Assessor Models with Reject Option for soccer result prediction

Daniel C. da Costa[1], **Ricardo Prudêncio**[1], **Alexandre Mota**[1]

[1]Centro de Informática – Universidade Federal de Pernambuco (UFPE)
Av. Jornalista Aníbal Fernandes, s/n – Cidade Universitária
Recife-PE – Brasil CEP: 50.740-560

dcc5@cin.ufpe.br, acm@cin.ufpe.br, rbcp@cin.ufpe.br

Abstract. *Soccer is a widely popular sport, both in Brazil and around the world, with a billion-dollar industry surrounding it. The use of data and Machine Learning (ML) algorithms has been explored as a tool to predict outcomes in this sport. However, the unpredictability of soccer makes it challenging to obtain accurate and reliable predictions. In this study, we propose to use a ML model called assessor, which analyzes the predictions returned by a classifier of match outcomes in order to select those ones with the highest reliability, discarding the others. We seek to optimize the relationship between the accuracy of accepted predictions and rejection rate, in order to maximize the reliability of the model adopted for match outcomes. We performed experiments with real data, identifying the championships, teams and rounds in which the proposed model presents the best performance. This innovative approach contributes to the improvement of soccer result predictions, using advanced ML techniques together with the selection of high-quality predictions.*

Resumo. *O futebol é um esporte amplamente popular, tanto no Brasil quanto em todo o mundo, com uma indústria bilionária ao seu redor. A utilização de dados e algoritmos de Aprendizagem de Máquina (AM) tem sido explorada como uma ferramenta para prever resultados nesse esporte. No entanto, a imprevisibilidade do futebol torna desafiador obter previsões precisas e confiáveis. Neste estudo, propomos a utilização de um modelo de AM chamado assessor, que analisa as predições de um classificador de resultados de partidas para selecionar aquelas com maior confiabilidade, descartando as demais. Buscamos otimizar a relação entre acurácia das predições aceitas e a taxa de rejeição, de forma a maximizar a confiabilidade no uso do modelo de AM para predição das partidas. Realizamos experimentos com dados reais de partidas, identificando os campeonatos, equipes e rodadas em que o modelo proposto apresenta melhor desempenho. Essa abordagem inovadora contribui para o aprimoramento das previsões de resultados de futebol, utilizando técnicas avançadas de AM em conjunto com a seleção de predições de alta qualidade.*

1. Introdução

O futebol é indiscutivelmente o esporte mais popular tanto no Brasil quanto em todo o mundo. De acordo com [FIFA 2018], a Copa do Mundo da FIFA de 2018, sediada na Rússia, cativou uma audiência de mais de 3,5 bilhões de pessoas, que acompanharam os

jogos pela televisão e por meio digital. A final entre França e Croácia, por exemplo, atraiu impressionantes 1,12 bilhão de telespectadores. Essa magnitude de interesse resulta na formação de uma indústria bilionária que engloba patrocínios, vendas de jogadores e produtos, direitos de transmissão, entre outros. Apenas os vinte clubes mais ricos do mundo geraram incríveis 10,6 bilhões de dólares em receitas no ano de 2020 [Deloitte 2020].

No contexto apresentado, o uso de algoritmos de Aprendizado de Máquina (AM) é uma alternativa viável para prever resultados de partidas de futebol, devido à habilidade desses algoritmos em generalizar padrões em dados complexos. Informações provenientes de plataformas especializadas em análises de futebol, que disponibilizam, por exemplo, indicadores de desempenho, estilos de jogo e dados históricos das equipes envolvidas, podem aprimorar os modelos gerados [Stübinger et al. 2019, Hucaljuk e Rakipović 2011, Godin et al. 2014, Tax e Joustra 2015].

A imprevisibilidade do futebol, que é um dos elementos mais fascinantes desse esporte, gera um nível significativo de incerteza ao se realizar predições de resultados por meio de modelos de AM. Nesse sentido, é importante identificar quais predições têm maior probabilidade de estarem incorretas e, se necessário, descartá-las. Esse tipo de estratégia é referenciado comumente na literatura como AM com opção de rejeição [Hendrickx et al. 2021, Chow 1970], explorada com o uso de diversas técnicas para filtragem de predições de alta incerteza [Geifman e El-Yaniv 2017, Jiang et al. 2018]. Em relação às predições esportivas, estudos anteriores adotaram a estratégia de descartar predições com maior incerteza, derivada a partir das probabilidades de resultados estimadas pelo próprio modelo de predição [Partida et al. 2021, Stübinger et al. 2019]. O sucesso dessa estratégia, no entanto, depende da qualidade do modelo de AM. De fato, um modelo de AM pode eventualmente estimar uma alta probabilidade para predições que se mostram erradas (i.e., o modelo é altamente confiante mas a predição é errada), o que se caracteriza como uma incerteza epistêmica associada ao modelo [Hüllermeier e Waegeman 2021].

