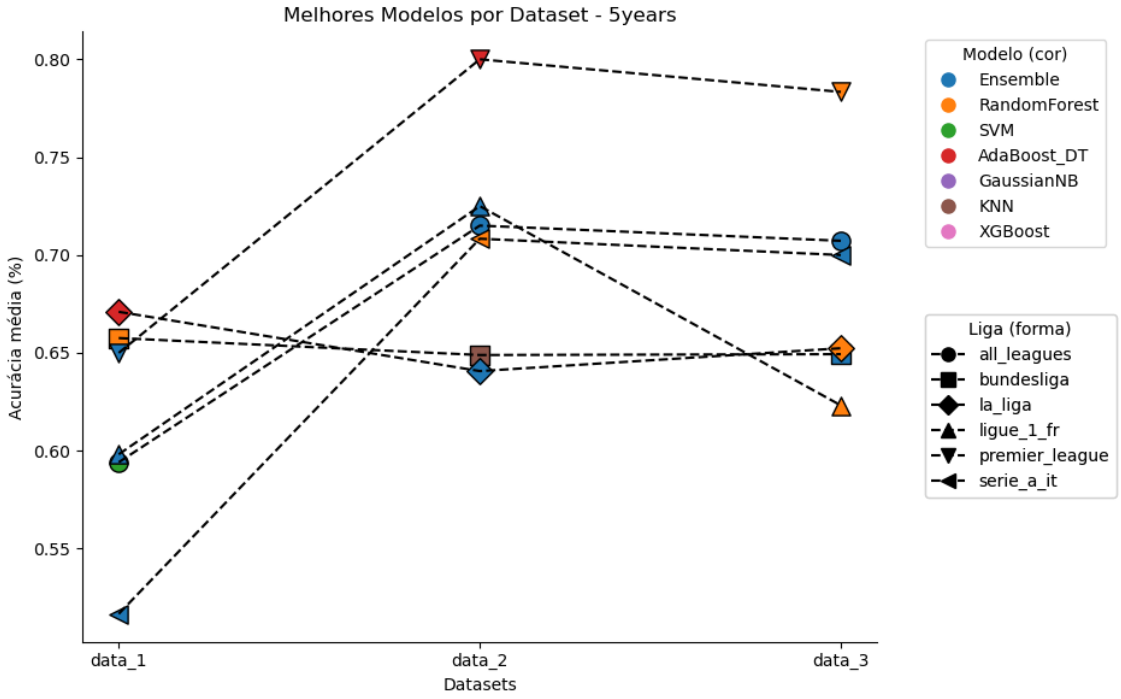

Figura 9: Acurácia média dos melhores modelos com corte temporal de 5 anos nos diferentes datasets

4.3 Comparação entre as ligas europeias

Para o primeiro conjunto de dados, relativo à transferência dos jogadores (`data_1`), o teste de Friedman para o corte de 5 anos indicou diferença global significativa entre as ligas analisadas ($\chi^2 = 14{,}71$,

$p = 0{,}0117$). No pós-hoc de Nemenyi, observouse diferenças significativas entre *La Liga* e *Série A Italiana* ($p = 0{,}0166$) e entre *Premier League* e *Série A Italiana* ($p = 0{,}0284$).

Para o segundo conjunto de dados (`data_2`), o teste de Friedman no corte de 1 ano revelou diferença global significativa ($\chi^2 = 13{,}33$, $p = 0{,}0204$). No pós-hoc, houve diferença significativa entre *Bundesliga* e *Premier League* ($p = 0{,}0218$) e entre *Ligue 1 Francesa* e *Premier League* ($p = 0{,}0366$).

Em todos os demais casos — para os cortes de 1 e 3 anos em `data_1`, para os cortes de 3 e 5 anos em `data_2`, bem como em qualquer duração para os demais conjuntos de dados — os testes de Friedman não revelaram diferenças globais significativas ($p > 0{,}05$).

5 Discussão

Em relação aos modelos de aprendizado de máquina, o ensemble dos melhores modelos foi o que mais se destacou. Isso era esperado levando em conta que eles utiliza a decisão dos três melhores modelos para a tomada de decisão. De forma geral pode-se observar que somente com os dados de transferências (data 1) os modelos apresentaram um desempenho inferior na predição da posição final nas temporadas seguintes. Acredita-se que isso aconteceu por fornecer informações insuficientes para predição. Quando observa-se a utilização do desempenho do time nas temporadas anteriores para predição (data 2) os modelos apresentaram um desempenho um pouco melhor, variando entre as diferentes ligas. Já quando se observa a predição utilizando os dados de desempenho junto com os de transferência (data 3) não foi apresentada uma melhora em relação à somente utilização dos dados de desempenho.

