UNIVERSIDADE FEDERAL DE PERNAMBUCO
CENTRO DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

DANIEL CABRAL DA COSTA

Modelos assessores como opção de rejeição para predição de resultados de partidas de futebol

Recife

2025

DANIEL CABRAL DA COSTA

Modelos assessores como opção de rejeição para predição de resultados de partidas de futebol

Trabalho apresentado ao Programa de Pós-graduação em Ciência da Computação do Centro de Informática da Universidade Federal de Pernambuco, como requisito parcial para obtenção do grau de Mestre em Ciência da Computação.

Área de Concentração: Inteligência Computacional

Orientador (a): Ricardo Bastos Cavalcante Prudêncio

Coorientador (a): Alexandre Cabral Mota

Recife

2025

.Catalogação de Publicação na Fonte. UFPE - Biblioteca Central

Costa, Daniel Cabral da.
Modelos assessores como opção de rejeição para predição de resultados de partidas de futebol / Daniel Cabral da Costa. - Recife, 2024.
71f.: il.

Dissertação (Mestrado) - Universidade Federal de Pernambuco, Centro de Informática, Programa de Pós-graduação em Ciência da Computação, 2024.
Orientação: Ricardo Bastos Cavalcante Prudencio.
Coorientação: Alexandre Cabral Mota.
Inclui referências e apêndice.

1. Aprendizagem de máquina; 2. Opção de rejeição; 3. Futebol. I. Prudencio, Ricardo Bastos Cavalcante. II. Mota, Alexandre Cabral. III. Título.

UFPE-Biblioteca Central

Daniel Cabral da Costa

“Modelos assessores como opção de rejeição para predição de resultados de partidas de futebol”

Dissertação de mestrado apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de Pernambuco, como requisito parcial para a obtenção do título de Mestre em Ciência da Computação. Área de Concentração: Inteligência Computacional.

Aprovado em: 09/12/2024.

BANCA EXAMINADORA

_____
Prof. Dr. George Darmiton da Cunha Cavalcanti
Centro de Informática / UFPE

_____
Prof. Dr. André Câmara Alves do Nascimento
Departamento de Computação / UFRPE

_____
Prof. Dr. Ricardo Bastos Cavalcante Prudêncio
Centro de Informática / UFPE
(**orientador**)

Dedico esta dissertação aos meus pais, Fernando e Carol.

AGRADECIMENTOS

Quero agradecer aos meus pais, Fernando e Carol, à minha irmã, Carina, e à minha namorada, Bia, por todo o apoio durante essa jornada. Passei por diversas mudanças durante o mestrado, como de área de atuação, emprego e cidade, mas o amor e o suporte oferecido por eles foram constantes e essenciais para que pudesse seguir em frente.

Também agradeço aos meus dois orientadores: Professor Alexandre Mota e Professor Ricardo Prudêncio. Sou grato a ambos pela paciência, compreensão e orientação nesse processo. O apoio de vocês foi essencial para a construção desse projeto e para que chegassemos até aqui.

Agredeço aos meus familiares e amigos, que foram fontes de apoio e alegria nos últimos três anos.

Por fim, agradeço aos CIn e a todos os seus professores pela formação e estrutura concedidas no programa de mestrado, assim como aos professores da UFPB com quem comecei minha jornada acadêmica.

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 alta incerteza do futebol torna desafiador obter previsões precisas e confiáveis. Nesta dissertação, investigamos técnicas de AM com opção de rejeição no contexto de predição de resultados de partidas de futebol. O objetivo principal deste trabalho é quantificar a incerteza de predições de modelos de AM e, assim, abster-se das predições consideradas mais incertas. Especificamente, propomos a utilização de modelos chamados assessores, modelos de AM que predizem o desempenho de um modelo base em tarefas específicas, para analisar as predições de um classificador de resultados de partidas e 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.

Palavras-chaves: Aprendizado de Máquina. Opção de Rejeição. Modelos Assessores. Futebol. Predição

ABSTRACT

Soccer is a widely popular sport, both in Brazil and worldwide, with a multi-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 high uncertainty of soccer makes it challenging to obtain accurate and reliable predictions. In this dissertation, we investigate ML techniques with rejection option in the context of soccer match prediction. The main goal of this topic is to quantify the uncertainty of ML model predictions and then abstain from predictions considered most uncertain. Specifically, we propose the use of models called assessors, ML models that predict the performance of a base model on specific tasks, to analyze the predictions of a match prediction classifier and select those with the highest reliability, discarding the others. We seek to optimize the relationship between the accuracy of accepted predictions and the rejection rate, in order to maximize confidence in the use of the ML model for match prediction. We experimented with real match data, identifying the championships, teams, and rounds in which the proposed model performs best. This innovative approach contributes to the improvement of soccer outcome predictions, using advanced ML techniques to select high-quality predictions.

Keywords: Machine Learning. Reject Option. Assessor Models. Football. Prediction.

LISTA DE FIGURAS

LISTA DE QUADROS

SUMÁRIO

1 INTRODUÇÃO

A área de Aprendizado de Máquina (AM) tem testemunhado um notável crescimento de interesse nos últimos anos, com suas técnicas sendo amplamente aplicadas em diversos domínios de estudo, resultando em significativos impactos tanto econômicos quanto humanos (IT, 2023; BRYANT, 2023). O aumento da relevância deste campo suscita questionamentos cruciais acerca da confiabilidade das decisões automáticas e, consequentemente, da necessidade de minimizar potenciais erros inerentes a essas técnicas.

Diante do cenário atual, em que a AM desempenha um papel crucial em várias indústrias, garantir predições confiáveis torna-se um ponto crítico. Existem problemas desafiadores em que se investiga o uso de métodos de AM com opção de rejeição (HENDRICKX et al., 2021) para quantificar a incerteza de predições e então se abster de realizar predições menos confiáveis. Essa dissertação investiga a aplicação de métodos de AM com opção de rejeição, mais especificamente baseados em modelos assessores (HERNÁNDEZ-ORALLO; SCHELLAERT; MARTÍNEZ-PLUMED, 2022), no contexto de predição de resultados de partidas de futebol. Veremos que de fato esse é um contexto de aplicação desafiador, com alto grau de incerteza de predições, dependendo das partidas (BUNKER; SUSNJAK, 2022).

Esse capítulo está organizado como se segue. Inicialmente, introduzimos na Seção 1.1 o contexto da pesquisa sobre AM com opção de rejeição. Em seguida na Seção 1.2, motivamos a aplicação de técnicas de rejeição no contexto de prediçao de partidas de futebol. A partir daí, são destacados os objetivos da dissertação (Seção 1.3), seguido pela proposta de trabalho (Seção 1.4). Finalmente na Seção 1.5, descrevemos o conteúdo do restante da dissertação

1.1 CONTEXTO DA PESQUISA

AM com opção de rejeição parte da ideia central de reconhecer situações em que modelos de AM apresentam baixa confiança para fazer uma previsão ou tomar uma decisão. Isso adiciona uma camada de cautela, permitindo que o sistema rejeite a resposta quando a incerteza é alta (CHOW, 1957; HENDRICKX et al., 2021; HERBEI; WEGKAMP, 2006; BARTLETT; WEGKAMP, 2008; JIANG et al., 2018; GEIFMAN; EL-YANIV, 2017). Essa característica é crucial em domínios onde erros de classificação podem acarretar consequências graves, como diagnósticos médicos (NETO et al., 2011; KOMPA; SNOEK; BEAM, 2021). A opção de rejeição permite que o modelo

transfira a decisão para um especialista humano em casos de incerteza ou que apenas se abstenha naquela situação, informando apenas predições que julgue confiáveis. De modo geral, a abstenção ocorre quando a instância a ser classificada apresenta ambiguidade em relação às classes aprendidas ou se distancia significativamente dos dados de treinamento, caracterizando uma novidade (HENDRICKX et al., 2021).

Ainda nesse contexto, Hernández-Orallo, Schellaert e Martínez-Plumed (2022) propõem o uso de modelos de AM chamados Assessores, que são estimadores que predizem o desempenho de um modelo de inteligência artificial em uma tarefa específica. Essa capacidade preditiva se baseia no perfil do sistema e na instância do problema em questão, permitindo uma avaliação da performance do modelo. No contexto da Aprendizagem de Máquina com opção de rejeição, os modelos assessores podem ser utilizados para aprender a relação entre instâncias e a probabilidade de sucesso de um modelo base e, assim, identificar regiões de incerteza que o próprio modelo não reconhece, rejeitando predições potencialmente falhas e melhorando a confiabilidade do sistema. Além disso, a natureza explicável desses modelos permite a inspeção de suas decisões, fornecendo insights valiosos para o aprimoramento contínuo do modelo base. Zhou et al. (2022) os utiliza no âmbito das LLMs para rejeitar prompts que geram respostas de baixa qualidade. Tal abordagem surge como uma alternativa promissora, visto que introduz um segundo modelo a um sistema de predição capaz de aprender os pontos fracos do estimador principal e, a partir daí, rejeitar suas predições mais incertas.

1.2 MOTIVAÇÃO E APLICAÇÃO

Uma área de aplicação de AM que convive com alto grau de incerteza devido à sua imprevisibilidade e que pode se beneficiar de uma opção de rejeição é o de predições de resultados de partidas de futebol (BEAL; NORMAN; RAMCHURN, 2019). Por ser o o esporte mais acompanhado nacional e internacionalmente, o futebol recebe muita atenção da indústria e da academia. Sua relevância é atestada pelos números da sua principal competição, a Copa do Mundo da FIFA, que, segundo a própria entidade, atingiu cinco bilhões de pessoas durante sua última edição, sediada no Qatar. Apenas a final entre Argentina e França obteve uma audiência global de 1,5 bilhão de espectadores (FIFA, 2022). Além disso, apenas os vinte clubes mais ricos do mundo geraram incríveis 10,6 bilhões de dólares em receitas no ano de 2020 (DELOITTE, 2020). Essa magnitude de interesse resulta na formação de uma indústria bilionária que engloba patrocínios, vendas de jogadores e produtos, direitos de transmissão,

entre outros.

Na área acadêmica, os primeiros estudos concentraram-se em modelos estatísticos tradicionais, como regressões a partir da distribuição de Poisson, para estimar resultados com base em variáveis como o número de gols (MAHER, 1982; DIXON; COLES, 1997; GODDARD, 2005). Com o tempo, as abordagens evoluíram para o uso de modelos de AM supervisionados. Esses modelos incorporam atributos variados, desde estatísticas de desempenho das equipes - como número de chutes, passes e vitórias - até métricas avançadas como indicadores de força relativa das equipes (*ratings* Elo) e probabilidades (múltiplos) fornecidas por casas de apostas. A combinação dessas técnicas tem permitido maior precisão nas previsões, destacando-se na literatura como ferramentas essenciais para a análise preditiva no futebol (HUCALJUK; RAKIPOVIĆ, 2011; HVATTUM; ARNTZEN, 2010; CONSTANTINOU, 2019; HUBÁČEK; ŠOUREK; ŽELEZNỲ, 2019; PARTIDA et al., 2021; BERRAR; LOPES; DUBITZKY, 2019).

Entretanto, a grande maioria dos estudos relacionados ao tema se limitam a gerar estimativas para todas as partidas analisadas, mesmo naquelas de grande equilibrio em que o modelo não consegue definir um favorito claro. (MAHER, 1982; DIXON; COLES, 1997; HUCALJUK; RAKIPOVIĆ, 2011; GODIN et al., 2014; TAX; JOUSTRA, 2015; CONSTANTINOU, 2019; HVATTUM; ARNTZEN, 2010; GODDARD, 2005; BERRAR; LOPES; DUBITZKY, 2019). Ainda que métodos de rejeição com regras baseadas nas saídas do próprio modelo já tenham sido utilizadas para esse fim (PARTIDA et al., 2021) (STÜBINGER; MANGOLD; KNOLL, 2019), percebe-se que são poucos os autores que exploram a opção de rejeitar as predições de menor qualidade do modelo de modo a melhorar a acurácia do sistema e diminuir riscos.

Tendo em vista o baixo número de trabalhos de AM com opção de rejeição para predições de resultados de partidas de futebol, esta dissertação propões expandir as análises realizadas anteriormente no assunto, introduzindo especificamente o uso de modelos assessores. Além disso, o presente estudo agrega diferentes tipos de atributos utilizados na literatura para a predição de resultados de partidas de futebol, como estatísticas históricas das equipes, medidas de força em relação aos competidores (*ratings* Elo) e múltiplos de casas de apostas, que indicam a probabilidade de cada resultado segundo as bancas. Esses atributos fornecem informações cruciais sobre o desempenho esperado das equipes, e sua combinação pode influenciar significativamente a acurácia do modelo. A análise da importância desses atributos no contexto dos modelos base e assessores permite uma compreensão mais aprofundada de quais fatores são mais determinantes para a predição precisa de resultados no futebol.

1.3 OBJETIVOS

1.3.1 Objetivo geral

A presente dissertação busca investigar o uso de modelos assessores para AM com opção de rejeição no âmbito das predições de resultados de partidas de futebol.

1.3.2 Objetivos específicos

- Investigar diferentes algoritmos no papel de modelo base e modelo assessor, de acordo com sua força (fraco, médio e forte). Nesse ponto, investigaremos uma hipótese não explorada na literatura de que o ganho proveniente do uso do assessor varia de acordo com a qualidade do modelo base e da complexidade do próprio assessor;
- Analisar o comportamento dos modelos experimentalmente para diferentes ligas e fases dos torneios utilizados, fazendo a comparação com *baselines* da literatura de AM com opção de rejeição (JIANG et al., 2018) e com um método heurístico.
- Observar quais são as diferenças entre os atributos mais importantes considerados pelo assessor e pelo modelo base.

1.4 DESENVOLVIMENTO DE TRABALHO

O presente estudo concentrou-se na aplicação de modelos assessores para a predição de resultados em partidas de futebol das cinco principais ligas europeias (Alemanha, Espanha, França, Inglaterra e Itália) na temporada 2017/2018. A base de dados utilizada compreendeu informações detalhadas sobre os eventos das partidas, estatísticas das equipes, *ratings* Elo e múltiplos de apostas.

A metodologia experimental envolveu o treinamento e comparação de modelos base e assessores com diferentes níveis de complexidade (fraco, médio e forte). Os modelos base foram treinados utilizando dados históricos das partidas, obtendo como saída a predição de resultado da partida. Por sua vez, os assessores foram treinados a partir dos atributos de uma partida e o resultado predito pelo modelo base, obtendo como rótulo um escore de confiança de que aquela predição foi correta. A avaliação do desempenho foi realizada através de métricas

específicas para sistemas com opção de rejeição, a curva de acurácia-rejeição (ARC) e da área sob a curva (AUARC), que permitiram analisar o trade-off entre a taxa de rejeição e a acurácia das predições aceitas.

Os resultados demonstraram que os modelos assessores podem melhorar as métricas analisadas, especialmente quando o modelo base apresenta um desempenho mais fraco. O assessor médio, por exemplo, aumentou a performance do sistema em relação ao modelo base fraco, enquanto o assessor forte obteve resultados superiores ao modelo base forte em determinados períodos do campeonato e ligas específicas. Além disso, os assessores superaram o modelo *Trust Score* (JIANG et al., 2018) e o método baseado na taxa de vitórias do time mandante em todas as situações analisadas. Por fim, a análise da importância das variáveis revelou que os múltiplos das casas de apostas e os *ratings* Elo foram cruciais para as decisões dos modelos, destacando a relevância da inclusão desses atributos na previsão de resultados no futebol.

1.5 ORGANIZAÇÃO

Este trabalho está organizado da seguinte forma. No Capítulo 2 apresentamos a fundamentação teórica e trabalhos relacionados ao problema analisado. No Capítulo 3, explicamos apresentamos o sistema proposto, seus componentes e a metodologia utilizada no trabalho. No Capítulo 4, detalhamos o processo de treinamento dos modelos e de avaliação de resultados e em seguida, no Capítulo 5, expomos os resultados obtidos. Por fim, no Capítulo 6, discutimos nossas conclusões e principais trabalhos futuros.

2 FUNDAMENTAÇÃO TEÓRICA

Nesta seção, fazemos uma breve revisão de abordagens e estudos relacionados ao uso de dados no futebol e em outros esportes (Seção 2.1), à utilização de métodos de AM com opção de rejeição (Seção 2.2) e aos modelos assessores (Seção 2.3)

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. Esses estudos se diferenciam principalmente pela variável alvo que se deseja predizer. Por exemplo, Stival et al. (2023) analisaram os primeiros cinco segundos de posse de bola para determinar se aquela jogada alcançaria o último quarto do campo. Na mesma linha, Brooks, Kerr e Guttag (2016) tentaram prever se uma posse de bola gerou um chute através da localização dos passes de uma jogada, além de ranquear os jogadores estudados de acordo com o valor adicionado pelos seus passes. Já Rossi et al. (2018) desenvolveram um modelo que utiliza dados de GPS de treinos para prever lesões em jogadores profissionais a partir da sua carga de trabalho, um problema comumente explorado.

Nesta dissertação, focamos no problema de predição dos resultados das partidas, que é um tema que recebe grande atenção entre os trabalhos de AM para futebol. Em geral, a maioria dos trabalhos relacionados ao tema seguem a abordagem de predizer diretamente o resultado da partida, seja empate, vitória do mandante ou vitória do visitante (DIXON; COLES, 1997; HUCALJUK; RAKIPOVIĆ, 2011; HVATTUM; ARNTZEN, 2010; CONSTANTINOU, 2019; HUBÁČEK; ŠOUREK; ŽELEZNỲ, 2019). Outra abordagem encontrada é de predizer o número de gols que cada equipe irá marcar (PARTIDA et al., 2021) ou diferença de gols marcados por cada um (STÜBINGER; MANGOLD; KNOLL, 2019) e, partir daí, derivar o vencedor do confronto. Em Goddard (2005), os autores compararam as duas abordagens e concluiram que há pouca diferença entre eles, embora a abordagem de predição direta seja considerada mais simples. Por isso, essa foi a escolhida para o presente estudo.

As primeiras tentativas de predição de resultados de futebol envolviam métodos puramente estáticos, que buscavam métodos para modelar o número de gols das equipes nas partidas a partir de uma distribuição de Poisson e suas variações (MAHER, 1982; DIXON; COLES, 1997).