Nesse trabalho, investigamos o uso de modelos assessores [Hernández-Orallo et al. 2022] para opção de rejeição em AM, no contexto de predições de partidas de futebol. Um modelo assessor monitora o desempenho de teste de um dado modelo base e com isso, generaliza situações onde o modelo base tem maior probabilidade de acerto. O assessor é um modelo de AM independente, treinado a partir dos acertos e erros do modelo base ao longo do tempo. Assessores já foram usados, por exemplo, para predizer o desempenho de modelos de linguagem dado um prompt de texto de entrada [Zhou et al. 2022]. No nosso contexto, um assessor é construído para predizer a confiança (probabilidade de acerto) das predições de um modelo base ao longo das rodadas de um campeonato de futebol. A cada rodada, a predição do modelo base é comparada com o resultado real da partida e o seu acerto ou erro é registrado. Um exemplo de treinamento para o assessor é então criado com: (1) atributos preditores - as caracteristicas da partida e as probabilidades de resultados preditas pelo modelo base para aquela partida; e (2) atributo classe, definido como valor positivo, se o modelo base acertou o resultado da partida, e valor negativo, caso contrário. Uma vez treinado, o assessor pode ser usado para predizer a probabilidade de acerto do modelo base para as partidas da rodada seguinte. Se essa probabilidade para uma dada partida for abaixo de um limiar de aceitação, a predição do modelo base é então rejeitada.

Na literatura consultada, não foi encontrado nenhum trabalho que utilize modelos assessores como opção de rejeição para previsões em partidas de futebol. Experimentos realizados com dados de cinco ligas nacionais europeias, mostraram resultados promissores dos assessores em relação à abordagem tradicional de opção de rejeição. As principais contribuições deste trabalho são então:

- A investigação original de modelos assessores, para analisar as predições de partidas e rejeitá-las caso não alcancem um valor mínimo de confiabilidade;
- A realização de experimentos que verificam a viabilidade da proposta, considerando para quais campeonatos, faixas de rodadas e equipes nossa proposta possui melhor desempenho.

Este trabalho está organizado da seguinte forma. Na Seção 2 apresentamos os trabalhos relacionados. Em seguida, na Seção 3, apresentamos o sistema proposto e explicamos seus componentes. Na Seção 4, explicamos a metodologia experimental utilizada no trabalho e em seguida, na Seção 5, expomos os resultados obtidos. Por fim, na Seção 6, discutimos nossas conclusões e principais trabalhos futuros.

2. Trabalhos relacionados

Nessa seção, fazemos uma breve revisão de abordagens e estudos relacionados ao uso de dados no futebol (Seção 2.1) e da utilização de métodos de AM com opção de rejeição (Seção 2.2).

2.1. AM no futebol

A utilização de algoritmos de AM para a resolução e auxílio de tarefas é uma realidade em todas as áreas do conhecimento. No esporte e, especificamente, no futebol, vários estudos têm explorado as aplicações de AM. Por exemplo, [Rossi et al. 2018] desenvolveu um modelo que utiliza dados de GPS de treinos para prever lesões de jogadores profissionais a partir da sua carga de trabalho. Saindo da área da saúde e indo para o campo, [Brooks et al. 2016] utilizou a locação de passes durante uma posse de bola para prever a ocorrência de chutes e classificar jogadores de acordo com o valor de seus passes.

Além de trabalhos que aplicam AM para a análise de ações dentro de um jogo, existem trabalhos com o objetivo de predizer o resultados das partidas, que é o foco do presente trabalho. Dentre os modelos desenvolvidos por outros autores, [Tax e Joustra 2015] utilizaram a técnica de Análise de Componentes Principais junto aos algoritmos de Naive Bayes e um Perceptron multicamadas para obter uma acurácia de 54,7%. Outros trabalhos (e.g., [Constantinou 2019], [Hubáček et al. 2019]) podem ser citados no contexto da competição Machine Learning for Football, de 2017, em que as equipes tiveram que predizer resultados de partidas a partir de uma base de 200.000 jogos que continha o nome dos times participantes e o resultado final do jogo.

Com foco na predição para a utilização em estratégias de apostas, [Stübinger et al. 2019] chegaram a um retorno financeiro de 1,58% por partida ao combinar diferentes técnicas de AM na hora de definir uma aposta e usar dados provenientes do videogame FIFA para obter características dos jogadores dos times. Outro estudo chegou a uma lucratividade de 30% ao utilizar tweets de torcedores para uma análise de sentimento junto a métodos estatísticos para predizer resultados das partidas da temporada 2017/2018 da Premier League [Godin et al. 2014].

2.2. AM com Rejeição

Vários métodos já foram testados para a rejeição de predições de AM com o objetivo de aumentar a confiabilidade do uso de sistemas de AM [Chow 1970] [Nicora et al. 2022]. [Geifman e El-Yaniv 2017] utilizam um algoritmo de busca binária que seleciona as predições de uma rede neural de forma a atingir uma taxa de risco pré-determinada pelo usuário. Já [Jiang et al. 2018], propuseram o Trust Score, derivado através de um escore fornecido por um modelo K-vizinhos próximos treinado sobre o conjunto de teste de um classificador binário e o próprio valor de confiança do modelo base. O novo escore obtido se mostrou mais relevante para identificar exemplos corretamente classificados.

No contexto do futebol, os vários trabalhos da literatura comumente aceitam previsões para todos os jogos, sem que haja uma filtragem dos resultados mais confiáveis [Hucaljuk e Rakipović 2011], [Tax e Joustra 2015], [Constantinou 2019], [Hubáček et al. 2019] [Godin et al. 2014]. No entanto, já existem trabalhos que realizam apostas baseadas em identificar predições de maior confiabildiade, i.e., com o objetivo de restringir as apostas para aquelas partidas cuja predição de resultado tem um maior grau de certeza. Esses trabalhos podem ser enquadrados com a estratégia de AM com opção de rejeição. Em [Stübinger et al. 2019], os autores consideram partidas filtradas por prediçao de vitória com pelo menos dois gols de diferença. Em [Partida et al. 2021], os autores usam diretamente as saídas do modelo de predição de resultados, filtrando partidas com mais de 50% ou 70% de confiança no resultado.

