Universidade Federal de Campina Grande
Centro de Engenharia Elétrica e Informática
Unidade Acadêmica de Sistemas e Computação
Programa de Pós-Graduação em Ciência da Computação

Ígor Barbosa da Costa

Modelagem e Predição de Resultados de Futebol Antes e Durante as Partidas Usando Aprendizagem de Máquina

Campina Grande-PB
2021

Universidade Federal de Campina Grande
Centro de Engenharia Elétrica e Informática
Coordenação de Pós-Graduação em Ciência da Computação

Modelagem e Predição de Resultados de Futebol Antes e Durante as Partidas Usando Aprendizagem de Máquina

Ígor Barbosa da Costa

Tese submetida à Coordenação do Curso de Pós-Graduação em Ciência da Computação da Universidade Federal de Campina Grande - Campus I como parte dos requisitos necessários para obtenção do grau de Doutor em Ciência da Computação.

Área de Concentração: Ciência da Computação
Linha de Pesquisa: Sistemas de Computação

Carlos Eduardo Santos Pires
(Orientador)

Campina Grande, Paraíba, Brasil

C837m Costa, Ígor Barbosa da.
Modelagem e predição de resultados de futebol antes e durante as partidas usando aprendizagem de máquina / Ígor Barbosa da Costa. – Campina Grande, 2021.
115 f. : il. color.

Tese (Doutorado em Ciência da Computação) – Universidade Federal de Campina Grande, Centro de Engenharia Elétrica e Informática, 2021.
"Orientação: Prof. Dr. Carlos Eduardo Santos Pires".
Referências.

1. Aprendizagem de Máquina. 2. Sistemas de Computação – Futebol – Análise de Dados. 3. Aprendizagem de Máquina – Futebol – Eficiência de Mercado. I. Pires, Carlos Eduardo Santos. II. Título.

CDU 004.85:796.332(043)

FICHA CATALOGRÁFICA ELABORADA PELA BIBLIOTECÁRIA SEVERINA SUELI DA SILVA OLIVEIRA CRB-15/225

MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DE CAMPINA GRANDE
POS-GRADUACAO CIENCIAS DA COMPUTACAO
Rua Aprigio Veloso, 882, - Bairro Universitario, Campina Grande/PB, CEP 58429-900

FOLHA DE ASSINATURA PARA TESES E DISSERTAÇÕES

IGOR BARBOSA DA COSTA

MODELAGEM E PREDIÇÃO DE RESULTADOS DE FUTEBOL ANTES E DURANTE AS PARTIDAS USANDO APRENDIZAGEM DE MÁQUINA

Tese apresentada ao Programa de Pós-Graduação em Ciência da Computação como pré-requisito para obtenção do título de Doutor em Ciência da Computação.

Aprovada em 09/04/2021

Prof. Dr. Carlos Eduardo Santos Pires - Orientador - UFCG

Prof. Dr. Leandro Balby Marinho - Examinador Interno - UFCG

Prof. Dr. João Arthur Brunet Monteiro - Examinador Interno - UFCG

Prof. Dr. Cláudio Elízio Calazans Campelo - Examinador Interno - UFCG

Prof. Dr. Diego Furtado Silva - Examinador Externo - UFSCar

Prof. Dr. Paulo Salgado Gomes de Mattos Neto - Examinador Externo - UFPE

Documento assinado eletronicamente por **CARLOS EDUARDO SANTOS PIRES**, **PROFESSOR 3 GRAU**, em 23/06/2021, às 17:59, conforme horário oficial de Brasília, com fundamento no art. 8º, caput, da Portaria SEI nº 002, de 25 de outubro de 2018.

Documento assinado eletronicamente por **Paulo Salgado Gomes de Mattos Neto**, **Usuário Externo**, em 24/06/2021, às 10:09, conforme horário oficial de Brasília, com fundamento no art. 8º, caput, da Portaria SEI nº 002, de 25 de outubro de 2018.