Com o passar do tempo, outros métodos de aprendizagem de máquina começam a ser testados com sucesso na literatura, como em Tax e Joustra (2015), que utilizam algoritmos como Redes Neurais, *Naive Bayes* e Análises de Componentes Principais para predizer resultados do campeonato holandês. Recentemente, algoritmos de ensemble baseados em árvores, como o *Random Forest* e o *xgBoost* têm se destacado em competições e estudos pela sua boa capacidade para problemas desse tipo (HUBÁČEK; ŠOUREK; ŽELEZNỲ, 2019; BERRAR; LOPES; DUBITZKY, 2019; BABOOTA; KAUR, 2019).

Um aspecto muito importante na predição é a definição dos atributos utilizados, cujos principais tipos serão apresentados na seção a seguir.

2.1.1 Atributos utilizados para predição no futebol

Os atributos utilizados para predição de resultados tentam mensurar de alguma forma a força e as características das equipes participantes na partida. Os atributos preditores podem ser calculados usando os dados históricos dos times, em longo prazo (e.g., número médio de passes do time no último campeonato), ou usando somente as últimas atuações dos times (e.g., número médio de passes do time nas últimas seis partidas).

Nesta subseção, apresentamos três categorias distintas de atributos: as estatísticas das equipes, os *ratings* Elo e os múltiplos de casas de apostas.

2.1.1.1 Estatísticas das equipes

Vários atributos podem ser construídos a partir das partidas jogadas anteriormente por cada equipe. As mais simples envolvem o cálculo de taxas históricas de desempenho dos times, como as porcentagens de vitórias, empates e derrotas, que são bons indicadores do desempenho esperado. Outra estatística comumente calculada é da média de gols marcados e sofridos, que podem indicar uma maior força ou fragilidade das equipes, além de dar mais informações sobre o seu estilo de jogo. Por exemplo, duas equipes podem ter taxas de vitórias similares, porém uma pode obtê-las através de uma grande força ofensiva e a outra através de sua força defensiva. Além disso, diversos estudos separam o cálculo de seus atributos a partir do mando de campo das partidas, visto que esse é um fator relevante no esporte, sendo comum que o desempenho das equipes varie quando jogam fora ou dentro de casa (MAHER, 1982; DIXON; COLES, 1997; GODDARD, 2005; TAX; JOUSTRA, 2015; HUBÁČEK; ŠOUREK; ŽELEZNỲ,

2019).

Dentro dessa categoria de atributos, também pode-se destacar as estatísticas de jogo, calculadas em cima dos eventos em uma dada partida, como o número de passes de cada time, a posição onde ocorreu cada evento, número de finalizações detalhadas, cruzamentos, dribles, dentre outros. Essas informações podem ser indicativos da qualidade de uma equipe e de seu estilo de jogo. Obviamente, certas informações são mais ou menos efetivas dependendo do estilo de jogo do adversário. Portanto, mapear o estilo de jogo histórico e recente de cada equipe pode ser útil para predizer o vencedor quando há uma confronto entre elas.

Apesar da utilidade dessas informações, é difícil encontrar bases de dados detalhadas com estatísticas do que ocorreu dentro de uma partida de futebol. Devido a isso, poucos estudos conseguem incluir esses atributos nas suas modelagens. Pode-se destacar Baboota e Kaur (2019), que incluem os números de chutes aos gols e escanteios cobrados nos jogos mais recentes de cada equipe. A presente dissertação buscou aferir se esses atributos têm de fato grande importância para predizer partidas de futebol.

2.1.1.2 Rating Elo

Um sistema amplamente utilizado como indicador de força relativa de um atleta ou equipe é *rating* Elo (HVATTUM; ARNTZEN, 2010) (ANGELINI; CANDILA; ANGELIS, 2022). Os *ratings* Elo são medidas numéricas que representam a habilidade relativa dos participantes em jogos de estratégia (e.g., xadrez) e em outros esportes, incluindo futebol. Eles foram inventados por Arpad Elo e são amplamente utilizados para ranquear partipantes de competições.

O sistema Elo é baseado em princípios estatísticos e matemáticos simples. A ideia principal é que a probabilidade de um jogador A derrotar um jogador B em uma partida é uma função logística da diferença de *rating* entre os dois jogadores. Quanto maior a diferença de *rating*, menor a probabilidade de o jogador com o *rating* mais baixo vencer.

A fórmula básica para calcular a probabilidade de vitória de um jogador A em relação ao outro é dada por:

$$P(A) = \frac{1}{1 + 10^{(R_B - R_A)/400}} \tag{2.1}$$

onde:

- $P(A)$ é a probabilidade de o jogador A vencer.

- R_A é o *rating* do jogador A.
- R_B é o *rating* do jogador B.

A fórmula para ajustar o *rating* Elo após uma partida é dada por:

$$R_{\text{atual}} = R_{\text{antigo}} + K \times (S - P(A)) \tag{2.2}$$

onde:

- R_{atual} é o novo *rating* do jogador.
- R_{antigo} é o *rating* anterior do jogador.
- K é um fator de ponderação que controla a magnitude da mudança na pontuação.
- S é o resultado real da partida (1 para vitória, 0 para derrota, 0.5 para empate).

No contexto do futebol, Hvattum e Arntzen (2010) utilizam a diferença entre os *ratings* Elo das equipes como variáveis de uma Regressão Logística para prever os resultados de partidas, superando outros modelos similares baseados no histórico das partidas. Entretanto, não foi capaz de superar métodos baseados nos múltiplos fornecidos por casas de apostas.

Embora não utilizem o Elo, Hubáček, Šourek e Železnỳ (2019)) e Constantinou (2019) foram vencedores da competição Machine Learning for Football, organizada pela revista Machine Learning, e deram destaque a utilização de outros atributos relacionados a força relativa das equipes, como o *pi-ratings* (CONSTANTINOU; FENTON, 2013) e o *PageRank* (BRIN; PAGE, 1998), reforçando a relevância de atributos desse tipo para a predição de resultados no futebol.

2.1.1.3 Múltiplos de apostas

Múltiplos de casas de apostas, comumente referidos como *odds*, são valores numéricos que refletem a probabilidade implícita de um determinado desfecho em um evento esportivo, como uma partida de futebol. Essas *odds* são estipuladas pelas casas de apostas e servem como base para os apostadores estimarem o potencial retorno financeiro de suas apostas (FORREST; GODDARD; SIMMONS, 2005; STÜBINGER; MANGOLD; KNOLL, 2019).

A *odd* representa o valor total que será pago ao apostador por uma unidade apostada, incluindo a devolução da aposta original. Por exemplo, se as *odds* para uma equipe vencer são de 2,50, isso indica que, para cada unidade monetária apostada, o retorno será de 2,50

unidades caso a aposta seja bem-sucedida. Embora as *odds* reflitam a probabilidade calculada pela casa de aposta para determinado evento, elas não correspondem aos valores de fato, visto que são ajustadas com o objetivo de incluir uma margem de lucro para a operação por parte da casa.

Como esses valores são frutos de modelos elaborados pelas casas de apostas e influenciadas pela opinião de apostadores, ou seja, do público geral, eles já possuem um alto nível de informação sobre as partidas resumidos em si. Assim, pode ser interessante incluí-los como atributos para predizer resultados (TAX; JOUSTRA, 2015; STÜBINGER; MANGOLD; KNOLL, 2019).

2.1.2 AM para predição de resultados em outros esportes

Outros esportes, principalmente os americanos, também possuem trabalhos relacionados ao uso de dados para predição de resultados. O beisebol, como mostrado notoriamente pelo filme "*Moneyball* - O homem que mudou o jogo", foi um dos primeiros a passar por essa revolução. Valero (2016) utiliza dez anos de dados históricos da *Major League Baseball* para predizer resultados de partidas da liga. O autor testou diferentes algoritmos, modelando o problema como classificação - se o time mandante venceu ou não - e como regressão - a diferença de pontos entre as equipes, e obteve melhores resultados utilizando um SVM para classificação.

Já no basquete, Miljković et al. (2010) realiza predições de resultados de partidas da temporada 2009/2010 da NBA. Os autores utilizam atributos relacionados a estatísticas das partidas, como número médio de arremessos de dois e três pontos, assistências, faltas e pontos por partida. Também são utilizados números relativos às campanhas do time, como número e porcentagem de vitórias e número de derrotas. Além disso, foram utilizados atributos relacionados a forma mais recente da equipe, de forma a capturar o chamado "momento", através do número de vitórias e derrotas apenas nos últimos dez jogos e do número de derrotas ou vitórias em sequência até aquele dia. Isso mostra que a utilização de atributos que mensurem a forma recente extrapola o futebol e também é importante para a predição de resultados em outros esportes coletivos.

Indo para um esporte individual, Kovalchik (2016) compara onze diferentes modelos de predição de vencedores de partidas de tênis e mostra que o modelo baseado no *rating* ELO calculado pela plataforma *FiveThirtyEight* foi a que mais chegou próximo do desempenho do consenso entre as casas de aposta.

Esses trabalhos mostram que atributos baseados nos *ratings* Elo e que refletem não apenas o histórico geral das equipes, mas também o desempenho recente são de alta relevância em diferentes esportes.

2.2 AM COM OPÇÃO DE REJEIÇÃO

Modelos de AM são ferramentas poderosas e podem ter alta acurácia dependendo do problema e se treinados adequadamente. Entretanto, não se pode garantir que o desempenho preditivo de um modelo seja homogêneo, com boa qualidade para qualquer instância a ser predita. Modelos podem gerar predições para qualquer instância de entrada que receba, mesmo que caia em uma região de dados onde não se consiga diferenciar bem as classes do problema ou em regiões de dados não vistos no período de treinamento. Nesses casos, pode ser mais interessante que o sistema se abstenha de predizer do que retornar uma predição errada. Assim, é interessante o uso de modelos de AM com opção de rejeição para casos em que o risco de erro sobre uma predição seja demasiadamente alto. (HENDRICKX et al., 2021) (GEIFMAN; EL-YANIV, 2017)

Chow (1970), Chow (1957) foi pioneiro ao discutir a opção de rejeição em sistemas de aprendizado de máquina (AM), aplicando o conceito inicialmente a cenários de reconhecimento de caracteres. Essa abordagem foi posteriormente explorada em áreas como a medicina, onde erros de predição podem gerar custos humanos e financeiros significativos (NICORA et al., 2022). A literatura destaca a existência de um trade-off entre a taxa de rejeição e a taxa de erro (HANSEN; LIISBERG; SALAMON, 1997), evidenciando que uma maior taxa de rejeição tende a aumentar a confiabilidade das predições aceitas, reduzindo a taxa de erro. Contudo, rejeitar muitas predições pode demandar intervenção humana para a classificação final, o que pode resultar em custos operacionais elevados.

2.2.1 Tipos de rejeição

De modo geral, existem dois casos em que um modelo possui maiores riscos de classificar incorretamente uma amostra. A primeira delas é a ambiguidade. que ocorre quando um modelo tem dificuldade em distinguir entre as classes de um problema de predição. Por exemplo, um modelo que classifica imagens de animais como “gato” ou “cachorro” pode apresentar baixa confiança em sua previsão se uma imagem contiver um animal com características ambíguas.

Um método simples de representar uma opção de rejeição para casos ambíguos em um sistema de decisão é através da regra do *plug-in* (CHOW, 1970; HERBEI; WEGKAMP, 2006). Considere, por exemplo, um problema de classificação binário, onde um modelo base retorna como saída uma probabilidade $\hat{p}$ para a classe positiva. A nova saída do modelo usando opção de rejeição pode ser definida da seguinte forma:

$$\hat{f}(x) = \begin{cases} 1 & : \ \hat{p} \geq 1 - \theta \\ 0 & : \ \hat{p} \leq \theta \\ R & : \ 1 - \theta > \hat{p} > \theta \end{cases} \tag{2.3}$$

onde θ é um valor entre 0 e 0.5 que define os limiares da região de $\hat{p}$ em que a decisão possui risco excessivo, de acordo com o que é definido pelo modelador. A nova saída R é a rejeição do modelo que ocorre quando a probabilidade de classe cai em uma região de incerteza definada pelo limiar. Na literatura, a regra do *plug-in* é utilizada como um baseline de comparação para outros métodos de rejeição.

Nesse caminho, Geifman e El-Yaniv (2017) propõem um algoritmo para determinar o valor de θ ótimo dada uma taxa de tolerância a risco - ou erro - estabelecida pelo usuário. Outros estudos exploram características próprias dos algoritmos, como árvores de decisão (SHAKER; HÜLLERMEIER, 2020) e SVMs (GRANDVALET et al., 2008), para estabelecer um valor de confiabilidade para a predição, ou os modificam de modo a determinar uma função de rejeição durante o treinamento (FUMERA; ROLI, 2002; CORTES; DESALVO; MOHRI, 2016).

Por fim, uma abordagem pouco estudada é a utilização de outros modelos de AM para melhorar o processo de rejeição (ZHOU et al., 2022). Nesse sentido, Jiang et al. (2018) utilizou uma versão modificada do algoritmo de kNN para estimar a confiança sobre as predições de um modelo base. O valor gerado para isso, chamado *Trust Score*, é calculado como a razão entre a distância de uma amostra de teste ao conjunto de alta densidade da classe mais próxima (diferente da classe predita) e a distância ao conjunto de alta densidade da classe predita. Esse método se mostrou mais relevante na hora de selecionar as predições mais - e menos - confiáveis do que o valor de confiança dado pelo próprio estimador e foi escolhido como benchmark para comparação dos resultados obtidos pelo assessores no presente estudo.

A segunda condição é a novidade. Ela ocorre quando o exemplo a ser predito é muito diferente dos dados de treinamento (MARKOU; SINGH, 2003; PLAS et al., 2021), i,e., padrões mal representados nos dados de treinamento. Por exemplo, um modelo de detecção de fraudes

treinado com transações normais pode não estar preparado para identificar uma nova tática de fraude que nunca foi vista antes. Novamente, a opção de rejeição permite que o modelo recuse fazer uma previsão quando encontra exemplos novos e não familiares. Urahama e Furukawa (1995) modificam um k-NN para que o classificador se abstenha em casos em que a amostra esteja muito distante dos exemplos de treino. O mesmo pode ser feito para outros algoritmos baseados em protótipos, como o LVQ (CORDELLA et al., 1995).

2.2.2 Arquiteturas para sistemas com opção de rejeição

Ainda segundo (HENDRICKX et al., 2021), os modelos de AM com opção de rejeição podem possuir três tipos diferentes de arquiteturas:

- Arquitetura de rejeição separada: é uma abordagem na qual o modelo base e o rejeitador são treinados separadamente. O rejeitador atua de forma independente, filtrando as amostras sem que haja interação com as saídas resultantes do modelo classificador. Dentre as vantagens dessa arquitetura, pode-se destacar o fato de ser agnóstico quanto ao modelo base, sua simplicidade e, devido a sua função de filtro, ser capaz de diminuir o uso do modelo base, o que pode ser importante quando a inferência deste é cara computacionalmente;

- Arquitetura de rejeição dependente: nesse caso, o rejeitador depende das saídas fornecidas pelo modelo base. O rejeitador funciona como uma extensão do modelo base, geralmente através da produção de uma métrica de confiança associada a predição. Desse modo, o rejeitador pode atuar tanto em cima dos casos de novidade quanto de ambiguidade. Esse é o caso do sistema proposto na presente dissertação;

- Arquitetura de rejeição integrada: o rejeitador e o modelo base são indissociáveis, sendo treinados de maneira conjunta. Combina a previsão com a incerteza associada para tomar uma decisão final. Se a incerteza for alta, a previsão é rejeitada; caso contrário, ele faz a classificação normalmente (GEIFMAN; EL-YANIV, 2019).

Os mesmos autores diferenciam os tipos de sistemas de opção de rejeição de acordo com o período de treinamento dos rejeitadores. Eles usam o termo de Aprendizagem Sequencial se o rejeitador e o classificador são treinados separadamente, sendo o classificador treinado inicialmente. Essa abordagem tem como vantagem a flexibilidade, pois pode-se adicionar o

fator de rejeição a modelos já existentes ou reutilizar os rejeitadores ao adicioná-los a um processo existente, como o *plug-in*. Porém, esses sistema possui a desvantagem de que o apredizado só ocorrem uma direção, do classificador para o rejeitador. Assim, o classificador não consegue otimizar seu treinamento e seus resultados de acordo com o que seria rejeitado pelo rejeitador.

O outro tipo definido é o de Aprendizagem Simultânea, que adereça o problema citado acima. Nesse caso, os dois componentes são treinados em conjunto e vão se adaptando um ao outro. Contudo, a arquitetura desses sistemas tendem a ser mais complexas e caras computacionalmente.

2.2.3 Métricas para sistemas com opção de rejeição

A métrica mais utilizada para avaliar sistemas com opção de rejeição é a Curva de Acurácia-Rejeição, do inglês *Accuracy-Rejection Curve* (ARC) (NADEEM; ZUCKER; HANCZAR, 2009; GUAN et al., 2020; GIACINTO; ROLI; BRUZZONE, 2000; CONDESSA; BIOUCAS-DIAS; KOVAČEVIĆ, 2017; GEIFMAN; EL-YANIV, 2017; NETO et al., 2011). Na ARC, o eixo horizontal corresponde à taxa de rejeição, enquanto o eixo vertical representa a acurácia do modelo base, calculada para os exemplos de teste que não foram rejeitados. Para criar essa curva, são estabelecidos limiares progressivos de confiança, determinando a aceitação ou rejeição de uma predição. Para cada limiar, são calculadas a taxa de rejeição e a acurácia para os exemplos aceitos.

Quando a taxa de rejeição é zero, a acurácia obtida reflete o desempenho do modelo base ao ser avaliado em todos os exemplos de teste, servindo como um ponto de referência. Com o aumento da taxa de rejeição, espera-se uma melhoria na acurácia, já que o modelo é aplicado apenas aos exemplos de teste considerados menos incertos. A partir disso, é possível calcular a área sob a curva de acurácia-rejeição (AUARC), integrando os valores de acurácia ao longo das diferentes taxas de rejeição. Essa métrica consolidada fornece uma medida abrangente de desempenho do modelo, independente do valor específico da taxa de rejeição.