3. Proposta

Trabalhos anteriores da literatura já demonstraram a viabilidade do uso de AM para a predição de resultados de jogos no contexto de futebol. No entanto, esse é um problema difícil, que pode envolver um alto grau de incerteza dependendo dos times envolvidos. Nesse trabalho, propomos um sistema de AM que adota modelos assessores [Hernández-Orallo et al. 2022] para aceitar ou rejeitar as predições de um preditor de partidas. Trabalhos anteriores tentaram estimar incerteza das predições de partidas usando informações geradas pelo próprio preditor, como as probabilidades estimadas dos resultados [Partida et al. 2021]. No entanto, isso pode ser problemático no caso de preditores mal estimados ou baixa competência para contextos específicos de uso. O assessor, por sua vez, é um modelo independente, aprendido a partir dos erros e acertos do modelo base ao longo de um campeonato. Assim, ele deve ser capaz de identificar características das partidas em que o modelo base não é competente para realizar predições e nesses casos, recomendar a rejeição das predições.

O sistema proposto está representado na Figura 1. Dada uma partida, um modelo base de predição é usado para predizer probabilidades associadas aos resultados possíveis da partida, a partir das estatisticas dos times envolvidos e outras informações. Um modelo assessor recebe como entrada as características da partida, assim como as probabilidades estimadas pelo modelo base. A partir daí, o assessor estima a probabilidade de acerto do modelo base e usa essa informação para então aceitar ou rejeitar a predição. O modelo base é construído a partir de dados históricos de resultados das partidas de um campeonato (ver seção 3.1). Já o modelo assessor aprende a partir dos erros obtidos pelo modelo base, e desta forma, monitora e prediz a confiança associada às predições geradas pelo modelo base (ver seção 3.2).

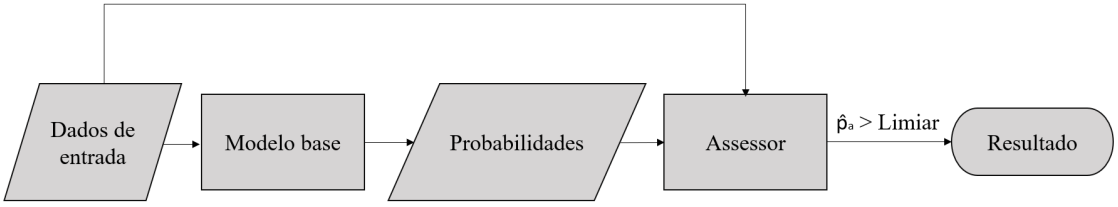

Figura 1. Sistema de predição proposto

3.1. Modelo Base

Cada partida é descrita por um vetor de características $\mathbf{x} = (x_1, \ldots, x_p)$. Essas características envolvem comumente estatísticas dos times envolvidos na partida. Para cada partida, o modelo base retorna a probabilidade de ocorrência de cada um dos resultados possíveis da partida, $\mathbf{p} = (p_m, p_e, p_v)$, dentre vitória do mandante (m), empate (e) e vitória do visitante (v). A predição final $\hat{y}$ do modelo base é o resultado possível da partida com maior probabilidade estimada:

$$\hat{y} = arg \max_{y \in \{m,e,v\}} \hat{p}_y \tag{1}$$

O modelo base é então uma estimativa da função real $y = f_b(\mathbf{x})$, aprendida por um algoritmo de aprendizagem supervisionada a partir de exemplos de partidas fornecidas durante o treinamento. Na literatura, pode-se utilizar algoritmos de classificação ou de regressão, dependendo da modelagem escolhida. Neste estudo, o modelo base é um classificador multiclasses, que retorna propabilidades associadas a cada resultado da partida. Essa modelagem é mais adequada para identificar partidas de maior incerteza, refletida através de probabilidades de classes.

3.2. Assessor

O modelo assessor recebe como entrada o vetor de características da partida $\mathbf{x}$ e as probabilidades $\hat{\mathbf{p}}$ estimadas pelo modelo base. Como variável alvo, o modelo assessor tenta predizer o acerto do modelo base. Nesse caso, o assessor é um modelo de classificação treinando sobre os resultados do modelo base para partidas já realizadas. Mais especificamente, a cada rodada de um campeonato, o modelo base é monitorado e os seus acertos ou erros são registrados para as partidas realizadas. No exemplo de treinamento para o assessor, a variável alvo é um atributo categórico indicando se o modelo base acertou ou errou o resultado da partida.

Após o processo de treinamento, espera-se que o assessor seja capaz de identificar partidas em que o modelo base tenha tido maior dificuldade de traçar uma fronteira de decisão adequada ou tenha uma confiança exagerada na sua predição. Um escore final de confiança é gerado usando a probabilidade de acerto estimada pelo assessor:

$$\hat{p}_a = f_a(\mathbf{x}, \hat{\mathbf{p}}) \tag{2}$$

onde $\hat{p}_a$ é a probabilidade da classe positiva (i.e., probabilidade de acerto do modelo base) estimada pelo assessor f_a. A partir desse conhecimento, utiliza-se o assessor para avaliar

Tabela 1. Campeonatos analisados.