Documento assinado eletronicamente por **JOAO ARTHUR BRUNET MONTEIRO**, **PROFESSOR(A) DO MAGISTERIO SUPERIOR**, em 24/06/2021, às 13:33, conforme horário oficial de Brasília, com fundamento no art. 8º, caput, da Portaria SEI nº 002, de 25 de outubro de 2018.

Documento assinado eletronicamente por **CLAUDIO ELIZIO CALAZANS CAMPELO**, **PROFESSOR(A) DO MAGISTERIO SUPERIOR**, em 25/06/2021, às 09:31, conforme horário oficial de Brasília, com fundamento no art. 8º, caput, da Portaria SEI nº 002, de 25 de outubro de 2018.

Documento assinado eletronicamente por **Diego Furtado Silva**, **Usuário Externo**, em 25/06/2021, às 10:07, conforme horário oficial de Brasília, com fundamento no art. 8º, caput, da Portaria SEI nº 002, de 25 de outubro de 2018.

Documento assinado eletronicamente por **LEANDRO BALBY MARINHO**, **PROFESSOR 3 GRAU**, em 08/07/2021, às 17:32, conforme horário oficial de Brasília, com fundamento no art. 8º, caput, da Portaria SEI nº 002, de 25 de outubro de 2018.

A autenticidade deste documento pode ser conferida no site https://sei.ufcg.edu.br/autenticidade, informando o código verificador **1573689** e o código CRC **086A81A7**.

Referência: Processo nº 23096.017374/2021-81 SEI nº 1573689

MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DE CAMPINA GRANDE
POS-GRADUACAO CIENCIAS DA COMPUTACAO
Rua Aprigio Veloso, 882, - Bairro Universitario, Campina Grande/PB, CEP 58429-900

REGISTRO DE PRESENÇA E ASSINATURAS

ATA Nº 003/2021 (TESE Nº 113)

Aos nove (9) dias do mês de abril do ano de dois mil e vinte e um (2021), às nove horas (09:00), de maneira REMOTA, na sala virtual do Google Meet, reuniu-se a Comissão Examinadora composta pelos Professores CARLOS EDUARDO SANTOS PIRES, Dr., lotado(a) no(a) UNIDADE ACADÊMICA DE SISTEMAS E COMPUTAÇÃO da(o) UNIVERSIDADE FEDERAL DE CAMPINA GRANDE, funcionando neste ato como Presidente, LEANDRO BALBY MARINHO, Dr., lotado(a) no(a) UNIDADE ACADÊMICA DE SISTEMAS E COMPUTAÇÃO da(o) UNIVERSIDADE FEDERAL DE CAMPINA GRANDE, JOÃO ARTHUR BRUNET MONTEIRO, Dr., lotado(a) no(a) UNIDADE ACADÊMICA DE SISTEMAS E COMPUTAÇÃO da(o) UNIVERSIDADE FEDERAL DE CAMPINA GRANDE, CLÁUDIO ELÍZIO CALAZANS CAMPELO, PhD., lotado(a) no(a) UNIDADE ACADÊMICA DE SISTEMAS E COMPUTAÇÃO da(o) UNIVERSIDADE FEDERAL DE CAMPINA GRANDE, DIEGO FURTADO SILVA, Dr., lotado(a) no(a) DEPARTAMENTO DE COMPUTAÇÃO, do CENTRO DE CIÊNCIAS EXATAS E DE TECNOLOGIA. da(o) UNIVERSIDADE FEDERAL DE SÃO CARLOS, PAULO SALGADO GOMES DE MATTOS NETO, Dr., lotado(a) no(a) CENTRO DE INFORMÁTICA da(o) UNIVERSIDADE FEDERAL DE PERNAMBUCO. Constituída a mencionada Comissão Examinadora pela Portaria Nº 004/2021 do Coordenador do Programa de Pós-Graduação em Ciência da Computação, tendo em vista a deliberação do Colegiado do Curso, tomada em reunião de 16 de Março de 2021 e com fundamento no Regulamento Geral dos Cursos de Pós-Graduação da Universidade Federal de Campina Grande - UFCG, juntamente com o Sr(a) IGOR BARBOSA DA COSTA, candidato(a) ao grau de DOUTOR em Ciência da Computação, comigo Lyana Silva e Cavalcante Nascimento, assistente em administração, presentes ainda professores e alunos do referido centro e demais presentes. Abertos os trabalhos, o(a) Senhor(a) Presidente da Comissão Examinadora anunciou que a reunião tinha por finalidade a apresentação e julgamento da tese "MODELAGEM E PREDIÇÃO DE RESULTADOS DE FUTEBOL ANTES E DURANTE AS PARTIDAS USANDO APRENDIZAGEM DE MÁQUINA", elaborada pelo(a) candidato(a) acima designado, sob a orientação do(s) Professor(es) CARLOS EDUARDO SANTOS PIRES, com o objetivo de atender às exigências do Regulamento Geral dos Cursos de Pós-Graduação da Universidade Federal de Campina Grande - UFCG. A seguir, concedeu a palavra, pelo prazo regulamentar de sessenta minutos, ao (a) candidato(a), o qual, após salientar a importância do assunto desenvolvido, defendeu o conteúdo da tese. Concluída a exposição e defesa do(a)