Outro meio de comparação entre sistemas de rejeição é atráves da análise gráfica das ARCs, de modo a identificar pontos em que uma curva está acima da outra, ou seja, taxas de rejeição para qual a acurácia de um sistema se torna maior que a de outro. Esse tipo de análise também é feito nos resultados do presente estudo.

2.2.4 AM com rejeição para predição de partidas de futebol

Poucos estudos exploram o conceito de rejeição no contexto do futebol ou do esporte em geral. Aqueles que aplicam algum tipo de filtragem de resultados geralmente utilizam adaptações da regra do *plug-in*. Partida et al. (2021) melhora a acurácia de um modelo para apostas de 50,65% para 60,74% ao rejeitar partidas cujo provável vencedor teve menos de 50% de chance de vitória definido pelo algoritmo. O mesmo teste é feito para os que tiveram menos de 70%, o que aumenta a acurácia do modelo para 70,00%. Já Stübinger, Mangold e Knoll (2019) segue o mesmo princípio, porém rejeita as partidas cuja diferença entre gols preditos para cada time seja maior que dois.

Entretanto, devido a limitações da regra do *plug-in*, ambos os casos dependem da qualidade do modelo base utilizado e não conseguem identificar instâncias em que o sistema foi exageradamente confiante. Esses são os casos em que a utilização de modelos de AM, como os assessores (HERNÁNDEZ-ORALLO; SCHELLAERT; MARTÍNEZ-PLUMED, 2022), são um diferencial no processo de rejeição de uma predição, como é proposto no presente estudo.

2.3 ASSESSORES

Segundo Hernández-Orallo, Schellaert e Martínez-Plumed (2022), um modelo assessor é um estimador de probabilidade condicional, denotado como $\hat{R}(r|\pi, \mu)$, onde r representa o resultado ou desempenho de um sistema de inteligência artificial em uma tarefa específica. O termo π refere-se ao perfil do sistema, que pode incluir características internas como arquitetura, hiperparâmetros ou o estado atual do sistema, enquanto μ representa a situação ou uma instância do problema.

O principal objetivo de um modelo assessor é prever a distribuição de probabilidade dos resultados r com base nos valores de π e μ. Desse modo, é possível antecipar o desempenho de um sistema para uma determinada instância ou conjunto de instâncias. Um exemplo de uso é uma situação em que há quatro modelos distintos disponíveis para predição de um problema. Um modelo assessor pode ser utilizado para antecipar qual dos estimadores π tem a maior chance de acertar sua previsão para aquela entrada μ e indicá-lo para uso.

O treinamento desses sistemas é realizado com base no conjunto de teste de um ou mais modelos base. Como atributos, utiliza-se a instância predita, juntamente com eventuais variáveis que adicionem contexto à situação. Já os rótulos correspondem a um indicador de

acerto da previsão, como 1 (correto) ou 0 (errado). A partir disso, espera-se que o assessor aprenda as relações entre o sistema, a instância e a sua probabilidade de acerto, identificando cenários em que o modelo base possui maior ou menor chance de realizar previsões corretas.

Os autores listam uma série de características que podem ser alcançadas ao se utilizar um modelo assessor, como:

- Antecipativa: o assessor prevê o desempenho de um sistema antes da implantação desse sistema;
- Autônoma: o assessor funciona independentemente do sistema original;
- Granular: o assessor prevê desempenho para cada instância de uma tarefa;
- Comportamental: o assessor pode aprender representações do comportamento emergente de um sistema.

Pensando no contexto da AM com opção de rejeição, fica clara a aptidão desses estimadores como rejeitadores em um sistema, podendo ser utilizados para aferir um escore de confiança para as predições de um modelo base. Como eles são treinados para aprender a relação entre uma instância e a chance de sucesso da predição do estimador base, o assessor pode encontrar regiões de dificuldade do modelo base que o próprio não identifica. Assim, enquanto o método plug-in, por exemplo, só consegue descartar as predições que o próprio modelo base aponta como incertos, o assessor pode ir além e rejeitar casos em que o modelo base foi excessivamente confiante. Zhou et al. (2022), por exemplo, utiliza um assessor para predizer o desempenho de modelos de linguagem dado um prompt de texto de entrada, rejeitando aquelas com baixa chance de sucesso e economizando os recursos necessários para a inferência de uma LLM que seriam desperdiçadas em uma má predição.

Além disso, por também ser um modelo de AM, pode-se aplicar técnicas de explicabilidade existentes para inspecionar as predições e entender quais são as suas maiores causas de incerteza. Essas informações podem ser utilizadas para conhecer e aprimorar o modelo base. Assim, o uso do assessor como opção de rejeição em um sistema é capaz de melhorar a sua confiabilidade ao rejeitar possíveis falhas e contribuir para a sua melhoria contínua.

3 ASSESSORES COM OPÇÃO DE REJEIÇÃO PARA FUTEBOL

A predição de resultados de partidas de futebol já se mostrou um desafio viável, porém complexo. Assim como outros contextos de aplicação, modelos de predição de resultados podem ter alta incerteza dependendo da instância a ser predita, no caso, da partida de futebol. No capítulo anterior, discutimos que alguns autores já usaram as probabilidades estimadas dos próprios preditores para tentar filtrar as partidas com maior grau de incerteza (PARTIDA et al., 2021) (STÜBINGER; MANGOLD; KNOLL, 2019). De uma forma geral, essa abordagem é limitada uma vez que modelos podem ter probabilidade de classe mal calibradas, ou com grau de confiança superestimado para certas instâncias.

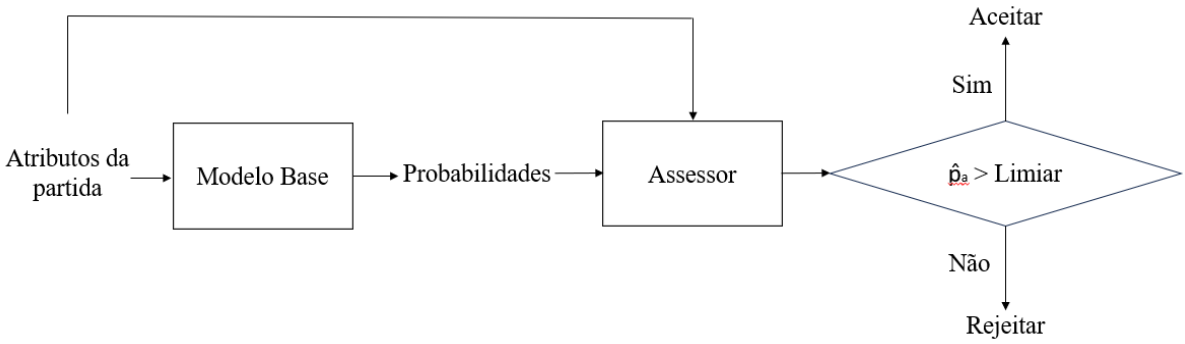
Assim, a presente dissertação investiga métodos alternativos usados como opção de rejeição no contexto de predição de resultados de partidas de futebol. Mais especificamente, a dissertação explora o uso dos modelos assessores (HERNÁNDEZ-ORALLO; SCHELLAERT; MARTÍNEZ-PLUMED, 2022), que já apresentaram resultados promissores como opção de rejeição em outros contextos (ZHOU et al., 2022).

3.1 ARQUITETURA GERAL PROPOSTA

A figura 1 representa a arquitetura geral da solução proposta. Dada uma partida, um modelo base de predição é usado para predizer probabilidades associadas aos resultados possíveis da partida, a partir das estatísticas 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. Portanto, o assessor proposto é uma adaptação do formulado por Hernández-Orallo, Schellaert e Martínez-Plumed (2022), pois recebe as probabilidades do modelo base como entrada, além das próprias características das partidas.

Formalmente, cada partida é descrita por um vetor de características $\vec{x} = (x_1, \ldots, x_p)$. 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, $\vec{p} = (\hat{p}_m, \hat{p}_e, \hat{p}_v)$, dentre vitória do mandante (m), empate (e) e vitória do visitante (v). A predição $\hat{y}$ do modelo base, a princípio, é o resultado possível da partida

Figura 1 – Sistema de predição proposto



Fonte: Elaborada pelo autor (2024)

com maior probabilidade estimada:

$$\hat{y} = arg \max_{y \in \{m,e,v\}} \hat{p}_y \tag{3.1}$$

O modelo base é então uma estimativa da função real $y = f_b(\vec{x})$, 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 abordagem escolhida para predição dos resultados das partidas. Nesta dissertação, o modelo base é um classificador multiclasses, que retorna probabilidades associadas a cada resultado possível da partida.

O modelo assessor, por sua vez, recebe como entrada o vetor de características da partida $\vec{x}$ e as probabilidades $\vec{p}$ 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 treinado 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:

$$\hat{p}_a = f_a(\vec{x}, \vec{p}) \tag{3.2}$$

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 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 $\hat{y}$ é rejeitada se o escore retornado pelo assessor $\hat{p}_a$ for menor que um dado limiar de aceitação θ.

$$\hat{f}(x) = \begin{cases} \hat{y} & : \hat{p}_a \geq \theta \\ R & : \hat{p}_a < \theta \end{cases} \tag{3.3}$$

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.

Observa-se que na solução proposta, o modelo assessor usado não é independente do modelo base e nem antecipativo, como em (HERNÁNDEZ-ORALLO; SCHELLAERT; MARTÍNEZ-PLUMED, 2022), uma vez que requer o uso das probabilidades de classe estimadas pelo modelo base. Porém, dessa maneira espera-se caracterizar suas características granulares e comportamentais, adicionando informação relativa a interação entre o modelo base e a instância.

Além disso, seguindo as classificações apresentadas por Hendrickx et al. (2021), a solução proposta possui as seguintes propriedades:

- Arquitetura de rejeição separada: recebe as probabilidades desse estimador como uma das suas variáveis de entrada;
- Aprendizagem sequencial: é treinado após o modelo base, pois assim é possível obter as saídas do estimador base do sistema.

O modelo base é construído a partir de dados históricos de resultados das partidas de um campeonato. 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. De modo geral, espera-se que o assessor aprenda regiões de dificuldade e de facilidade do modelo base e, no primeiro caso, auxiliar na rejeição de predições com baixa probabilidade de estarem corretas. Desse modo, o rejeitador atua no caso de identificação de ambiguidades.

3.2 ESTUDOS DE CASO

Nesta dissertação, a solução proposta de AM com opção de rejeição baseada em modelos assessores foi investigada em cinco campeonatos europeus: as ligas da Alemanha, Espanha, França, Inglaterra e Itália. Essas são as chamadas cinco grandes ligas europeias, que possuem maior poder financeiro, atenção do público e, consequentemente, mais dados disponíveis de maneira pública e confiável.

Iremos investigar a relação entre a qualidade dos modelos base e assessores. Observa-se aqui que uma das capacidades do modelo assessor é identificar situações em que o modelo base tem um desempenho ruim. O uso de um modelo assessor como opção de rejeição compensaria a baixa qualidade do modelo base em certas predições. Quando o modelo base tem uma alta qualidade para um dado problema no geral, é possível que o modelo assessor tenha menor utilidade. Assim, para o presente estudo, buscou-se avaliar o desempenho de diferentes modelos assessores e modelos bases, variando seus graus de complexidade, os dividindo entre modelos fracos, médios e fortes. A partir daí, analisou-se como os resultados variaram de acordo com a força dos modelos utilizados e em quais casos os assessores trouxeram maior ganho de performance para o sistema.

Na seção seguinte, são apresentadas as bases de dados usadas nos estudos de caso. Mais detalhes sobre a construção dos modelos base e dos assessores são apresentados nas Seções 3.2.2 e 3.2.3.

3.2.1 Base de dados

A base de dados disponibilizada por Pappalardo et al. (2019) foi a fonte utilizada na presente dissertação. Ela consiste em dados dos eventos de cada uma das partidas de cinco ligas europeias na temporada 2017/2018, da Copa do Mundo de 2018 e da Eurocopa de 2016. Para esse estudo, só foram consideradas as partidas dos campeonatos da Alemanha, Espanha, França, Inglaterra e Itália, totalizando 1826 jogos, distribuidos como mostra o quadro 1.

Nas seções seguintes, são apresentados os atributos descritores das partidas, usados pelos modelos como atributos preditores dos resultados. A descrição está organizada por tipo de atributo. Observa-se que cada partida é rotulada com uma das três classes equivalentes ao resultado de vitória do time visitante, empate ou vitória do time mandante.

Quadro 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**

Fonte: Elaborada pelo autor (2024)

3.2.1.1 Estatísticas das equipes

Cada evento corresponde a uma ação realizada dentro da partida, por exemplo: Atacante A passa para Atacante B no minuto X; Atacante B chuta para fora no minuto Y; e assim por diante. Cada liga possui um arquivo JSON identificando as partidas realizadas na temporada e um outro descrevendo cada um dos eventos realizados em suas partidas, de onde foram retiradas atributos preditores encontrada na base, obtendo-se valores como o número de chutes, passes, oportunidades, acelerações e cruzamentos para cada um dos times participantes. Uma amostra do arquivo de eventos é apresentada na Figura 2.

Após a coleta dos arquivos de eventos, o próximo passo consistiu em calcular os seguintes atributos descritores para cada uma das partidas:

- Calculo do desempenho médio geral - no campeonato todo, até aquela partida - de cada um dos times a partir de uma taxa de vitórias, empates, derrotas, gols marcados e sofridos como mandante e visitante, como descrito no Quadro 2;
- Calculo do desempenho médio, mas calculadas para os últimos 6 jogos de cada time para capturar o momento recente das equipes, como descrito no Quadro 3;
- Estatísticas médias de partida - número de passes, chutes, desarmes, dentre outros - para os últimos seis jogos do time, como descritos no Quadro 4 e Quadro 5.

Para garantir que os atributos preditores referentes ao momentos dos times sempre se referissem a seis partidas (HUCALJUK; RAKIPOVIĆ, 2011), descartou-se da base processada as seis primeiras rodadas de cada campeonato. Foi um escolhido um número par de rodadas

Figura 2 – Amostra de eventos da liga inglesa


```
[{'eventId': 8,
  'subEventName': 'Simple pass',
  'tags': [{'id': 1801}],
  'playerId': 25413,
  'positions': [{'y': 49, 'x': 49}, {'y': 78, 'x': 31}],
  'matchId': 2499719,
  'eventName': 'Pass',
  'teamId': 1609,
  'matchPeriod': '1H',
  'eventSec': 2.7586489999999912,
  'subEventId': 85,
  'id': 177959171},
 {'eventId': 8,
  'subEventName': 'High pass',
  'tags': [{'id': 1801}],
  'playerId': 370224,
  'positions': [{'y': 78, 'x': 31}, {'y': 75, 'x': 51}],
  'matchId': 2499719,
  'eventName': 'Pass',
  'teamId': 1609,
  'matchPeriod': '1H',
  'eventSec': 4.946850000000012,
  'subEventId': 83,
  'id': 177959172},
 {'eventId': 8,
  'subEventName': 'Head pass',
  'tags': [{'id': 1801}],
  'playerId': 3319,
  'positions': [{'y': 75, 'x': 51}, {'y': 71, 'x': 35}],
  'matchId': 2499719,
  'eventName': 'Pass',
  'teamId': 1609,
  'matchPeriod': '1H',
  'eventSec': 6.54218800000001,
  'subEventId': 82,
  'id': 177959173},
```


Fonte: Elaborada pelo autor (2024)

anteriores para que houvesse um equilíbrio entre partidas dentro e fora de casa. Portanto, a base de experimentos começa a partir da sétima rodada das ligas.

3.2.1.2 ELO

Em seguida, novamente para cada partida, coletou-se o *rating* ELO das equipes na semana anterior ao jogo através de consultas a uma API. Esse procedimento foi viável pois os valores da fonte dos dados do *rating* Elo, o ClubELO, são atualizados semanalmente e disponibilizados em http://www.clubelo.com/API. Assim, obteve-se um atributo quantificador da força relativa de cada equipe em relação às demais competidoras.