Campeonato	N° de rodadas	N° de partidas
Alemanha	34	306
Espanha	38	380
França	38	380
Inglaterra	38	380
Itália	38	380
TOTAL	**186**	**1826**

futuras instâncias apresentadas ao modelo base, junto às saídas associadas a elas, e, se necessário, barrar predições julgadas como de má qualidade. Mais especificamente, a predição de uma partida é rejeitada se o escore retornado pelo assessor for menor que um dado limiar de aceitação. Opcionalmente, pode-se rejeitar um determinado percentual das partidas menos confiáveis. Essa seria uma opção, por exemplo, quando os modelos de predição fossem usados para realizar um número pré-definido de apostas, e nesse caso, as apostas priorizadas seriam aquelas para as partidas mais confiáveis, conforme o assessor.

4. Experimentos

Nessa seção, descrevemos os experimentos realizados para avaliar o uso de assessores como opção de rejeição no problema de predição de resultados de partidas. Inicialmente, apresentamos as bases de dados para experimentos (seção 4.1). Em seguida, apresentamos o modelo base que será monitorado pelos assessores (seção 4.2). Em seguida, apresentamos o treinamento dos modelos assessores e o baseline de comparação (seção 4.3). Por fim, apresentamos as métricas de avaliação (seção 4.4).

4.1. Bases de dados

As bases de dados utilizadas nos experimentos foram coletadas por [Pappalardo et al. 2019], através da plataforma WyScout, que contém dados sobre partidas da temporada 2017/2018 de várias competições europeias de futebol. No nosso trabalho, foram consideradas as partidas dos campeonatos nacionais da primeira divisão de cinco países: Alemanha, Espanha, França, Inglaterra e Itália.

Para caracterização das partidas, foram considerados nos experimentos três tipos distintos de atributos, apresentados a seguir.

(a) Estatísticas gerais e de desempenho

- Informações gerais da partida: o time da casa, o visitante, a rodada e a data do confronto, o vencedor e o campeonato;
- Estatísticas médias das equipes: média de vitórias, derrotas, gols sofridos, gols marcados, dentre outras informações calculadas até a rodada da partida;
- Estatísticas médias recentes: mesmas estatísticas do item anterior, porém calculadas para os últimos seis jogos da equipe;
- Estatísticas médias de jogo: número de passes trocados, chutes ao gol, passes no terço final, oportunidades criadas, dentre outras, também calculadas para os seis jogos mais recentes de cada clube.

Para garantir que as estatísticas referentes ao momento recente das equipes englobassem seis partidas inteiras, optou-se por descartar as a partidas anteriores à sétima rodada de cada campeonato.

(b) Rating ELO

O índice utilizado para medir a força geral das equipes foi o rating ELO. Em cada uma das partidas, foi adicionado o valor referente ao ELO das equipes na semana que antecedeu o confronto, de modo a garantir que os valores não tivessem incorporado o resultado do evento. Esses dados foram obtidos por meio de chamadas para API *ClubElo*.

(c) Casas de Apostas

Finalmente, coletou-se, no agregador de dados online *www.football-data.co.uk*, os múltiplos de oito casas de apostas para cada um dos resultados possíveis das partidas escolhidas (vitória do mandante, empate, vitória do visitante) na véspera do jogo. Por fim, calculou-se a média dos múltiplos para cada um dos desfechos das partidas e incorporou-os à tabela utilizada.

Como resultado do processamento, produziu-se uma base de dados com 1571 partidas, cada uma delas caracterizadas por 89 atributos.

4.2. Modelo base

O modelo base escolhido para fazer as predições dos resultados dos jogos foi o *Random Forest Classifier*, disponibilizado pela biblioteca *Scikit-Learn*. O treinamento do modelo base foi realizado a cada rodada, utilizando como conjunto de treino todas as partidas ocorridas, de todos os campeonatos presentes na base de dados, até o dia anterior ao começo da rodada. Assim, garantiu-se que o algoritmo não recebesse informações referentes a partidas ocorridas após os eventos de teste. As métricas de avaliação de desempenho preditivo são calculadas para os resultados para cada uma das rodadas dos campeonatos.

Como dito, o modelo base é um classificador para três classes de resultados: vitória do mandante (m), empate (e) ou vitória do visitante (v). A cada predição, foram guardadas a classe predita $\hat{y}$ e as probabilidades retornadas pelo modelo $\hat{\mathbf{p}}$ para cálculo dos erros do modelo após a finalização da rodada.

4.3. Modelo assessor e baseline

O modelo assessor foi treinado utilizando uma metodologia similar ao do modelo base. A cada rodada, foram utilizados para amostra de treinamento todas as partidas anteriores a rodada. No caso dos assessores, os dados de entrada são os mesmos utilizados pelo modelo base, acrescidos pelas probabilidades dadas para cada resultado possível da partida. Como saída, temos um assessor que estima a probabilidade do modelo base acertar o resultado de uma partida. Assim como no modelo base, o assessor foi construído usando o algoritmo de Random Forests.

Como baseline de comparação, utilizou-se o próprio valor de confiança fornecido pelo modelo base para as predições, definido neste trabalho como a máxima probabilidade estimada pelo modelo base: $\max_y \hat{p}_y$, onde $y \in \{m, e, v\}$.