Já quando se observa o efeito dos cortes temporais nas predições apesar de visualmente parecer que há uma melhora com um corte temporal maior de forma geral não houve diferença significativa entre os cortes de 1, 3 e 5 anos, somente com a excessão da Ligue 1 que apresentou diferença entre a predição no data 2 entre 1 e 5 anos, com melhor predição com o corte temporal maior. Por fim, ao se observar as diferenças entre as predições nas diferentes ligas europeias, pode-se observar claramente que na premier league houve melhores valores de predição o que ficou evidente também nas análises estatísticas.

Acredita-se que este resultado reflete uma maior constância dos times da Premier League, com menores flutuações entre as equipes ao longo dos anos. Ou seja, os time que se destacam acabam tendo esse bom desempenho nos anos seguinte também, o mesmo vale para os times de meio da tabela e aqueles que brigam contra o rebaixamento. Outro estudo também demonstra essa tendência de constância da Premier League, aplicando algoritmos de árvore de decisão para classificar resultados da partida com base na qualidade dos oponentes obteve-se 67,9, 73,9 e 78,4% contra oponentes balanceados, mais fortes e mais fracos, respectivamente[8].

Entre as limitações do presente estudo, destaca-se o corte temporal de 10 anos devido à exclusão de dados das temporadas em que houve a COVID-19. Além disso, acredita-se que podem ser testadas outras estratégias de gridsearch para melhor otimização do algoritmos. Outra estratégia que pode ser interessante é a realização de seleção de atributos antes do treinamento e avaliação final dos modelos. As maiores dificuldades do presente estudo foram a definição das estratégias de pré-processamento e o balizamento entre custo computacional e estratégias de treinamento e avaliação dos modelos. Por fim, acredita-se que técnicas de aprendizado de máquina para análise esportiva e auxilio na tomada de decisão devem ser empregadas no esporte por demonstrar uma capacidade de predição de fenômenos complexos e descoberta de conhecimento[4].

6 Conclusão

O ensemble se destacou para predição da classificação final dos times nas temporadas seguintes. Os diferentes cortes temporais não demonstraram diferença na predição. Em relação à comparação da capacidade de predição entre as ligas europeias, a premier league foi a que demonstrou melhores resultados. Por fim, utilizar dados do desempenho nas temporadas anteriores foi mais efetivo para predição do que a utilização de dados de transferência de mercado.

7 Material Suplementar

Para informações mais detalhadas dos resultados e acessar os scripts desse projeto, acesse o repositório Abel-Chinaglia/Footbal_prediction-Introducao_AM

Referências

[1] European Club Association (ECA), PricewaterhouseCoopers (PwC), LIUC Università Cattaneo (represented by Emanuele Grasso and Ernesto Paolillo). STUDY ON THE TRANSFER SYSTEM IN EUROPE; 2013. Acessado em abril 8, 2025. Available from: https://www.ecaeurope.com/media/2731/eca-study-on-transfer-system-in-europe.pdf.

[2] Veliz E. Here We Go! Predicting Transfer Market Valuations of Premier League Footballers; 2025. Acessado em abril 8, 2025. https://medium.com/stanford-cs224w/here-we-go-predicting-transfer-market-valuations-of-premier-league-footballers-1774131946ff.

[3] FIFA. FIFA Global Transfer Report 2023. FIFA; 2023. Available from: https://digitalhub.fifa.com/m/114622e4e17cf6a8/original/FIFA-Global-Transfer-Report-2023.pdf.

[4] Claudino JG, Capanema DdO, de Souza TV, Serrão JC, Machado Pereira AC, Nassis GP. Current approaches to the use of artificial intelligence for injury risk assessment and performance prediction in team sports: a systematic review. Sports medicine-open. 2019;5:1-12.

[5] Shen Q. Predicting the value of football players: machine learning techniques and sensitivity analysis based on FIFA and real-world statistical datasets. Applied Intelligence. 2025;55:265. Available from: https://doi.org/10.1007/s10489-024-06189-0.

[6] ewenme. transfers: data on European football player transfers; 2025. Repositório sem licença explícita; acesso em 10 nov. 2023. https://github.com/ewenme/transfers.

[7] Sports Reference LLC. FBref.com: Football Statistics and History; 2018. Acesso em 10 nov. 2023; uso automatizado proibido, extração manual permitida (cláusula 5). https://fbref.com.

[8] Bilek G, Ulas E. Predicting match outcome according to the quality of opponent in the English premier league using situational variables and team performance indicators. International Journal of Performance Analysis in Sport. 2019;19(6):930-41.