Quadro 2 – Estatísticas gerais

Coluna	Descrição
home_avg_team_score	Média geral de gols em casa do mandante
home_avg_opp_score	Média geral de gols sofridos em casa do mandante
home_avg_win	Média geral de vitórias em casa do mandante
home_avg_loss	Média geral de derrotas em casa do mandante
home_avg_draw	Média geral de empates em casa do mandante
away_avg_team_score	Média geral de gols fora de casa do mandante
away_avg_opp_score	Média geral de gols sofridos fora de casa do mandante
away_avg_win	Média geral de vitórias fora de casa do mandante
away_avg_loss	Média geral de derrotas fora de casa do mandante
away_avg_draw	Média geral de empates fora de casa do mandante
opp_home_avg_team_score	Média geral de gols em casa do time visitante
opp_home_avg_opp_score	Média geral de gols sofridos em casa do time visitante
opp_home_avg_win	Média geral de vitórias em casa do time visitante
opp_home_avg_loss	Média geral de derrotas em casa do time visitante
opp_home_avg_draw	Média geral de empates em casa do time visitante
opp_away_avg_team_score	Média geral de gols fora de casa do time visitante
opp_away_avg_opp_score	Média geral de gols sofridos fora de casa do time visitante
opp_away_avg_win	Média geral de vitórias fora de casa do time visitante
opp_away_avg_loss	Média geral de derrotas fora de casa do time visitante
opp_away_avg_draw	Média geral de empates fora de casa do time visitante

Fonte: Elaborado pelo autor

Quadro 3 – Estatísticas para os últimos seis jogos

Coluna	Descrição
form_win	Média de vitórias pelo mandante nos últimos seis jogos
form_draw	Média de empates pelo mandante nos últimos seis jogos
form_loss	Média de derrotas pelo mandante nos últimos seis jogos
form_team_score	Média de gols feitos pelo mandante nos últimos seis jogos
form_opp_score	Média de gols sofridos pelo mandante nos últimos seis jogos
opp_form_win	Média de vitórias pelo time visitante nos últimos seis jogos
opp_form_draw	Média de empates pelo time visitante nos últimos seis jogos
opp_form_loss	Média de derrotas pelo time visitante nos últimos seis jogos
opp_form_team_score	Média de gols feitos pelo time visitante nos últimos seis jogos
opp_form_opp_score	Média de gols sofridos pelo time visitante nos últimos seis jogos

Fonte: Elaborado pelo autor

Quadro 4 – Estatísticas de jogo para os últimos seis jogos do mandante

Coluna	**Descrição**
form_passes_forward	Média de passes para frente pelo mandante
form_passes_back	Média de passes para tráspelo mandante
form_passes_attack	Média de passes no terço de ataque pelo mandante
form_passes_mid	Média de passes no terço central do campo pelo mandante
form_passes_def	Média de passes no terço defensivo pelo mandante
form_passes_smart	Média de passes inteligentes pelo mandante
form_passes_key	Média de passes chave pelo mandante
form_tackle	Média de desarmes tentados pelo mandante
form_tackle_cplt	Média de desarmes bem sucedidos pelo mandante
form_tackles_attack	Média de desarmes no terço de ataque pelo mandante
form_tackles_mid	Média de desarmes no terço central time pelo mandante
form_tackles_def	Média de desarmes no terço defensivo pelo mandante
form_off_duels	Média de dribles tentados pelo mandante
form_off_duels_cplt	Média de dribles completados pelo mandante
form_opportunities	Média de oportunidades criadas pelo mandante
form_crosses	Média de cruzamentos tentados pelo mandante
form_crosses_cplt	Média de cruzamentos completos pelo mandante
form_shot_onTarget	Média de chutes no gol pelo mandante
form_accelerations	Média de acelerações pelo mandante
form_pct_pass_complete	Média de % de passes completos pelo mandante
form_passes_complete	Média de passes completos pelo mandante
form_passes	Média de passes tentados pelo mandante
form_duels	Média de duelos pelo mandante
form_shots	Média de chutes pelo mandante

Fonte: Elaborado pelo autor

3.2.1.3 Múltiplos de apostas

Como último passo, obteve-se os múltiplos de sete casas de apostas para cada final possível das partidas: vitória do mandante, empate ou vitória visitante. Os múltiplos se referem aos valores no momento de abertura, ou seja, são as estimativas iniciais das casas para aquela partida. Esses valores foram coletados do site *football-data.co.uk*, que agrega essas informações para diferentes campeonatos e temporadas e as disponibiliza para o público. Para a tabela final, foi incluída a média dos valores. Desse modo, espera-se incorporar o conhecimento imbutido nesses múltiplos pelos sistemas de apostas.

Quadro 5 – Estatísticas de jogo para os últimos seis jogos do visitante

Coluna	Descrição
opp_form_passes_forward	Média de passes para frente pelo time visitante
opp_form_passes_back	Média de passes para trás visitante
opp_form_passes_attack	Média de passes no terço de ataque pelo time visitante
opp_form_passes_mid	Média de passes no terço central do campo pelo time visitante
opp_form_passes_def	Média de passes no terço defensivo pelo time visitante
opp_form_passes_smart	Média de passes inteligentes pelo time visitante
opp_form_passes_key	Média de passes chave pelo time visitante
opp_form_tackle	Média de desarmes tentados pelo time visitante
opp_form_tackle_cplt	Média de desarmes bem sucedidos pelo time visitante
opp_form_tackles_attack	Média de desarmes no terço de ataque pelo time visitante
opp_form_tackles_mid	Média de desarmes no terço central time pelo visitante
opp_form_tackles_def	Média de desarmes no terço defensivo pelo time visitante
opp_form_off_duels	Média de dribles tentados pelo time visitante
opp_form_off_duels_cplt	Média de dribles completados pelo time visitante
opp_form_opportunities	Média de oportunidades criadas pelo time visitante
opp_form_crosses	Média de cruzamentos tentados pelo time visitante
opp_form_crosses_cplt	Média de cruzamentos completos pelo time visitante
opp_form_shot_onTarget	Média de chutes no gol pelo time visitante
opp_form_accelerations	Média de acelerações pelo time visitante
opp_form_pct_pass_complete	Média de % de passes completos pelo time visitante
opp_form_passes_complete	Média de passes completos pelo time visitante
opp_form_passes	Média de passes tentados pelo time visitante
opp_form_duels	Média de duelos pelo time visitante
opp_form_shots	Média de chutes pelo time visitante

Fonte: Elaborado pelo autor

3.2.1.4 Análise exploratória da base final

A base final obtida contém 1581 partidas, cada uma representada por uma linha da tabela, com 83 atributos preditores e não-nulos, um rótulo indicando o resultado final da partida, além de cinco colunas auxiliares com informações gerais da partida, mas não utilizadas no treinamento, como o nome dos times, rodada e e competição. As colunas da tabela estão consolidadas e descritas no Apêndice A.

O Quadro 6 mostra a distribuição dos resultados possíveis para as partidas. Há um ligeiro desbalanceamento a favor da vitória dos mandantes, como esperado dado a influência do fator

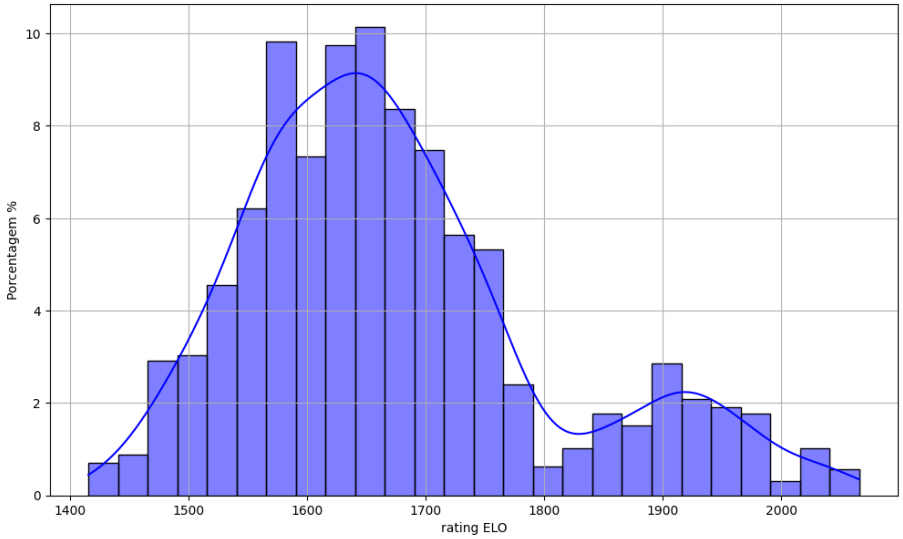
Quadro 6 – Distribuição de resultados das partidas de futebol.

Resultado	Porcentagem
Vitória do Mandante (2)	45,41%
Vitória do Visitante (0)	29,77%
Empate (1)	24,83%

Fonte: Elaborada pelo autor (2024)

casa nas partidas.

Figura 3 – Distribuição dos valores de *rating* Elo



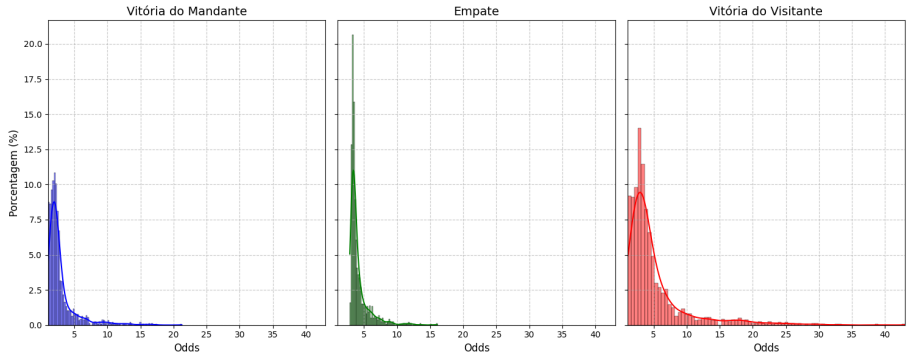
Fonte: Elaborada pelo autor (2024)

As Figuras 3, 4, 5 e 6 mostram os histogramas para os valores de diferentes colunas da base de dados. A Figura 3 indica que a maioria das equipes se concentram com Elos entre 1500 e 1700, mas que há uma outra concentração perto da marca de 1900. Esse segundo grupo agrega os chamados "superclubes", como o Real Madrid, Bayern de Munique, Juventus, PSG e Manchester City, que dominam o cenário europeu e suas respectivas ligas.

Já a Figura 4 mostra como estão distribuídos os múltiplos das casas de apostas. Os valores para vitórias de mandante ficam mais concentrados no início do eixo X, refletindo o favoritismo dos times da casa. Também percebe-se que os valores mais baixos da *odd* para empate começam a partir de 2,5, enquanto as de vitórias iniciam em valores mais baixos, próximos de 1. Isso indica que as casas não possuíram grande confiança em apontar empates em nenhum caso.

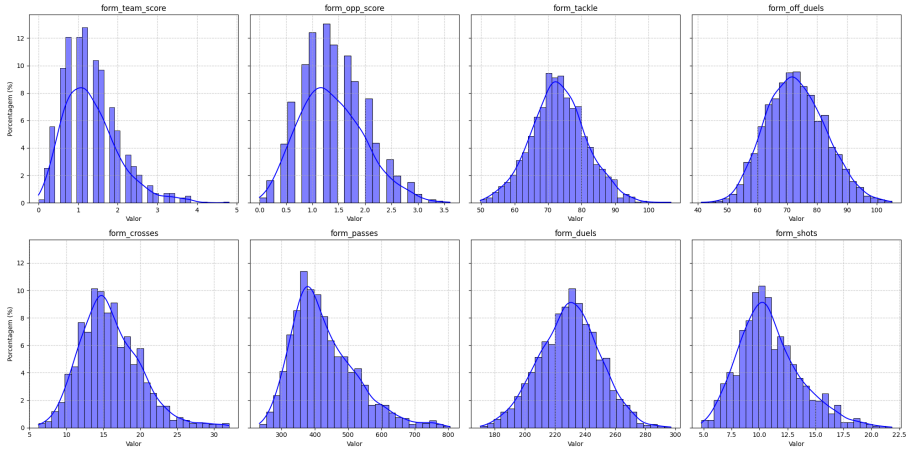
Por fim, as Figuras 5 e 6 mostram os histogramas de variáveis relacionadas a estatísticas de jogo e de resultados das equipes. Os gráficos relacionados às variáveis de número de vitórias

Figura 4 – Distribuição dos valores de múltiplos das casas de apostas



Fonte: Elaborada pelo autor (2024)

Figura 5 – Distribuição dos valores de estatísticas de jogo para os últimos seis jogos do mandante



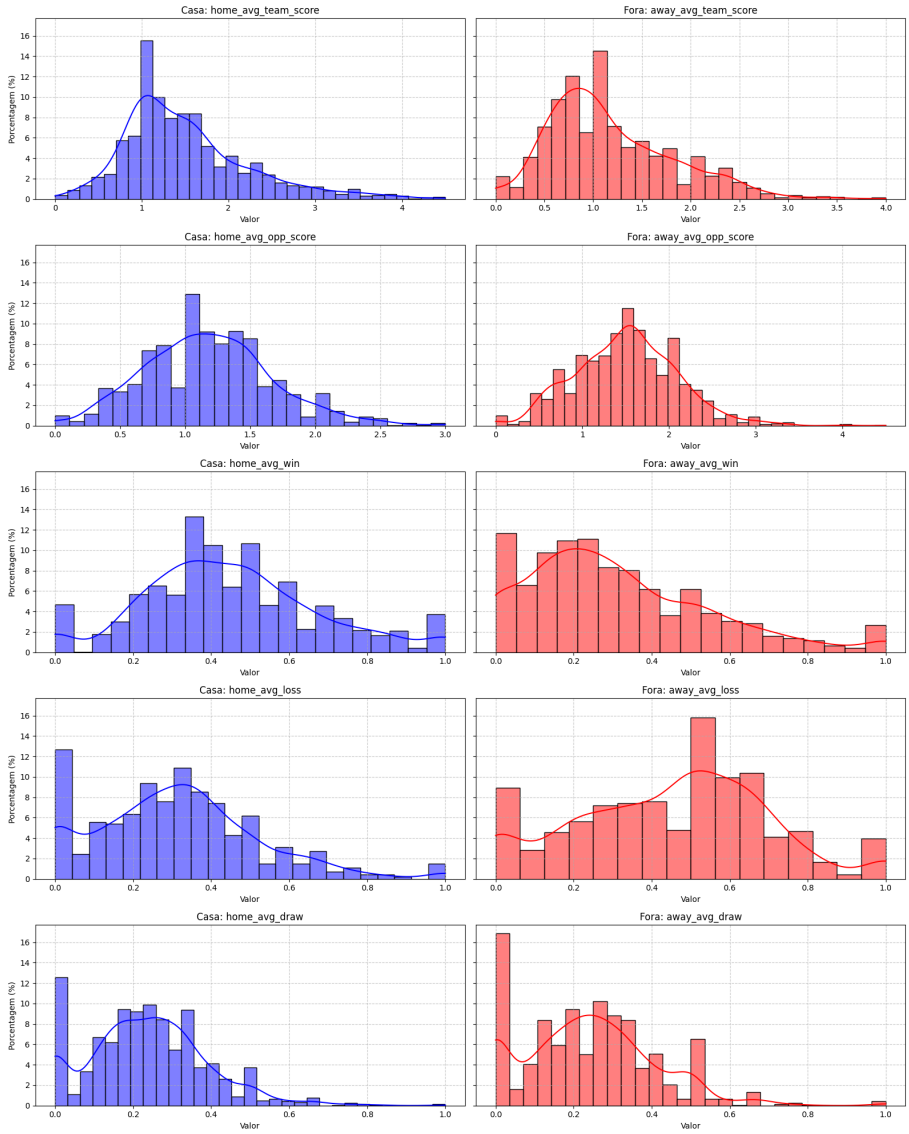
Fonte: Elaborada pelo autor (2024)

(*avg_win*) e de derrotas (*avg_loss*), na Figura 6 evidenciam a diferença de performance das equipes dependendo do fator casa, de modo que as equipes tendem a ter taxas de vitórias maiores quando jogam em casa.

3.2.2 Modelo Base

Os modelos foram treinados para cada rodada de cada liga. Em cada uma delas, utilizou-se como base de treino todas as partidas das rodadas anteriores do próprio torneio, mais todas as partidas de outras ligas ocorridas até o dia anterior à primeira partida da rodada a ser predita, de modo que o modelo receba o máximo possível de dados históricos disponíveis antes do período de teste. Escolheu-se usar partidas de outras ligas para aumentar o tamanho da base de treino, especialmente para as rodadas iniciais, quando o histórico de partida dentro dos próprios campeonatos ainda é baixo.

Figura 6 – Distribuição dos valores de estatísticas gerais do mandante



Fonte: Elaborada pelo autor (2024)

Feito isso, treinou-se o modelo com a base de treino construída para aquele momento e estimou-se as probabilidades de cada resultado da rodada atual, considerada de teste. Após cada repetição desse processo, ou seja, a cada rodada passada, armazenou-se as probabilidades obtidas para a rodada de teste em uma tabela que, ao fim, continha informações das partidas e as probabilidades geradas para cada uma delas. Essa tabela é a utilizada para a avaliação e análise dos resultados.

Cinco classificadores diferentes, de diferentes graus de complexidade, foram testados para o estudo de caso: a Regressão Logística com regularização L2, *Naive Bayes* Gaussiano, Perceptron Multicamadas (MLP) e *Random Forest*, da implementação do *Scikit-Learn* (PEDREGOSA et al., 2011), e o XGBoost, implementado pela biblioteca xgboost do Python. Todos eles foram

utilizados na literatura no contexto de predição de resultados para futebol e foram testados primeiramente na sua forma padrão das bibliotecas. No caso da Regressão Logística, *Naive Bayes* e MLP, os atributos preditores passaram por uma normalização pelo *z-score*.

3.2.3 Modelo assessor

Os modelos assessores foram treinados de modo similar aos modelos bases, com a diferença que recebiam como entrada adicional as probabilidades geradas por um modelo base e tinham como variável alvo se o modelo base havia acertado ou não a sua previsão. Como seu treinamento ocorre após o do modelo base, pode-se treinar diferentes assessores utilizando probabilidades geradas por modelos bases distintos.

Os algoritmos utilizados foram os mesmos que os testados para os modelos base.

4 EXPERIMENTOS

Neste capítulo, as Seções 4.1 e 4.2 explicam o processo de avaliação dos modelos. O processo de seleção de atributos é descrito na Seção 4.1.1 e o de otimização de hiperparâmetros em Seção 4.1.2. Por fim, os benchmarks de comparação são apresentados na Seção 4.3.

4.1 MODELO BASE

Os treinamentos começaram a partir da rodada 8, de modo que cada campeonato tivesse duas rodadas anteriores completas como base de treinamento, e todos os experimentos foram feitos com uma mesma semente de aleatoriedade. Além das predições feitas a cada rodada, também foram salvas as importâncias de variáveis após cada retreino, que foram utilizadas para análises.

Os experimentos foram avaliados através do cálculo da área abaixo da curva de acurácia e rejeição, explicado na Subseção 2.2.3. O primeiro passo para calculá-la foi de selecionar o resultado com maior probabilidade estimada pelo modelo para cada partida predita. Em seguida, as partidas foram ordenadas de maneira decrescente a partir do grau de confiança do estimador para o seu resultado final. Por fim, calculou-se as acurácias para diferentes taxas de rejeição, obtendo assim a curva ARC e, consequentemente, a área abaixo dela (AUARC).