4.4. Métricas de avaliação

A curva de Acurácia x Rejeição (ARC) é utilizada para avaliação e comparação de modelos com opção de rejeição [Geifman e El-Yaniv 2017] [da Rocha Neto et al. 2011]. Na ARC, o eixo X representa a taxa de rejeição, enquanto que o eixo Y representa a acurácia do modelo base calculada para os exemplos de teste que não foram rejeitados. Para gerar a curva, são definidos limiares crescentes de confiança para que uma predição possa ser aceita e, para cada limiar, é calculada a taxa de rejeição e a acurácia para exemplos aceitos.

Quando a taxa de rejeição é zero, a acurácia obtida é uma referência de desempenho quando um modelo base é avaliado utilizando todos os exemplos de teste. À medida que a taxa de rejeição aumenta, a tendência é que a acurácia aumente, uma vez que o modelo base será aplicado a exemplos de teste supostamente menos incertos. Em compensação, uma taxa de rejeição muito alta não é ideal em aplicações práticas. A área sob a ARC pode ser calculada então para resolver esse conflito, agregando assim a acurácia ao longo dos valores de taxa de rejeição. Desta forma, tem-se uma medida de desempenho agregada, de maneira agnóstica à taxa de rejeição.

5. Resultados

Nessa seção, apresentamos os resultados obtidos com o modelo assessor. Inicialmente, apresentamos os resultados gerais e por liga (Seção 5.1). Em seguida, detalhamos os resultados obtidos pelo modelo assessor ao longo dos campeonatos (Seção 5.2). Finalmente, apresentamos uma análise da acurácia obtida pelo modelo base, para as top-10 partidas aceitas pelo assessor (Seção 5.3).

5.1. Desempenho geral e por campeonato

A Tabela 2 (segunda coluna) apresenta os valores de Acurácia obtidos pelo modelo base por liga, além do resultado global. Observa-se que o desempenho do modelo base pode variar por liga. Por exemplo, para a liga alemã o modelo base teve uma acurácia de 0,414, enquanto que para o campeonato italiano, a acucácia foi 0,550. Existe um fator de incerteza alto e uma variabilidade da qualidade das predições do modelo base dependendo do contexto. Essa observação motiva o uso de modelos de rejeição.

A terceira e quarta coluna da tabela apresenta ainda o desempenho dos assessores e do baseline de comparação, considerando a métricas de AU-ARC. As curvas em si são apresentadas na Figura 2 (resultado geral) e na Figura 2 (resultado por liga). Observa-se que o modelo assessor melhorou a métrica AU-ARC em todos os casos. Isso significa que a seleção de predições de melhor qualidade foi mais efetiva quando feita através de um modelo de AM do que quando filtrado apenas pelo escore de confiança do modelo base. Embora os ganhos tenham sido pequenos em termos absolutos, eles foram consistentes, considerando melhoria para todas as ligas. O teste de Wilcoxon foi aplicado para comparar as diferenças de resultados ao longo das ligas obtidos pelo assessor e pelo baseline. Foi obtido um p-valor de 0,031, e assim, foi constatado que a diferença entre os métodos é estatisticamente relevante para uma significância estatística de 95%.

Considerando as curvas ARC, no resultado geral, essa diferença se torna mais clara perto da taxa de rejeição de 35%. Na Figura 3, as curvas com maiores ganhos para o assessor em relação ao baseline foram obtidas para as ligas Francesa, Inglesa e Italiana,

Tabela 2. Métricas do experimento.

Competição	Acurácia	AU-ARC	
	Modelo Base	Base	Assessor
Geral	0,497	0,630	**0,647**
Alemanha	0,414	0,554	**0,564**
Espanha	0,491	0,613	**0,616**
França	0,500	0,628	**0,657**
Inglaterra	0,518	0,650	**0,674**
Itália	0,550	0,675	**0,698**

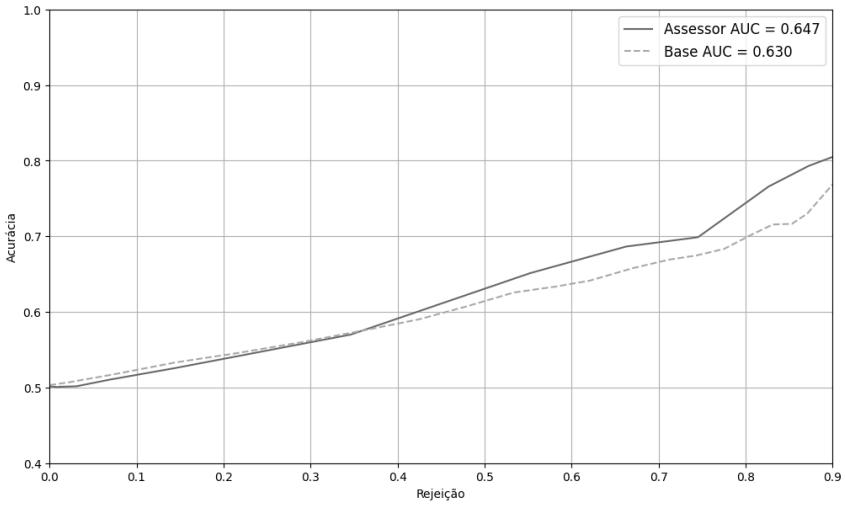

Figura 2. Curva de acurácia-rejeição para cada método avaliado

respectivamente. Já o resultado menos consistente ocorreu para a liga Espanhola. Para as ligas da Inglaterra e da França, nota-se que, mesmo saindo de um patamar mais alto de acurácia, o descarte das piores predições pelo assessor foi consistentemente melhor que pela confiança do modelo base. A liga Italiana possui comportamento semelhante, embora a curva do assessor só ultrapasse o do baseline perto dos 30% de rejeição. Para a liga Alemã, a curva do assessor começa em um nível baixo de acurácia, mas recupera seu desempenho em relação ao modelo base, passando-o quando a taxa de rejeição atinge cerca de 50%.