Quadro 7 – Exemplos de predições

Partida	**Rodada**	$\hat{p}_m$	$\hat{p}_e$	$\hat{p}_v$	**Resultado**
Time A x Time B	8	0,1	0,3	**0,6**	Empate
Time C x Time D	8	**0,7**	0,15	0,15	Mandante
Time B x Time C	9	**0,5**	0,3	0,2	Mandante
Time C x Time A	10	**0,4**	0,35	0,25	Visitante
Time D x Time A	11	**0,9**	0,08	0,02	Mandante

Fonte: Elaborada pelo autor (2024)

Utilizando como exemplo o Quadro 7, obtém-se o Quadro 8 ao se executar os dois primeiros passos da avaliação. Já para os pontos da curva, a acurácia para uma taxa de rejeição de 0% é equivalente a acurácia geral do estimador, ou seja, de 60%. Já para uma taxa de rejeição de 20%, rejeita-se a predição de menor confiança, aumentando a acurácia do sistema para 75%. Ao subir a taxa de rejeição para 60%, mantendo apenas as duas predições de maior confiança

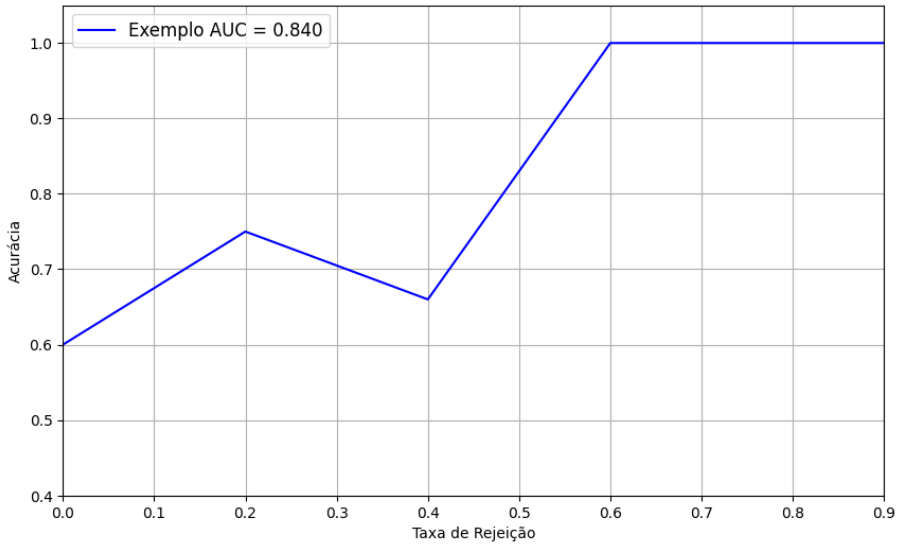
Quadro 8 – Exemplos de predições ordenadas pelo resultado mais provável

Partida	$max(\hat{p_y})$	**Acerto**
Time D x Time A	0,9	Sim
Time C x Time D	0,7	Sim
Time A x Time B	0,6	Não
Time B x Time C	0,5	Sim
Time C x Time A	0,4	Não

Fonte: Elaborada pelo autor (2024)

do modelo, a acurácia do sistema chega a 100%. A partir desses valores, pode-se montar a curva de acurácia-rejeição, representada na Figura 7 e a área abaixo dela, como no Quadro 9

Figura 7 – Curva de acurácia-rejeição (ARC)



Fonte: Elaborada pelo autor (2024)

Os cinco algoritmos especificados na Subseção 3.2.2 foram avaliados dessa maneira e, a partir dos seus resultados, mostrados na Seção 5.1, estabeleceu-se os três tipos de modelos base:

- Modelo base fraco: um algoritmo mais simples e com pior resultado, sem otimização de hiperparâmetros e seleção de atributos. No presente estudo, utilizou-se a Regressão Logística;
- Modelo base médio: algoritmo de melhor desempenho dentre os testados, porém sem otimização de hiperparâmetros e seleção de atributos. No presente estudo, utilizou-se o Random Forest;

Quadro 9 – Exemplo de cálculo de área abaixo da ARC

Taxa de rejeição (%)	**Acurácia (%)**
0	60
20	75
40	66
60	100
80	100
AUARC	**0.84**

Fonte: Elaborada pelo autor (2024)

- Modelo base forte: mesmo algoritmo que o modelo base médio, com hiperparâmetros otimizados e seleção de atributos.

As técnicas utilizadas para obter o modelo base forte são descritos nas subseções seguintes e foram aplicadas em conjunto com o o processo de treinamento descrito em 4.1 utilizando os conjuntos de treino e teste.

4.1.1 Seleção de atributos

Foi utilizado o algoritmo de *Backward Feature Elimination* (THARMAKULASINGAM et al., 2020), em que realizou-se o ciclo de treinamento da versão forte do modelo base e do assessor, repetidas vezes, sempre com a remoção de um atributo. Se essa remoção resultasse em um aumento do valor de AUARC, esse atributo era retirado definitivamente e continuava-se a repetição. Já caso a métrica não aumentasse, o atributo era mantido no estimador. Isso foi feito até que restassem apenas os atributos cuja remoção não impactaram negativamente na desempenho do modelo. Desse modo, buscou-se a combinação de atributos que maximiza o valor da métrica utilizada.

4.1.2 Otimização de hiperparâmetros

A otimização de hiperparâmetros do assessor e modelo base fortes foi feita através da técnica de Otimização Bayesiana (MARTINEZ-CANTIN, 2014) para obter os melhores valores de *n_estimators*, *min_samples_split*, *min_samples_leaf*, *max_samples* e *ccp_alpha*.

A otimização bayesiana de hiperparâmetros é uma técnica usada para encontrar os melhores

conjuntos de hiperparâmetros para um modelo de aprendizado de máquina. A abordagem bayesiana é usada porque ela fornece uma maneira eficiente de encontrar a configuração ótima de hiperparâmetros, mesmo quando o espaço de busca é grande e complexo. Em vez de tentar todas as combinações possíveis de hiperparâmetros, a otimização bayesiana tenta modelar a função de desempenho (ou função objetivo) e, com base nesse modelo, decide quais conjuntos de hiperparâmetros explorar a seguir.

No caso estudado, a função de desempenho buscou maximizar a métrica AUARC para avaliar os algoritmos utilizados. Para isso, fez-se o processo de treinamento completo do modelo e buscou-se maximizar o AUARC avaliando todas as rodadas preditas. O processo de otimização decide qual conjunto de valores dos hiperparâmetros explorar a seguir com base nas informações fornecidas pelo modelo probabilístico.

Ao longo de várias iterações, a otimização bayesiana ajusta continuamente seu modelo probabilístico com base nos resultados das avaliações dos hiperparâmetros. Isso permite que ela se adapte e concentre-se nas regiões mais promissoras do espaço de hiperparâmetros, eventualmente convergindo para a configuração que otimiza o desempenho do modelo.

As faixas de valores testadas são indicadas na Tabela 10.

Quadro 10 – Faixas de valores de hiperparâmetros

Hiperparâmetro	**Valor mínimo**	**Valor máximo**
max_samples	0,2	1
n_estimators	200	1000
min_samples_leaf	1	20
min_samples_split	2	20
ccp_alpha	0	0,04

Fonte: Elaborada pelo autor (2024)

A otimização foi realizada em etapas. Primeiro, testou-se valores para o *max_samples* e para o *n_estimators*. Em seguida, repetiu-se o processo para o *min_samples_leaf* e *min_samples_split* e, por último, para o *ccp_alpha*. Os resultados estão nas Tabelas 12 e 13

4.1.3 Importância dos atributos

A importância dos atributos para o modelo são obtidos através do método "*feature_importances_*", do *Random Forest Classifier* do scikit-learn. Ele retorna a média de diminuição de impureza, do inglês *Mean Decrease Impurity* (MDI), que mede a quantidade média pela qual a impureza

(geralmente a entropia ou o índice de Gini) é reduzida por cada variável ao longo de todas as árvores na floresta.

Os valores fornecidos pelos modelos foram salvos a cada rodada e, depois, calculou-se a importância média dada para cada atributo para o período analisado.

4.2 ASSESSOR

Nesse caso, os treinamentos começaram a partir da rodada 10, seguindo a mesma lógica de ter pelo menos duas rodadas completas disponíveis como base de treinamento.

A avaliação seguiu o mesmo processo, com a diferença de que foi utilizado o escore de confiança do assessor no momento de ordenar as partidas.

Aqui, também são testados três tipos diferentes de assessores, de maneira similar aos modelos base e os algoritmos são os mesmos escolhidos para o modelo base:

- Assessor fraco: regressão logística, o mesmo algoritmo escolhido para o modelo base fraco;
- Assessor médio: *Random Forest*, o mesmo algoritmo escolhido para modelo base médio, também sem otimização de hiperparâmetros e seleção de atributos;
- Assessor forte: mesmo algoritmo que o assessor médio, porém com hiperparâmetros afinados e com a melhor combinação de atributos.

Cada um desses assessores pode utilizar como insumo as predições de qualquer dos três tipos de modelo base definidos, dada a arquitetura sequencial e separada escolhida. Assim, é feito um experimento avaliando o assessor médio treinado com as predições de um modelo base de mesma força. Os resultados desse experimento são explorados na Seção 5.2, junto a uma análise por liga e por rodadas.

O passo seguinte foi de avaliar como sua performance muda ao trocar o modelo base anterior por um de força fraca e outro de força forte. Esses resultados e as análises estão na Seção 5.3. Feito isso, repetiu-se os passos anteriores com os assessores fortes e fracos, possibilitando uma análise extensa sobre as diferentes interações possíveis entre os modelos bases e os assessores e como suas forças alteram os resultados obtidos, que estão na Seção 5.4. Por fim, escolheu-se os experimentos entre o assessor forte e o modelo base forte para analisar as variáveis mais importantes para o estudo de caso, mostrados na Seção 5.4.2.

4.3 BASELINES DE COMPARAÇÃO

Os *baselines* escolhidos como base de comparação para a eficácia dos assessores foram:

- Modelo base (*plug-in*): através das próprias probabilidades geradas por ele, utilizando a regra do *plug-in* (Equação 2.3) para tomada de decisão;
- *Trust Score*: modelo proposto por Jiang et al. (2018) e explicado na Seção 2.2. Foi utilizada a classe implementada pelos autores, sem alteração de hiperparâmetros.
- Taxa de vitórias do mandante: a previsão para as partidas foi de vitória do mandante com probabilidade de ocorrência igual a taxa (%) de vitórias do mandante ao jogar em casa.

5 RESULTADOS

Neste capítulo, apresentamos os resultados obtidos com os modelos base e assessores. Inicialmente, detalhamos os experimentos de definição e treinamento dos modelos base (Seção 5.1). Em seguida, exploramos o desempenho do assessor médio com diferentes forças de modelos base, incluindo análises por liga e rodadas (Seções 5.2 e 5.3). Posteriormente, avaliamos as interações entre assessores e modelos base de forças distintas (Seção 5.4). Por fim, destacamos as variáveis mais relevantes em experimentos com o assessor forte e modelo base forte (Seção 5.4.2).

5.1 AVALIAÇÃO DOS MODELOS BASES

Esta seção tem dois objetivos. O primeiro é identificar qual é o melhor algoritmo para uso no modelo base. A partir daí, busca-se selecionar os algoritmos usados para geração dos modelos fortes, médios e fracos.

Os resultados obtidos pelos cinco algoritmos candidatos são sumarizados pela Tabela 11 e pelas curvas da Figura 8. Assim, o classificador com melhor AUARC foi o *Random Forest*. Além disso, esse algoritmo obteve uma curva dominante com vantagem até uma taxa de rejeição de 70%, quando foi ultrapassada pelo algoritmo de *Naive Bayes*. Destaca-se o resultado positivo do *Naive Bayes* para identificar as predições de altíssima confiança, mesmo que não se possa verificar as condições de independência entre os atributos, assim como ocorreu em (BABOOTA; KAUR, 2019). Esse algoritmo, foi porém muito fraco para taxas de rejeição mais baixas. A Regressão Logística, por sua vez, foi um algoritmo que obteve uma curva mais estável (porém sub-ótima) ao longo das taxas de rejeição, tendo tido o melhor AUARC dentre os algoritmos não-baseados em ensembles.

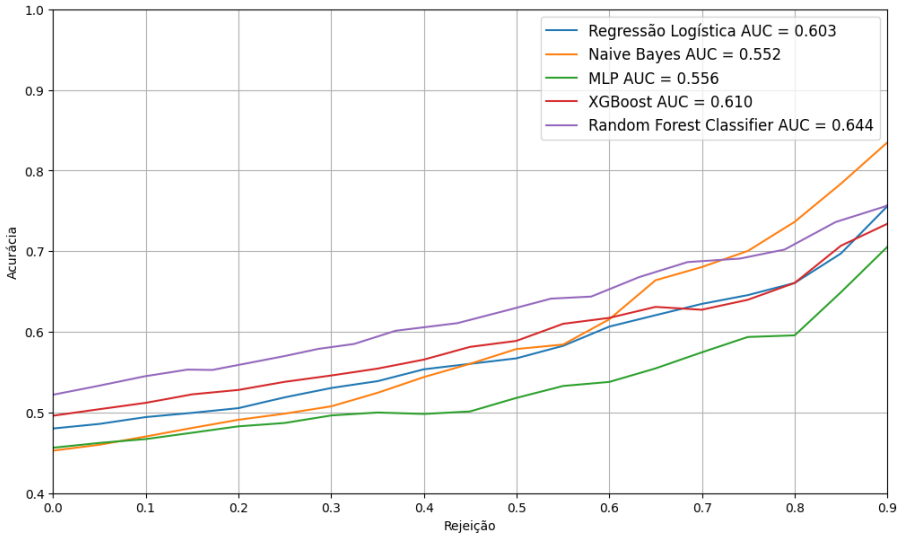
Conforme a discussão apresentada, decidiu-se seguir o restante das experimentações, utilizando o algoritmo *Random Forest* otimizado como modelo forte, o algoritmo *Random Forest* não otimizado como modelo médio e o algoritmo de Regressão Logística como modelo fraco. Os resultados dos modelos bases são apresentados na Figura 9. O base forte foi o que obteve o melhor valor de AUARC - 0,666 - para sua curva, que se manteve acima das outras curvas para todos valores de taxa de rejeição. Em seguida, como esperado, veio o base médio com valores próximos do forte e, por último, o base fraco, que obteve um AUARC de 0,603.

Quadro 11 – Resultados dos algoritmos testados para o Modelo Base

Algoritmo	**AUARC**
Regressão Logística	0,603
Naive Bayes	0,552
MLP	0,556
Xgboost	0,610
Random Forest	0,644

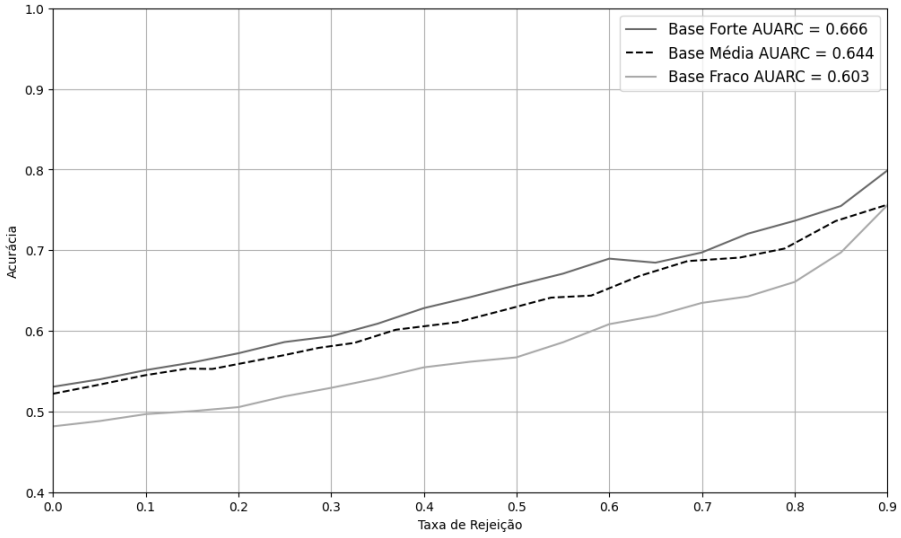
Fonte: Elaborada pelo autor (2024)

Figura 8 – Algoritmos testados para o modelo base



Fonte: Elaborada pelo autor (2024)

Figura 9 – Modelos bases

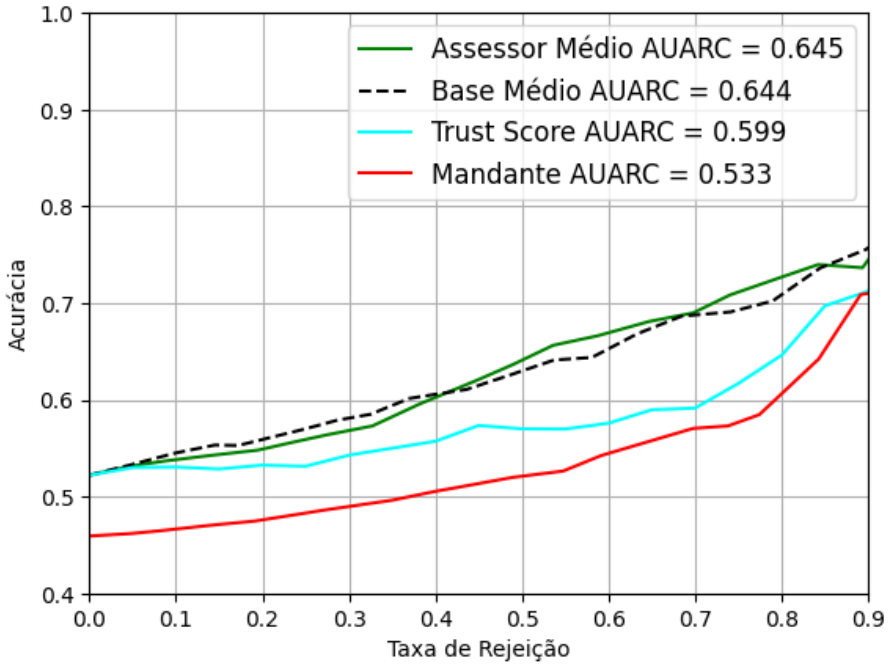


Fonte: Elaborada pelo autor (2024)

5.2 AVALIAÇÃO DO ASSESSOR MÉDIO

Esta seção busca analisar a qualidade do modelo assessor não-otimizado. Nesse caso, foram considerados os resultados obtidos pelo algoritmo *Random Forest* usado como modelo assessor médio, i.e., sem nenhuma otimização de parâmetros ou seleção de atributos. Inicialmente, apresenta-se os resultados usando tambem o modelo base médio. A partir da figura 10, nota-se que o assessor e o modelo base (*plug-in*) obtiveram desempenho similar em termos de AUARC, mas as curvas tiveram leves diferenças de acordo com a taxa de rejeição escolhida. A curva do modelo base estava acima da do assessor até chegar na taxa de rejeição de 40%, quando ocorre uma reversão que segue até em torno da taxa de 85%.

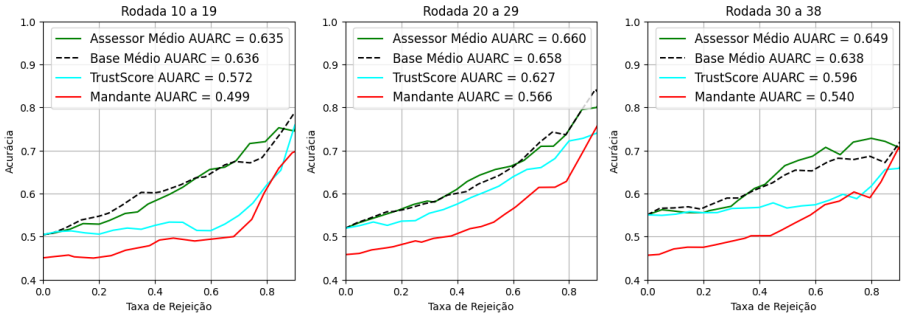
Figura 10 – Assessor Médio x Base Médio x *Trust Score*



Fonte: Elaborada pelo autor (2024)

Já a curva do método de *Trust Score* ficou abaixo das curvas dos dois métodos citados anteriormente em todas as faixas de rejeição, incluindo um período de estagnação entre as taxas de 45% e 70% de rejeição. Entretanto, todos modelos tiveram AUARC maior que o *baseline* mais básico, que sempre escolheu a vitória do mandante.

Figura 11 – Comparação por rodadas



Fonte: Elaborada pelo autor (2024)

5.2.1 Análise por rodadas

A análise por rodadas mostra uma tendência interessante, em que ocorre melhoras nas métricas das três curvas entre o primeiro e o segundo período analisados, mas há uma queda quando se analisa as partidas finais da base. A queda do modelo assessor é menor que do modelo base e a do *Trust Score*, mas observa-se que os valores de acurácia para as taxas de rejeição maiores sofrem uma estagnação.

Tal comportamento reflete o aumento de imprevisibilidade dos campeonatos nas retas finais, visto que equipes começam a ter motivações diferentes dependendo da sua posição na tabela, ocasionando que, por exemplo, times antes com baixo aproveitamento comecem a desempenhar melhor para fugir do rebaixamento e times de desempenho médio, que não brigam por nada ao fim da temporada, relaxem nos jogos finais.

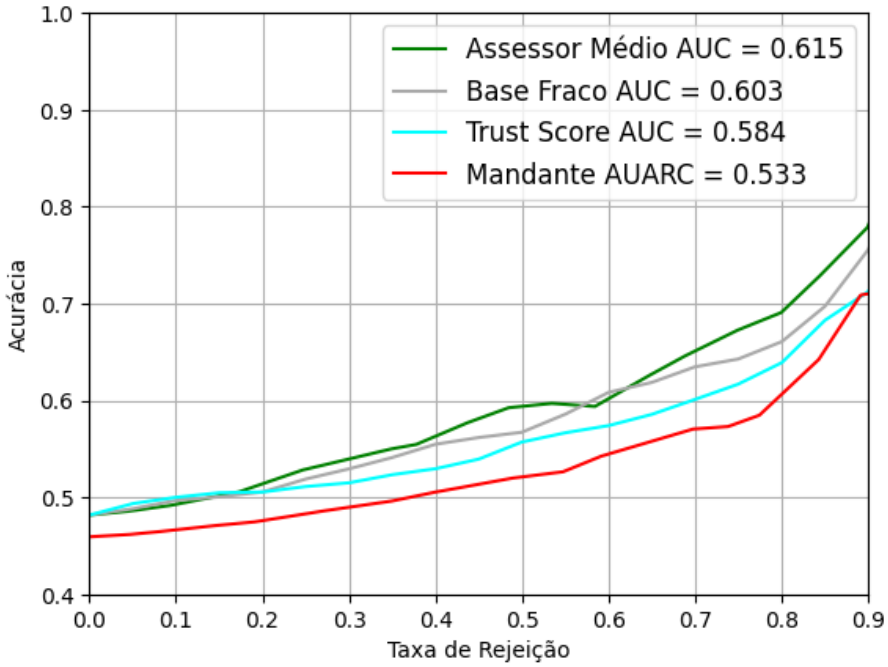
5.3 AVALIAÇÃO DO ASSESSOR MÉDIO COM DIFERENTES TIPOS DE MODELOS BASE

Nesta seção, expande-se as análises para os casos de uso do assessor médio junto aos modelos fraco e forte.

5.3.1 Base fraco

Nesse caso, percebe-se que a adição do assessor médio melhorou o desempenho do sistema, o que é esperado, visto que o modelo base fraco é um algoritmo de menor complexidade. A diferença do *Trust Score* pros modelos propostos é menor aqui, mas continua abaixo.

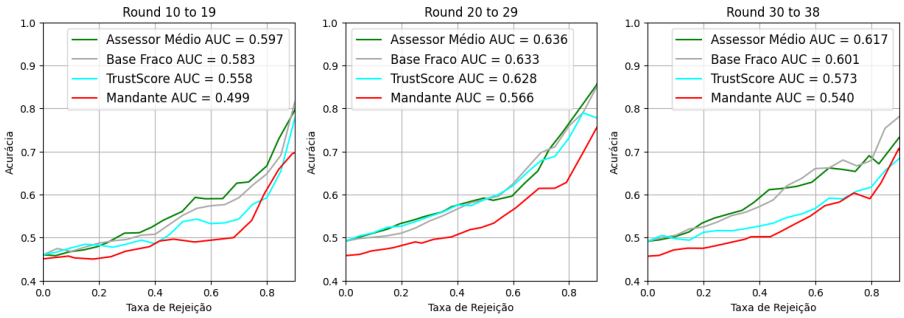
Figura 12 – Assessor médio x base fraco



Fonte: Elaborada pelo autor (2024)

5.3.1.1 Análise por rodadas

Figura 13 – Assessor médio x base fraco - Rodadas



Fonte: Elaborada pelo autor (2024)

A mesma tendência observada no modelo base médio é visto aqui, com o auge do desempenho ocorrendo na faixa central do campeonato, quando já há uma quantidade razoável de jogos para a base de treino e os times estão bem estabelecidos no campeonato.

Ao contrário das outras curvas, o *Trust Score* apresenta um AUARC maior quando aplicado sobre o modelo base fraco que sobre o modelo médio, para a faixa intermediária do campeonato.

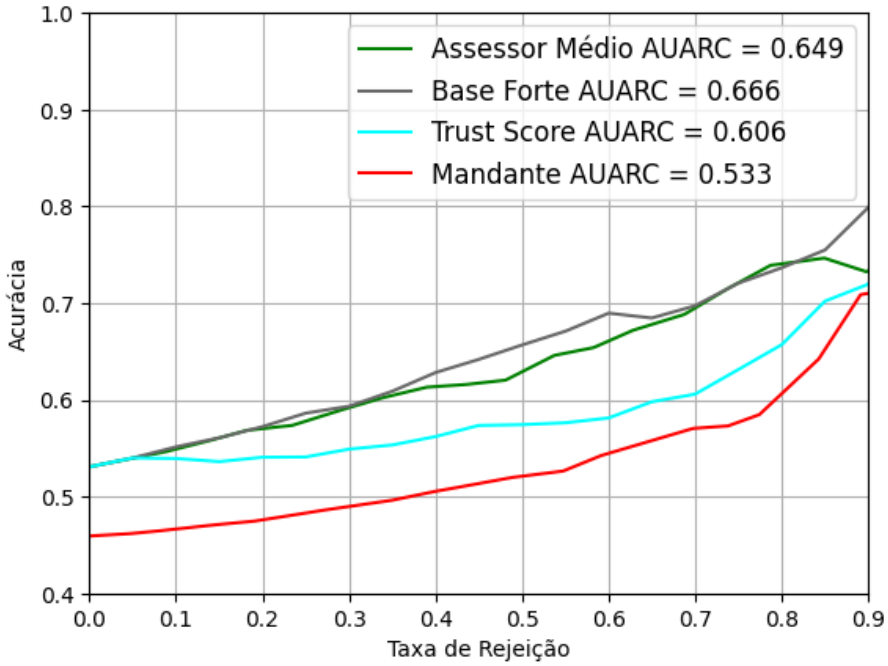
Quadro 12 – Hiperparâmetros do modelo base forte.

Hiperparâmetro	Valor
max_samples	0,237
n_estimators	710
min_samples_leaf	10
min_samples_node	10
ccp_alpha	0,01453

Fonte: Elaborada pelo autor (2024)

5.3.2 Base Forte

Figura 14 – Assessor médio x base forte



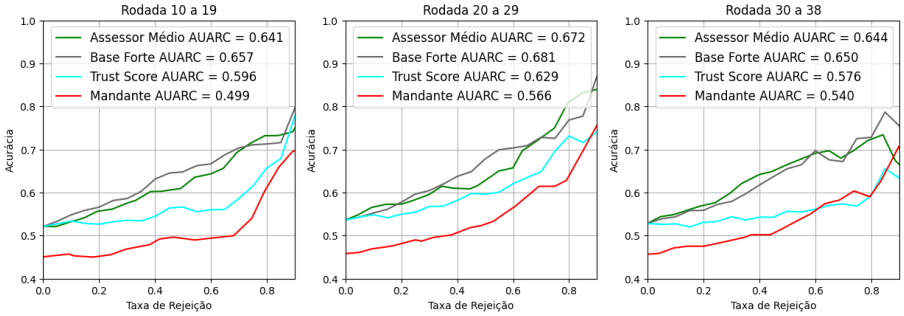
Fonte: Elaborada pelo autor (2024)

O modelo base forte é foi treinado sem as colunas "*form_team_score*", "*form_tackles_attack*", "*form_opportunities*", "*form_crosses*"e com os hiperparâmetros indicador no Quadro 12.

A Figura 14 mostra que o assessor médio não foi capaz de melhorar os resultados do modelo base forte, que apresentou um valor de AUARC maior e cuja curva ficou consistentemente acima do assessor médio e do *Trust Score*. Tal padrão indica que o benefício dos assessores diminui quando o modelo base é mais robusto.

5.3.2.1 *Análise por rodadas*

Figura 15 – Assessor médio x base forte - Rodadas



Fonte: Elaborada pelo autor (2024)

O base forte apresenta melhores resultados em todas as faixas do campeonato, embora a diferença diminua com o passar do tempo. Na última faixa de rodadas, a curva do assessor chega a ficar acima da do assessor até uma taxa de rejeição de 60%. Nesse período também chama atenção que o *Trust Score* fica abaixo da predição no mandante nas taxas de rejeição mais altas, evidenciando um desempenho fraco.

5.4 O CASO DO ASSESSOR FORTE E MODELO BASE FORTE

Esta seção mostra os resultados dos experimentos realizados para o assessor forte em comparação com o modelo base forte.

O assessor forte é foi treinado sem as colunas "*form_team_score*", "*form_passes_attack*", "*form_passes_smart*", "*form_passes_key*", "*opp_form_off_duels*", "opp_form_win"e com os hiperparâmetros indicados no Quadro 13.

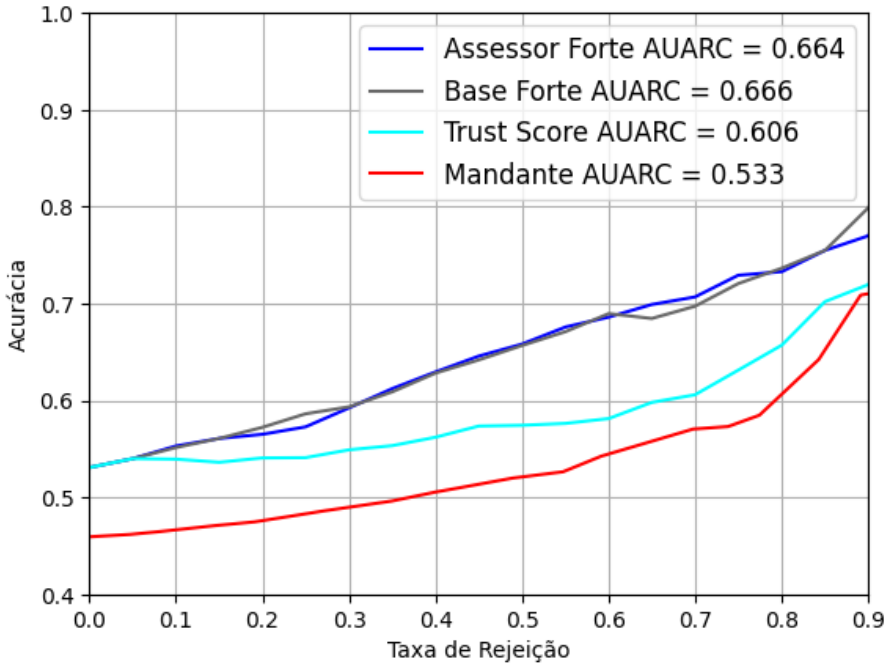
O assessor forte não conseguiu melhorar o desempenho geral do modelo base forte, entre-

Quadro 13 – Hiperparâmetros do assessor forte.

Hiperparâmetro	**Valor**
max_samples	0,4273
n_estimators	347
min_samples_leaf	12
min_samples_split	15
ccp_alpha	0,0

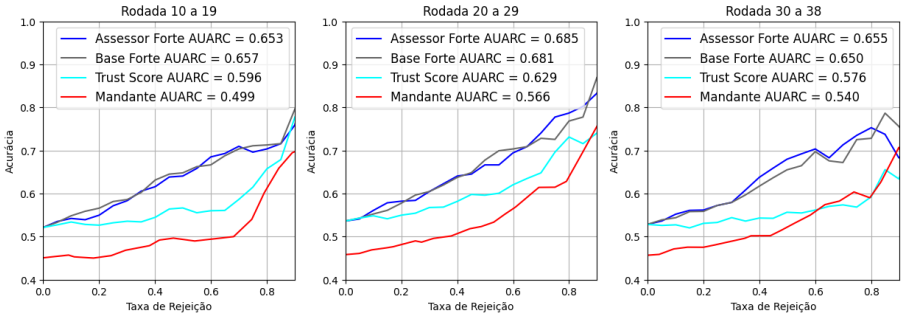
Fonte: Elaborada pelo autor (2024)

Figura 16 – Comparação entre assessor forte e modelo base forte



Fonte: Elaborada pelo autor (2024)

Figura 17 – Comparação entre o assessor forte e o base forte durante diferentes momentos dos campeonatos



Fonte: Elaborada pelo autor (2024)

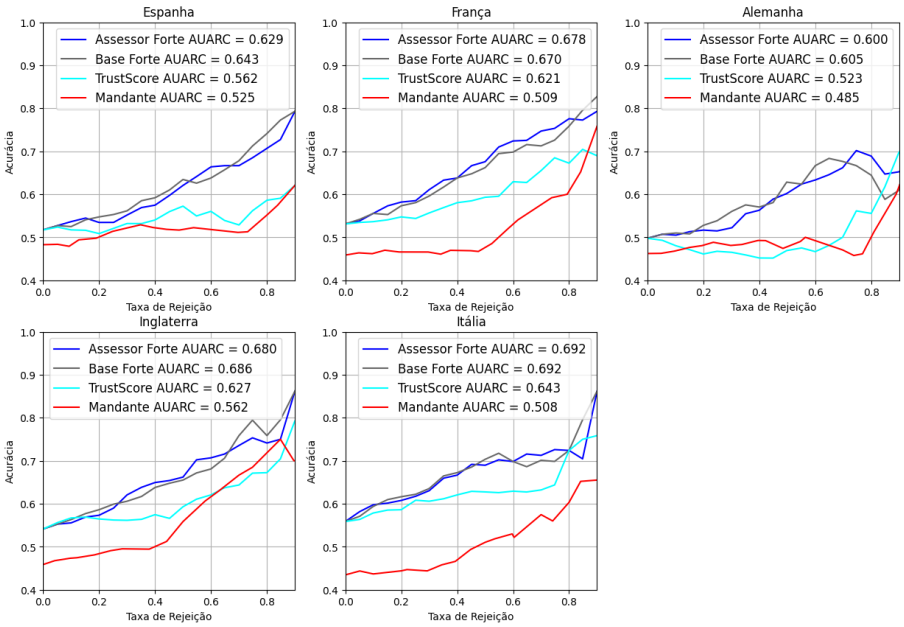
tanto, o fato de o modelo base possuir duas rodadas de vantagem sobre o assessor em uma base que já é pequena acaba influenciando na comparação geral entre os modelos. Assim, quando se analisa os gráficos da Figura 17, percebe-se que o assessor forte apresenta AUARC maior que o modelo base nas faixas de rodada mais avançadas, 20° até a 29° e 30° até a 38° rodadas, quando a vantagem inicial do modelo base devido às duas rodadas adicionais no seu treinamento se torna menos relevante. O mesmo comportamento é observado na Figura 10, indicando uma melhor capacidade do assessor de aprender com novos dados e de lidar com instâncias mais desafiadoras, como as do final do campeonato.

Além disso, o assessor proposto apresenta curvas melhores que o método de *Trust Score*,

que não conseguiu melhorar os resultados do modelo base em nenhuma ocasião.

5.4.1 Análise por campeonatos

Figura 18 – Comparação entre o assessor forte e o base forte em diferentes ligas



Fonte: Elaborada pelo autor (2024)

Na análise segregada por ligas, mostrado na Figura 18, o assessor só possui AUARC maior que o modelo base na liga francesa. Nesse caso, a curva do assessor se mantém acima até atingir uma taxa de rejeição de 80%. Nos outros países, há um empate na liga italiana e uma superioridade do modelo base nas ligas espanhola, alemã e inglesa, embora a análise gráfica das curvas mostre alguns pontos de rejeição em que o assessor possui maior acurácia.