5.1.1. Desempenho por faixa de rodadas

Nessa seção, avaliamos a qualidade do assessor em identificar situações de incerteza ao longo das rodadas. Para isso geramos as curvas ARC, mas restritas a partidas realizadas em faixas diferentes de rodadas: (1) rodadas 7 a 14; (2) rodadas 15 a 22; (3) rodadas 23 a 30; e (4) rodadas 31 a 38. O pior resultado do assessor foi obtido para a primeira faixa avaliada, com um AU-ARC de 0,608, abaixo da média geral. Já o melhor desempenho se dá na faixa intermediária de 23 a 30, com a área de 0,688. O melhor resultado seguinte foi obtido para a última faixa, apresentando uma métrica similar a geral, porém a sua curva apresenta uma estagnação a partir da taxa de rejeição de 45% e possui a menor acurácia

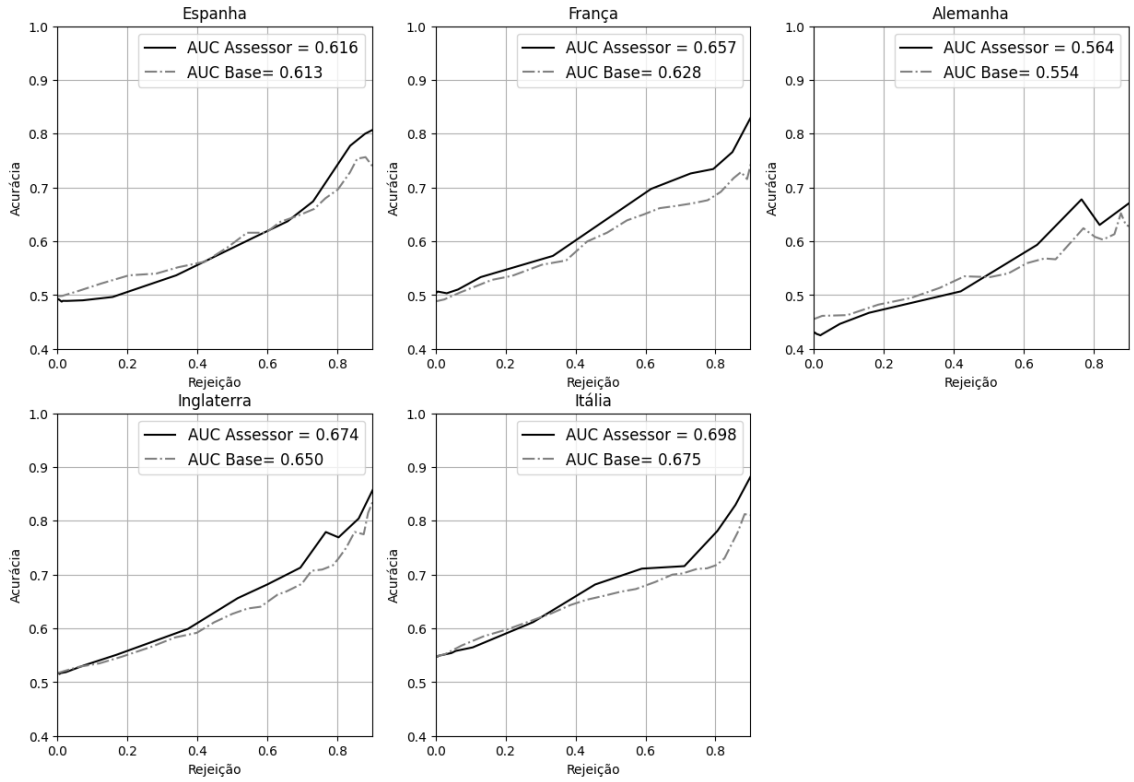

Figura 3. Curvas de acurácia-rejeição por liga

para uma taxa de 90% de descartes.

Os resultados por rodada indicam que o assessor ganha desempenho à medida que a base de dados de treino aumenta, mas que essa tendência é quebrada ao fim do campeonato, provavelmente devido à própria dinâmica dos torneios e um aumento na incerteza nas partidas. Isso ocorre porque, ao fim da competição, times de desempenho médio normalmente estão fora das disputas mais importantes, enquanto times nas últimas colocações ficam mais motivados, pois precisam escapar de um eventual rebaixamento. Esses fatores externos ocorrem com menor intensidade no resto da temporada e diminuem a previsibilidade dos jogos.

5.1.2. Ranqueamento de predições

Na Figura 5, apresentamos a acurácia pelo modelo base para as partidas em cada posição do ranking de confiança gerado pelo assessor. Nota-se uma maior acurácia nas primeiras posições. Além disso, observa-se um decréscimo gradual na acurácia da predição entre os cinco primeiras posições do ranking. Isso indica um bom desempenho do modelo ao graduar sua confiança nesses resultados.