Outro ponto que chama atenção é a diferença das métricas para cada liga. A liga alemã se mostrou a mais desafiadora para todos os casos analisados com uma grande margem de diferença, enquanto a italiana foi a mais previsível, seguida de perto pela inglesa.

5.4.2 Variáveis mais importantes para o modelo base e assessor fortes

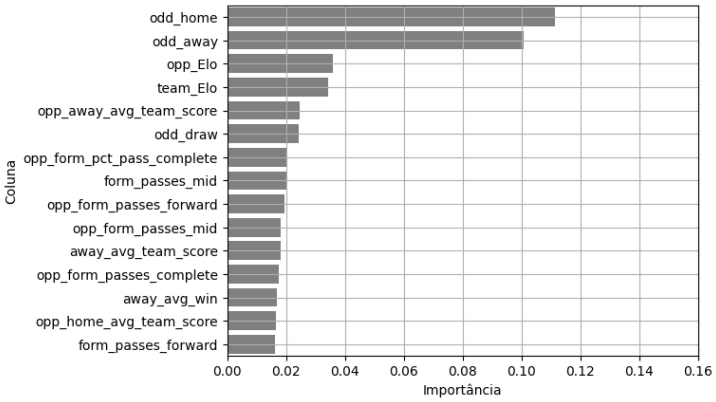
5.4.2.1 Atributos mais importantes para o modelo base forte

A Figura 19 mostra que as múltiplas relacionadas à vitória do mandante e do visitante fornecidas pelas casas de apostas foram, de longe, os atributos de maior importância para

o modelo. Também destaca-se os atributos relacionados ao *rating* Elo, que foram o terceiro e quarto atributos mais relevantes, mostrando que a adição desses atributos foi importante para o treinamento do modelo. Já dentre as estatísticas de jogo, percebe-se que os atributos relacionados aos passes foram os únicos que apareceram entre os quinze principais.

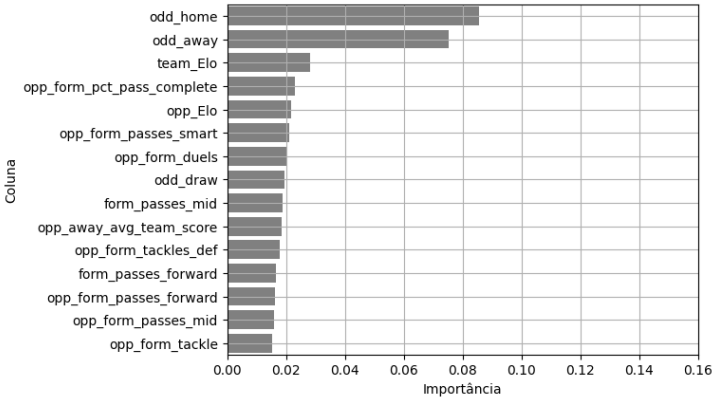
As Figuras 20, 21 e 22 mostram como os valores foram se alterando de acordo com o passar das rodadas. Nota-se que o modelo foi dando cada vez mais peso para os múltiplos das casas de aposta.

Figura 19 – Atributos com maior importância média para o modelo base forte



Fonte: Elaborada pelo autor (2024)

Figura 20 – Atributos com maior importância média para o modelo base forte entre as rodadas 10 e 19

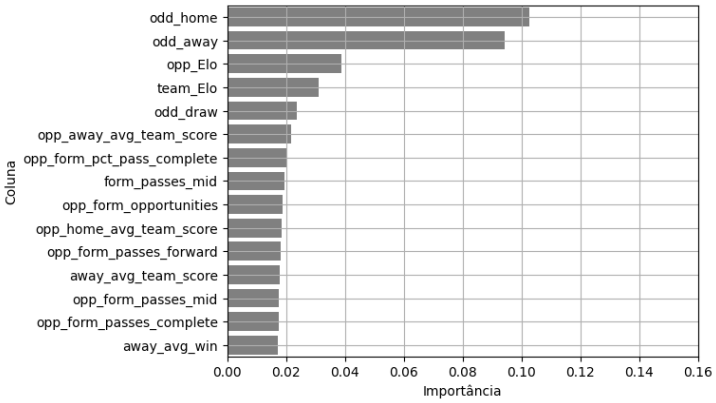


Fonte: Elaborada pelo autor (2024)

5.4.2.2 Atributos mais importantes para o assessor forte

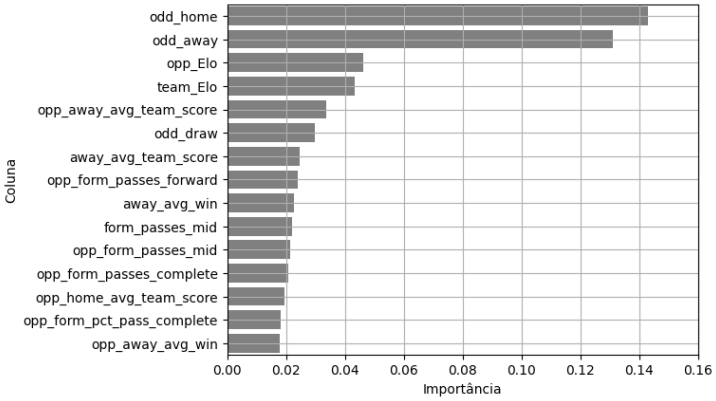
No caso do assessor, na Figura 23, a múltipla para empate surge como o atributo mais importante para o modelo. Também surgem as variáveis "*home*", "*away*"e "*draw*", que são

Figura 21 – Atributos com maior importância média para o modelo base forte entre as rodadas 20 e 29



Fonte: Elaborada pelo autor (2024)

Figura 22 – Atributos com maior importância média para o modelo base forte entre as rodadas 30 e 38



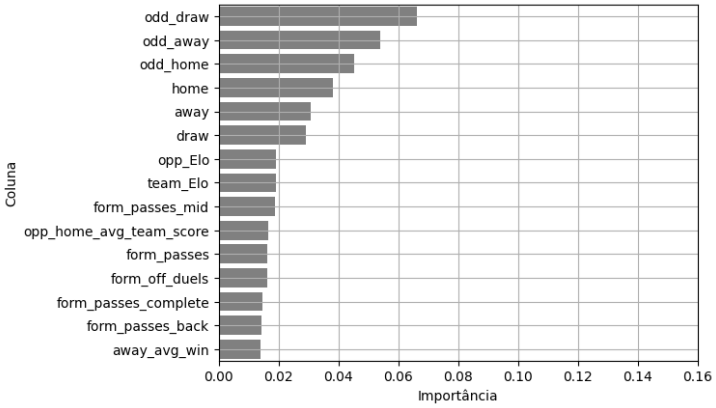
Fonte: Elaborada pelo autor (2024)

as probabilidades geradas pelo modelo base. Através das Figuras 24, 25 e 26 que a coluna "*odd_draw*"foi consistentemente a de maior relevância para o assessor.

De fato, o modelo base forte teve grande dificuldade em identificar empates. Embora esse tenha sido o resultado de 341 das partidas da base, o modelo só o indicou em duas ocasiões, errando ambas. Portanto, o assessor identificou essa fraqueza do modelo base e que a coluna "*odd_draw*"indicava uma maior chance desse tipo de resultado ocorrer.

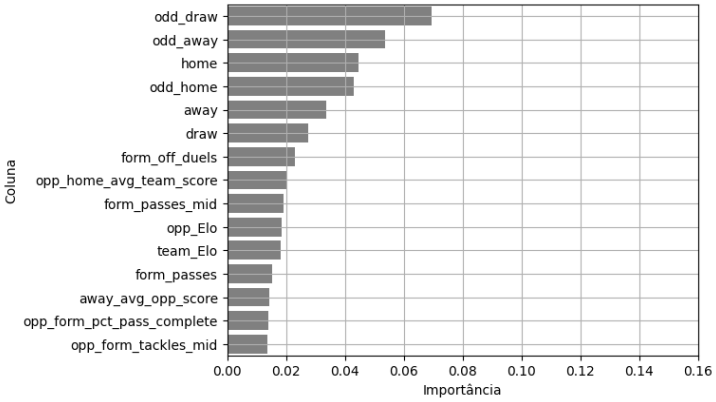
Também chama atenção que uma variável relacionada ao número de passes do time da casa chega a ter mais relevância que a variável ELO do time visitante no último período analisado.

Figura 23 – Atributos com maior importância média para o assessor forte



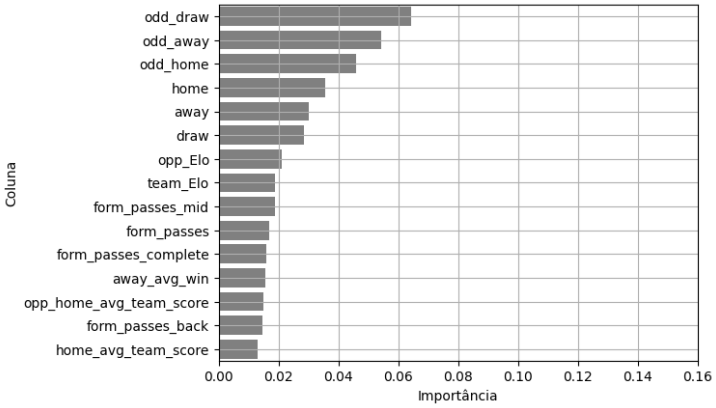
Fonte: Elaborada pelo autor (2024)

Figura 24 – Atributos com maior importância média para o assessor forte entre as rodadas 10 e 19



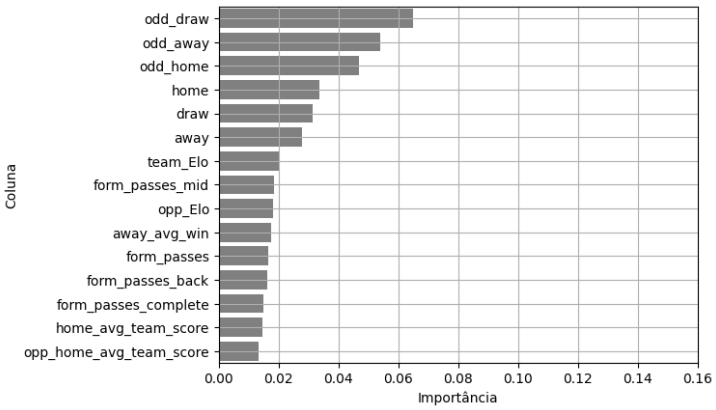
Fonte: Elaborada pelo autor (2024)

Figura 25 – Atributos com maior importância média para o assessor forte entre as rodadas 20 e 29



Fonte: Elaborada pelo autor (2024)

Figura 26 – Atributos com maior importância média para o assessor forte entre as rodadas 30 e 38



Fonte: Elaborada pelo autor (2024)

6 CONCLUSÕES

Esta dissertação investigou a utilização de modelos assessores como opção de rejeição no contexto de predições de partidas de futebol, um domínio caracterizado por grande imprevisibilidade e incerteza. Nessa proposta, modelos assessores são treinados a partir da avaliação de um modelo preditor de partidas ao longo das rodadas de um torneio. O modelo assessor é capaz então de antecipar nas rodas seguintes a incerteza das predições do modelo base, descartando aquelas predições menos confiáveis.

6.1 RESULTADOS PRINCIPAIS

Os experimentos realizados sobre dados reais de diferentes campeonatos europeus permitiram explorar variados aspectos desta proposta, levando a conclusões relevantes. Primeiramente, verificou-se que o assessor desenvolvido é capaz de filtrar situações de baixa confiança do modelo base, melhorando sua acurácia global ao melhorar a capacidade de rejeição de predições duvidosas, além de superar um algoritmo da literatura e um método heurístico na mesma tarefa.

Apesar de terem performances globais semelhantes, o assessor médio o *Random Forest* padrão, tendeu a superar o modelo base médio com o passar dos campeonatos investigados. Isso indica que o assessor analisado aproveita melhor o aumento de disponibilidade dos dados. Outro fator que chama a atenção é que a maior diferença entre as abordagens ocorre na fase final do campeonato, a mais imprevisível, evidenciando o potencial da solução em casos em que a opção de rejeição é interessante. Entretanto, observou-se que o benefício do assessor é reduzido à medida que a qualidade do modelo base melhora. De fato, enquanto o assessor médio melhorou os resultados em relação ao base fraco, foi superado por uma grande margem pelo base forte. Porém, a distância entre os dois diminuiu à medida que o campeonato avançou, reforçando o potencial visto anteriormente.

Do mesmo modo, o assessor forte conseguiu melhorar os resultados em relação ao modelo base equivalente nos dois últimos períodos do campeonato. Esses resultados mostram que, a longo prazo e com maiores volumes de dados, o assessor é capaz de melhorar até mesmo um modelo base já bem ajustado.

Além disso, o estudo indicou que o desempenho dos modelos varia bastante de acordo com

a liga a qual está sendo aplicada. A liga alemã se mostrou a mais difícil em todos os contextos analisados, tanto para o assessor quanto para o modelo base. Já a liga italiana foi onde foram obtidos os melhores resultados, incluindo o maior AUARC, de 0,703, atingido pelo assessor forte nas suas últimas oito rodadas.

Contudo, a obtenção de bases de dados que sejam extensas e contenham um grande número de características para cada partida é um aspecto desafiador e foi um fator limitante para o presente estudo, visto que escolheu-se privilegiar um grande detalhamento das partidas em detrimento de mais partidas para análise.

Por fim, a análise dos atributos mais importantes para cada método mostrou que os múltiplos médios das casas de aposta tiveram grande peso nas decisões dos algoritmos. Outras variáveis consistentemente entre as principais foram as baseados nos *ratings* ELO das equipes participantes na partida. Também destaca-se que, no caso dos assessores, o múltiplo para o resultado de empate foi o atributo mais relevante, pois ele identificou que o modelo base teve grande dificuldade de predizer esse fim para as partidas. Essa análise corrobora a capacidade e o potencial dos assessores para identificar as dificuldades de um modelo base e ajustar sua predição com base nisso.

6.2 TRABALHOS FUTUROS

Para trabalhos futuros, pode-se explorar a mesma aplicação, mas com um volume maior de dados, contendo mais ligas e mais temporadas. Os dados utilizados são provenientes do *WyScout* através de Pappalardo et al. (2019), porém, pode-se explorar o repositório público do *StatsBomb*, empresa concorrente, que possui atualizações de temporadas mais recentes. Além disso, pode-se criar e testar novos atributos para representar as partidas e equipes, buscando características mais informativas para a predição dos resultados. Nesse caminho, um indicador bastante em evidência nas análises esportivas ultimamente é o de gols esperados (xG) (GREEN, 2012), que calcula a probabilidade de um chute resultar em gol e cuja sumatória indica a qualidade das chances criadas por uma equipe ou jogador em um determinado período. Também pode-se aplicar técnicas de seleção de atributos mais avançadas para identificar o conjunto de variáveis mais relevantes para o problema estudado, evitando dados com alta correlação entre si e com baixa capacidade preditiva. Isso pode ser feito através de técnicas como a eliminação recursiva de atributos (GUYON et al., 2002) e análises de correlação.

Outras características possíveis dos assessores podem ser exploradas, como a antecipativa,

ao treinar os assessores sem que esse receba as predições geradas pelo modelo base, desse modo, pode-se utilizar o assessor como um filtro anterior, assim como feito por Zhou et al. (2022). Por fim, também há a oportunidade de se analisar o uso dos assessores em outros esportes e, principalmente, em outros domínios onde erros do classificador tenham alto custo e seja interessante rejeitar predições de baixa confiabilidade, como em prevenção à fraudes, diagnósticos de doenças e predição de falhas em equipamentos mecânicos.

REFERÊNCIAS

ANGELINI, G.; CANDILA, V.; ANGELIS, L. D. Weighted elo rating for tennis match predictions. *European Journal of Operational Research*, Elsevier, v. 297, n. 1, p. 120–132, 2022.

BABOOTA, R.; KAUR, H. Predictive analysis and modelling football results using machine learning approach for english premier league. *International Journal of Forecasting*, Elsevier, v. 35, n. 2, p. 741–755, 2019.

BARTLETT, P. L.; WEGKAMP, M. H. Classification with a reject option using a hinge loss. *Journal of Machine Learning Research*, v. 9, n. 8, 2008.

BEAL, R.; NORMAN, T. J.; RAMCHURN, S. D. Artificial intelligence for team sports: a survey. *The Knowledge Engineering Review*, Cambridge University Press, v. 34, p. e28, 2019.

BERRAR, D.; LOPES, P.; DUBITZKY, W. Incorporating domain knowledge in machine learning for soccer outcome prediction. *Machine learning*, Springer, v. 108, p. 97–126, 2019.

BRIN, S.; PAGE, L. The anatomy of a large-scale hypertextual web search engine. *Computer networks and ISDN systems*, Elsevier, v. 30, n. 1-7, p. 107–117, 1998.

BROOKS, J.; KERR, M.; GUTTAG, J. Using machine learning to draw inferences from pass location data in soccer. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, Wiley Online Library, v. 9, n. 5, p. 338–349, 2016.

BRYANT, K. How ai is impacting society and shaping the future. 2023. Disponível em: <https://www.forbes.com/sites/kalinabryant/2023/12/13/how-ai-is-impacting-society-and-shaping-the-future/>. Acesso em: 15 dez. 2024.

BUNKER, R.; SUSNJAK, T. The application of machine learning techniques for predicting match results in team sport: A review. *Journal of Artificial Intelligence Research*, v. 73, p. 1285–1322, 2022.

CHOW, C. On optimum recognition error and reject tradeoff. *IEEE Transactions on information theory*, IEEE, v. 16, n. 1, p. 41–46, 1970.

CHOW, C.-K. An optimum character recognition system using decision functions. *IRE Transactions on Electronic Computers*, IEEE, n. 4, p. 247–254, 1957.

CONDESSA, F.; BIOUCAS-DIAS, J.; KOVAČEVIĆ, J. Performance measures for classification systems with rejection. *Pattern Recognition*, Elsevier, v. 63, p. 437–450, 2017.

CONSTANTINOU, A. C. Dolores: a model that predicts football match outcomes from all over the world. *Machine Learning*, Springer, v. 108, n. 1, p. 49–75, 2019.