A partir desse ordenamento, também foi possível detectar os times que o assessor mais recomendou entre os três vencedores mais prováveis da rodada. Nota-se pela Tabela 6 que alguns times foram menos suscetíveis a resultados surpreendentes, as famosas ”zebras”, como o Bayern de Munique, Manchester City, Lyon e Juventus, pois venceram mais de 80% das partidas em que foram indicados como grandes favoritos da rodada. Já o

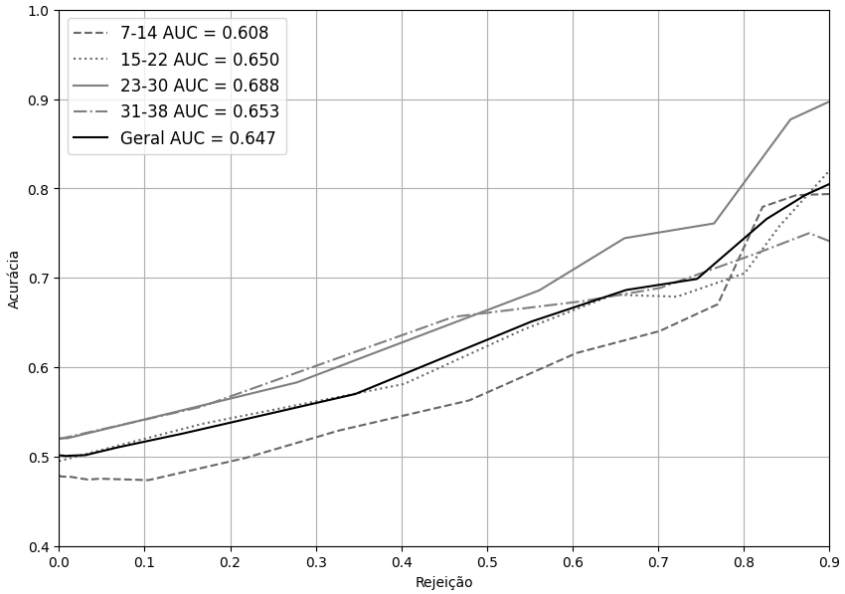

Figura 4. Curvas de acurácia-rejeição por momento do campeonato

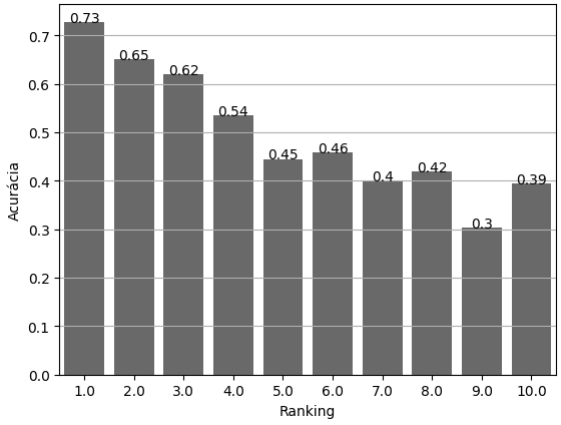

Figura 5. Acurácia para cada posição do ranqueamento por rodada

Borussia Dortmund não correspondeu às expectativas do modelo em mais da metade das ocasiões, visto que ganhou apenas 44% das partidas quando esteve entre os três favoritos da rodada de sua liga.

6. Conclusão

A utilização do modelo assessor mostrou-se mais eficaz na identificação e rejeição das previsões de menor qualidade do que o método de rejeição baseado em limiares diretamente aplicados aos valores de confiança fornecidos pelo modelo base. Isso resultou em um aumento na AU-ARC, tanto no geral (de 0,630 para 0,647), quanto ao analisar cada liga individualmente. O melhor desempenho foi obtido na liga italiana, alcançando um AU-ARC de 0,698 com a introdução do modelo de rejeição, superando os 0,675 alcançados pela seleção baseada na confiança do modelo base. Outras ligas que se beneficiaram com o modelo de avaliação foram as ligas francesa e inglesa, cujas curvas de AU-ARC do modelo de avaliação superaram as do modelo principal para todas as taxas

Tabela 3. Equipes mais vezes no top 3 vencedores mais confiáveis da rodada na sua liga.

Time	N° de vezes no top 3	% de acerto
Barcelona	30	73.33%
Napoli	29	72.41%
PSG	29	75.86%
Bayern de Munique	28	82.14%
Real Madrid	27	66.67%
Manchester City	25	84.00%
Liverpool	19	63.15%
Borussia Dortmund	18	44.44%
Lyon	17	81.25%
Juventus	16	81.25%

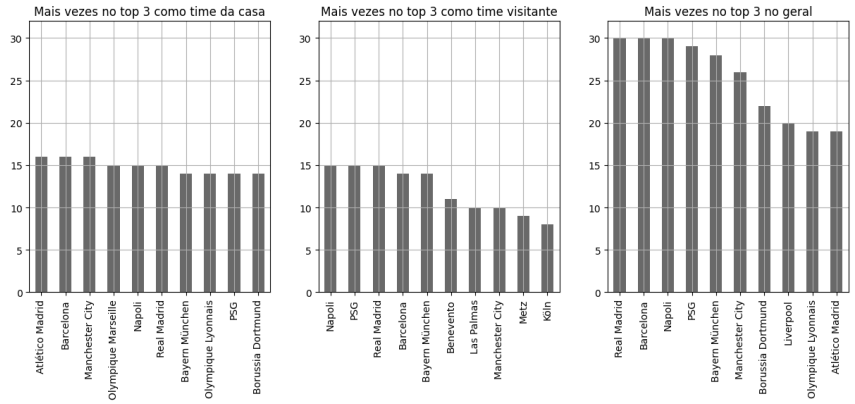

Figura 6. Aparições entre os 3 vencedores mais confiáveis da rodada

de rejeição analisadas.