CONSTANTINOU, A. C.; FENTON, N. E. Determining the level of ability of football teams by dynamic ratings based on the relative discrepancies in scores between adversaries. *Journal of Quantitative Analysis in Sports*, De Gruyter, v. 9, n. 1, p. 37–50, 2013.

CORDELLA, L. P.; STEFANO, C. D.; SANSONE, C.; VENTO, M. An adaptive reject option for lvq classifiers. In: SPRINGER. *Image Analysis and Processing: 8th International Conference, ICIAP'95 San Remo, Italy, September 13–15, 1995 Proceedings 8*. [S.l.], 1995. p. 68–73.

CORTES, C.; DESALVO, G.; MOHRI, M. Boosting with abstention. *Advances in Neural Information Processing Systems*, v. 29, 2016.

DELOITTE, U. *Deloitte Football Money League*. [S.l.]: Manchester, 2020.

DIXON, M. J.; COLES, S. G. Modelling association football scores and inefficiencies in the football betting market. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, Wiley Online Library, v. 46, n. 2, p. 265–280, 1997.

FIFA. Fifa world cup qatar 2022 commercial. 2022. Disponível em: <https://inside.fifa.com/fifa-world-cup-qatar-2022-commercial>. Acesso em: 13 ago. 2024.

FORREST, D.; GODDARD, J.; SIMMONS, R. Odds-setters as forecasters: The case of english football. *International journal of forecasting*, Elsevier, v. 21, n. 3, p. 551–564, 2005.

FUMERA, G.; ROLI, F. Support vector machines with embedded reject option. In: SPRINGER. *Pattern Recognition with Support Vector Machines: First International Workshop, SVM 2002 Niagara Falls, Canada, August 10, 2002 Proceedings*. [S.l.], 2002. p. 68–82.

GEIFMAN, Y.; EL-YANIV, R. Selective classification for deep neural networks. *Advances in neural information processing systems*, v. 30, 2017.

GEIFMAN, Y.; EL-YANIV, R. Selectivenet: A deep neural network with an integrated reject option. In: PMLR. *International conference on machine learning*. [S.l.], 2019. p. 2151–2159.

GIACINTO, G.; ROLI, F.; BRUZZONE, L. Combination of neural and statistical algorithms for supervised classification of remote-sensing images. *Pattern Recognition Letters*, Elsevier, v. 21, n. 5, p. 385–397, 2000.

GODDARD, J. Regression models for forecasting goals and match results in association football. *International Journal of forecasting*, Elsevier, v. 21, n. 2, p. 331–340, 2005.

GODIN, F.; ZUALLAERT, J.; VANDERSMISSEN, B.; NEVE, W. D.; WALLE, R. Van de. Beating the bookmakers: leveraging statistics and twitter microposts for predicting soccer results. In: ACM NEW YORK, NY, USA. *KDD Workshop on large-scale sports analytics*. [S.l.], 2014. p. 2–14.

GRANDVALET, Y.; RAKOTOMAMONJY, A.; KESHET, J.; CANU, S. Support vector machines with a reject option. *Advances in neural information processing systems*, v. 21, 2008.

GREEN, S. Assessing the performance of premier league goalscorers. 2012. Disponível em: <https://www.statsperform.com/resource/assessing-the-performance-of-premier-league-goalscorers/>. Acesso em: 15 dez. 2024.

GUAN, H.; ZHANG, Y.; CHENG, H.-D.; TANG, X. Bounded-abstaining classification for breast tumors in imbalanced ultrasound images. *International Journal of Applied Mathematics and Computer Science*, v. 30, n. 2, 2020.

GUYON, I.; WESTON, J.; BARNHILL, S.; VAPNIK, V. Gene selection for cancer classification using support vector machines. *Machine learning*, Springer, v. 46, p. 389–422, 2002.

HANSEN, L. K.; LIISBERG, C.; SALAMON, P. The error-reject tradeoff. *Open Systems & Information Dynamics*, Springer, v. 4, n. 2, p. 159–184, 1997.

HENDRICKX, K.; PERINI, L.; PLAS, D. Van der; MEERT, W.; DAVIS, J. Machine learning with a reject option: A survey. *arXiv preprint arXiv:2107.11277*, 2021.

HERBEI, R.; WEGKAMP, M. H. Classification with reject option. *The Canadian Journal of Statistics/La Revue Canadienne de Statistique*, JSTOR, p. 709–721, 2006.

HERNÁNDEZ-ORALLO, J.; SCHELLAERT, W.; MARTÍNEZ-PLUMED, F. Training on the test set: Mapping the system-problem space in ai. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. [S.l.: s.n.], 2022. v. 36, n. 11, p. 12256–12261.

HUBÁČEK, O.; ŠOUREK, G.; ŽELEZNỲ, F. Learning to predict soccer results from relational data with gradient boosted trees. *Machine Learning*, Springer, v. 108, p. 29–47, 2019.

HUCALJUK, J.; RAKIPOVIĆ, A. Predicting football scores using machine learning techniques. In: IEEE. *2011 Proceedings of the 34th International Convention MIPRO*. [S.l.], 2011. p. 1623–1627.

HVATTUM, L. M.; ARNTZEN, H. Using elo ratings for match result prediction in association football. *International Journal of forecasting*, Elsevier, v. 26, n. 3, p. 460–470, 2010.

IT, A. Inside the nascent industry of ai-designed drugs. *Nature medicine*, v. 29, p. 1292–1295, 2023.

JIANG, H.; KIM, B.; GUAN, M.; GUPTA, M. To trust or not to trust a classifier. *Advances in neural information processing systems*, v. 31, 2018.

KOMPA, B.; SNOEK, J.; BEAM, A. L. Second opinion needed: communicating uncertainty in medical machine learning. *NPJ Digital Medicine*, Nature Publishing Group UK London, v. 4, n. 1, p. 4, 2021.

KOVALCHIK, S. A. Searching for the goat of tennis win prediction. *Journal of Quantitative Analysis in Sports*, De Gruyter, v. 12, n. 3, p. 127–138, 2016.

MAHER, M. J. Modelling association football scores. *Statistica Neerlandica*, Wiley Online Library, v. 36, n. 3, p. 109–118, 1982.

MARKOU, M.; SINGH, S. Novelty detection: a review—part 1: statistical approaches. *Signal processing*, Elsevier, v. 83, n. 12, p. 2481–2497, 2003.

MARTINEZ-CANTIN, R. Bayesopt: a bayesian optimization library for nonlinear optimization, experimental design and bandits. *J. Mach. Learn. Res.*, v. 15, n. 1, p. 3735–3739, 2014.

MILJKOVIĆ, D.; GAJIĆ, L.; KOVAČEVIĆ, A.; KONJOVIĆ, Z. The use of data mining for basketball matches outcomes prediction. In: IEEE. *IEEE 8th international symposium on intelligent systems and informatics*. [S.l.], 2010. p. 309–312.

NADEEM, M. S. A.; ZUCKER, J.-D.; HANCZAR, B. Accuracy-rejection curves (arcs) for comparing classification methods with a reject option. In: PMLR. *Machine Learning in Systems Biology*. [S.l.], 2009. p. 65–81.

NETO, A. R. da R.; SOUSA, R.; BARRETO, G. de A.; CARDOSO, J. S. Diagnostic of pathology on the vertebral column with embedded reject option. In: SPRINGER. *Pattern Recognition and Image Analysis: 5th Iberian Conference, IbPRIA 2011, Las Palmas de Gran Canaria, Spain, June 8-10, 2011. Proceedings 5*. [S.l.], 2011. p. 588–595.

NICORA, G.; RIOS, M.; ABU-HANNA, A.; BELLAZZI, R. Evaluating pointwise reliability of machine learning prediction. *Journal of Biomedical Informatics*, Elsevier, v. 127, p. 103996, 2022.

PAPPALARDO, L.; CINTIA, P.; ROSSI, A.; MASSUCCO, E.; FERRAGINA, P.; PEDRESCHI, D.; GIANNOTTI, F. A public data set of spatio-temporal match events in soccer competitions. *Scientific data*, Nature Publishing Group UK London, v. 6, n. 1, p. 236, 2019.

PARTIDA, A.; MARTINEZ, A.; DURRER, C.; GUTIERREZ, O.; POSTA, F. Modeling of football match outcomes with expected goals statistic. *Journal of Student Research*, v. 10, n. 1, 2021.

PEDREGOSA, F.; VAROQUAUX, G.; GRAMFORT, A.; MICHEL, V.; THIRION, B.; GRISEL, O.; BLONDEL, M.; PRETTENHOFER, P.; WEISS, R.; DUBOURG, V. et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, JMLR. org, v. 12, p. 2825–2830, 2011.

PLAS, D. Van der; MEERT, W.; VERBRAECKEN, J.; DAVIS, J. A reject option for automated sleep stage scoring. In: *Workshop on Interpretable ML in Healthcare at International Conference on Machine Learning (ICML)*. [S.l.: s.n.], 2021.

ROSSI, A.; PAPPALARDO, L.; CINTIA, P.; IAIA, F. M.; FERNÁNDEZ, J.; MEDINA, D. Effective injury forecasting in soccer with gps training data and machine learning. *PloS one*, Public Library of Science San Francisco, CA USA, v. 13, n. 7, p. e0201264, 2018.

SHAKER, M. H.; HÜLLERMEIER, E. Aleatoric and epistemic uncertainty with random forests. In: SPRINGER. *Advances in Intelligent Data Analysis XVIII: 18th International Symposium on Intelligent Data Analysis, IDA 2020, Konstanz, Germany, April 27–29, 2020, Proceedings 18*. [S.l.], 2020. p. 444–456.

STIVAL, L.; PINTO, A.; ANDRADE, F. d. S. P. d.; SANTIAGO, P. R. P.; BIERMANN, H.; TORRES, R. d. S.; DIAS, U. Using machine learning pipeline to predict entry into the attack zone in football. *PloS one*, Public Library of Science San Francisco, CA USA, v. 18, n. 1, p. e0265372, 2023.

STÜBINGER, J.; MANGOLD, B.; KNOLL, J. Machine learning in football betting: Prediction of match results based on player characteristics. *Applied Sciences*, MDPI, v. 10, n. 1, p. 46, 2019.

TAX, N.; JOUSTRA, Y. Predicting the dutch football competition using public data: A machine learning approach. *Transactions on knowledge and data engineering*, v. 10, n. 10, p. 1–13, 2015.

THARMAKULASINGAM, M.; TOPAL, C.; FERNANDO, A.; RAGIONE, R. L. Backward feature elimination for accurate pathogen recognition using portable electronic nose. In: IEEE. *2020 IEEE International Conference on Consumer Electronics (ICCE)*. [S.l.], 2020. p. 1–5.

URAHAMA, K.; FURUKAWA, Y. Gradient descent learning of nearest neighbor classifiers with outlier rejection. *Pattern Recognition*, Elsevier, v. 28, n. 5, p. 761–768, 1995.

VALERO, C. S. Predicting win-loss outcomes in mlb regular season games–a comparative study using data mining methods. *International Journal of Computer Science in Sport*, v. 15, n. 2, p. 91–112, 2016.

ZHOU, L.; MARTINEZ-PLUMED, F.; HERNÁNDEZ-ORALLO, J.; FERRI, C.; SCHELLAERT, W. Reject before you run: Small assessors anticipate big language models. In: *Proceedings of the EBeM22,IJCAI Workshop on AI Evaluation Beyond Metrics Intelligence*. [S.l.: s.n.], 2022.

APÊNDICE A – DESCRIÇÃO DAS COLUNAS DA TABELA PROCESSADA

Coluna	Descrição
team_name	Nome do mandante
opp_name	Nome do visitante
result	Resultado da partida
competition	Competição
date	Data da partida
gameweek	Rodada da partida
odd_home	Múltipla média para vitória do mandante
odd_draw	Múltipla média para empate do mandante
odd_away	Múltipla média para derrota do mandante
team_Elo	Elo Rating do mandante
opp_Elo	Elo Rating do visitante
form_team_score	Média de gols feitos pelo mandante nos últimos seis jogos
form_opp_score	Média de gols sofridos pelo mandante nos últimos seis jogos
form_passes_forward	Média de passes para frente pelo mandante nos últimos seis jogos
form_passes_back	Média de passes para tráspelo mandante nos últimos seis jogos
form_passes_attack	Média de passes no terço de ataque pelo mandante nos últimos seis jogos
form_passes_mid	Média de passes no terço central do campo pelo mandante nos últimos seis jogos
form_passes_def	Média de passes no terço defensivo pelo mandante nos últimos seis jogos
form_passes_smart	Média de passes inteligentes pelo mandante nos últimos seis jogos
form_passes_key	Média de passes chave pelo mandante nos últimos seis jogos
form_tackle	Média de desarmes tentados pelo mandante nos últimos seis jogos
form_tackle_cplt	Média de desarmes bem sucedidos pelo mandante nos últimos seis jogos
form_tackles_attack	Média de desarmes no terço de ataque pelo mandante nos últimos seis jogos
form_tackles_mid	Média de desarmes no terço central time pelo mandante nos últimos seis jogos
form_tackles_def	Média de desarmes no terço defensivo pelo mandante nos últimos seis jogos
form_off_duels	Média de dribles tentados pelo mandante nos últimos seis jogos
form_off_duels_cplt	Média de dribles completados pelo mandante nos últimos seis jogos
form_opportunities	Média de oportunidades criadas pelo mandante nos últimos seis jogos
form_crosses	Média de cruzamentos tentados pelo mandante nos últimos seis jogos
form_crosses_cplt	Média de cruzamentos completos pelo mandante nos últimos seis jogos
form_shot_onTarget	Média de chutes no gol pelo mandante nos últimos seis jogos
form_accelerations	Média de acelerações pelo mandante nos últimos seis jogos
form_pct_pass_complete	Média de % de passes completos pelo mandante nos últimos seis jogos
form_passes_complete	Média de passes completos pelo mandante nos últimos seis jogos
form_passes	Média de passes tentados pelo mandante nos últimos seis jogos
form_duels	Média de duelos pelo mandante nos últimos seis jogos
form_shots	Média de chutes pelo mandante nos últimos seis jogos
form_win	Média de vitórias pelo mandante nos últimos seis jogos
form_draw	Média de empates pelo mandante nos últimos seis jogos
form_loss	Média de derrotas pelo mandante nos últimos seis jogos
opp_form_team_score	Média de gols feitos pelo time visitante nos últimos seis jogos
opp_form_opp_score	Média de gols sofridos pelo time visitante nos últimos seis jogos
opp_form_passes_forward	Média de passes para frente pelo time visitante nos últimos seis jogos
opp_form_passes_back	Média de passes para trás visitante nos últimos seis jogos
opp_form_passes_attack	Média de passes no terço de ataque pelo time visitante nos últimos seis jogos
opp_form_passes_mid	Média de passes no terço central do campo pelo time visitante nos últimos seis jogos
opp_form_passes_def	Média de passes no terço defensivo pelo time visitante nos últimos seis jogos
opp_form_passes_smart	Média de passes inteligentes pelo time visitante nos últimos seis jogos
opp_form_passes_key	Média de passes chave pelo time visitante nos últimos seis jogos

opp_form_tackle	Média de desarmes tentados pelo time visitante nos últimos seis jogos
opp_form_tackle_cplt	Média de desarmes bem sucedidos pelo time visitante nos últimos seis jogos
opp_form_tackles_attack	Média de desarmes no terço de ataque pelo time visitante nos últimos seis jogos
opp_form_tackles_mid	Média de desarmes no terço central time pelo visitante nos últimos seis jogos
opp_form_tackles_def	Média de desarmes no terço defensivo pelo time visitante nos últimos seis jogos
opp_form_off_duels	Média de dribles tentados pelo time visitante nos últimos seis jogos
opp_form_off_duels_cplt	Média de dribles completados pelo time visitante nos últimos seis jogos
opp_form_opportunities	Média de oportunidades criadas pelo time visitante nos últimos seis jogos
opp_form_crosses	Média de cruzamentos tentados pelo time visitante nos últimos seis jogos
opp_form_crosses_cplt	Média de cruzamentos completos pelo time visitante nos últimos seis jogos
opp_form_shot_onTarget	Média de chutes no gol pelo time visitante nos últimos seis jogos
opp_form_accelerations	Média de acelerações pelo time visitante nos últimos seis jogos
opp_form_pct_pass_complete	Média de % de passes completos pelo time visitante nos últimos seis jogos
opp_form_passes_complete	Média de passes completos pelo time visitante nos últimos seis jogos
opp_form_passes	Média de passes tentados pelo time visitante nos últimos seis jogos
opp_form_duels	Média de duelos pelo time visitante nos últimos seis jogos
opp_form_shots	Média de chutes pelo time visitante nos últimos seis jogos
opp_form_win	Média de vitórias pelo time visitante nos últimos seis jogos
opp_form_draw	Média de empates pelo time visitante nos últimos seis jogos
opp_form_loss	Média de derrotas pelo time visitante nos últimos seis jogos
home_avg_team_score	Média geral de gols em casa do mandante
home_avg_opp_score	Média geral de gols sofridos em casa do mandante
home_avg_win	Média geral de vitórias em casa do mandante
home_avg_loss	Média geral de derrotas em casa do mandante
home_avg_draw	Média geral de empates em casa do mandante
away_avg_team_score	Média geral de gols fora de casa do mandante
away_avg_opp_score	Média geral de gols sofridos fora de casa do mandante
away_avg_win	Média geral de vitórias fora de casa do mandante
away_avg_loss	Média geral de derrotas fora de casa do mandante
away_avg_draw	Média geral de empates fora de casa do mandante
opp_home_avg_team_score	Média geral de gols em casa do time visitante
opp_home_avg_opp_score	Média geral de gols sofridos em casa do time visitante
opp_home_avg_win	Média geral de vitórias em casa do time visitante
opp_home_avg_loss	Média geral de derrotas em casa do time visitante
opp_home_avg_draw	Média geral de empates em casa do time visitante
opp_away_avg_team_score	Média geral de gols fora de casa do time visitante
opp_away_avg_opp_score	Média geral de gols sofridos fora de casa do time visitante
opp_away_avg_win	Média geral de vitórias fora de casa do time visitante
opp_away_avg_loss	Média geral de derrotas fora de casa do time visitante
opp_away_avg_draw	Média geral de empates fora de casa do time visitante