Conclui-se, portanto, que o escore atribuído pelo modelo de rejeição refletiu de forma mais precisa a probabilidade de sucesso de uma previsão em comparação ao valor obtido diretamente pelo modelo base. Como resultado, torna-se possível realizar uma seleção mais eficaz das instâncias de maior qualidade e reduzir a quantidade necessária de rejeições para que o sistema atinja uma determinada acurácia.

Para estudos futuros sobre previsão de resultados no futebol com opção de rejeição, é possível aprimorar o desempenho preditivo ao utilizar uma base de treinamento mais ampla, abrangendo um maior número de campeonatos e temporadas, porém mantendo o alto nível de detalhamento presente na base de dados utilizada neste estudo. Além disso, pode-se explorar outras abordagens para o desenvolvimento do modelo de rejeição, como a previsão da taxa de erro das previsões por meio de algoritmos de regressão. Também é possível o aprimoramento do desempenho tanto do modelo base como do assessor, utilizando outros algoritmos de AM, assim como métodos de otimização de hiperparâmetros. Outra estratégia para melhoria de desempenho é selecionar os atributos mais relevantes tanto para o modelo assessor quanto para o modelo base.

Referências

Brooks, J., Kerr, M., e Guttag, J. (2016). Using machine learning to draw inferences from pass location data in soccer. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 9(5):338–349.

Chow, C. (1970). On optimum recognition error and reject tradeoff. *IEEE Transactions on information theory*, 16(1):41–46.

Constantinou, A. C. (2019). Dolores: a model that predicts football match outcomes from all over the world. *Machine Learning*, 108(1):49–75.

da Rocha Neto, A. R., Sousa, R., de A. Barreto, G., e Cardoso, J. S. (2011). Diagnostic of pathology on the vertebral column with embedded reject option. In *Pattern Recognition and Image Analysis: 5th Iberian Conference, IbPRIA 2011, Las Palmas de Gran Canaria, Spain, June 8-10, 2011. Proceedings 5*, pages 588–595. Springer.

Deloitte, U. (2020). Deloitte football money league.

FIFA (2018). More than half the world watched record-breaking 2018 world cup. *FIFA*.

Geifman, Y. e El-Yaniv, R. (2017). Selective classification for deep neural networks. *Advances in neural information processing systems*, 30.

Godin, F., Zuallaert, J., Vandersmissen, B., De Neve, W., e Van de Walle, R. (2014). Beating the bookmakers: leveraging statistics and twitter microposts for predicting soccer results. In *KDD Workshop on large-scale sports analytics*, pages 2–14. ACM New York, NY, USA.

Hendrickx, K., Perini, L., Van der Plas, D., Meert, W., e Davis, J. (2021). Machine learning with a reject option: A survey. *arXiv preprint arXiv:2107.11277*.

Hernández-Orallo, J., Schellaert, W., e Martínez-Plumed, F. (2022). Training on the test set: Mapping the system-problem space in ai. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 12256–12261.

Hubáček, O., Šourek, G., e Železnỳ, F. (2019). Learning to predict soccer results from relational data with gradient boosted trees. *Machine Learning*, 108:29–47.

Hucaljuk, J. e Rakipović, A. (2011). Predicting football scores using machine learning techniques. In *2011 Proceedings of the 34th International Convention MIPRO*, pages 1623–1627. IEEE.

Hüllermeier, E. e Waegeman, W. (2021). Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine Learning*, 110:457–506.

Jiang, H., Kim, B., Guan, M., e Gupta, M. (2018). To trust or not to trust a classifier. *Advances in neural information processing systems*, 31.

Nicora, G., Rios, M., Abu-Hanna, A., e Bellazzi, R. (2022). Evaluating pointwise reliability of machine learning prediction. *Journal of Biomedical Informatics*, 127:103996.

Pappalardo, L., Cintia, P., Rossi, A., Massucco, E., Ferragina, P., Pedreschi, D., e Giannotti, F. (2019). A public data set of spatio-temporal match events in soccer competitions. *Scientific data*, 6(1):236.

Partida, A., Martinez, A., Durrer, C., Gutierrez, O., e Posta, F. (2021). Modeling of football match outcomes with expected goals statistic. *Journal of Student Research*, 10(1).

Rossi, A., Pappalardo, L., Cintia, P., Iaia, F. M., Fernández, J., e Medina, D. (2018). Effective injury forecasting in soccer with gps training data and machine learning. *PloS one*, 13(7):e0201264.

Stübinger, J., Mangold, B., e Knoll, J. (2019). Machine learning in football betting: Prediction of match results based on player characteristics. *Applied Sciences*, 10(1):46.

Tax, N. e Joustra, Y. (2015). Predicting the dutch football competition using public data: A machine learning approach. *Transactions on knowledge and data engineering*, 10(10):1–13.

Zhou, L., Martinez-Plumed, F., Hernández-Orallo, J., Ferri, C., e Schellaert, W. (2022). Reject before you run: Small assessors anticipate big language models. In *Proceedings of the EBeM22,IJCAI Workshop on AI Evaluation Beyond Metrics Intelligence*.