candidato(a), passou cada membro da Comissão Examinadora a arguir o(a) doutorando sobre os vários aspectos que constituíram o campo de estudo tratado na referida tese. Terminados os trabalhos de arguição, o(a) Senhor(a) Presidente da Comissão Examinadora determinou a suspensão da sessão pelo tempo necessário ao julgamento da tese. Reunidos, em caráter secreto, no mesmo recinto, os membros da Comissão Examinadora passaram à apreciação da tese, analisando os aspectos concernentes ao domínio do tema, à originalidade, à capacidade, sistematização e pesquisa bibliográfica. Concluída a análise da tese, cada Examinador emitiu o seu julgamento do que se apurou o seguinte resultado: CARLOS EDUARDO SANTOS PIRES, nível APROVADO; LEANDRO BALBY MARINHO, nível APROVADO; JOÃO ARTHUR BRUNET MONTEIRO, nível APROVADO; CLÁUDIO ELÍZIO CALAZANS CAMPELO, nível APROVADO; DIEGO FURTADO SILVA, nível APROVADO; PAULO SALGADO GOMES DE MATTOS NETO, nível APROVADO, tendo assim, o(a) candidato(a) obtido o Conceito APROVADO. Reaberta a sessão, o(a) Presidente da Comissão Examinadora anunciou o resultado do julgamento, tendo, a seguir, encerrado a sessão, da qual lavrei a presente ata, que vai assinada por mim, Lyana Silva e Cavalcante Nascimento, pelos membros da Comissão Examinadora e pelo(a) candidato(a). Campina Grande, 9 de Abril de 2021.

Documento assinado eletronicamente por **CARLOS EDUARDO SANTOS PIRES**, **PROFESSOR 3 GRAU**, em 09/04/2021, às 15:15, conforme horário oficial de Brasília, com fundamento no art. 8º, caput, da Portaria SEI nº 002, de 25 de outubro de 2018.

Documento assinado eletronicamente por **CLAUDIO ELIZIO CALAZANS CAMPELO**, **PROFESSOR(A) DO MAGISTERIO SUPERIOR**, em 09/04/2021, às 15:39, conforme horário oficial de Brasília, com fundamento no art. 8º, caput, da Portaria SEI nº 002, de 25 de outubro de 2018.

Documento assinado eletronicamente por **LYANA SILVA E CAVALCANTE NASCIMENTO**, **ASSISTENTE EM ADMINISTRACAO**, em 09/04/2021, às 18:44, conforme horário oficial de Brasília, com fundamento no art. 8º, caput, da Portaria SEI nº 002, de 25 de outubro de 2018.

Documento assinado eletronicamente por **JOAO ARTHUR BRUNET MONTEIRO**, **PROFESSOR(A) DO MAGISTERIO SUPERIOR**, em 09/04/2021, às 19:49, conforme horário oficial de Brasília, com fundamento no art. 8º, caput, da Portaria SEI nº 002, de 25 de outubro de 2018.

Documento assinado eletronicamente por **Igor Barbosa da Costa**, **Usuário Externo**, em 11/04/2021, às 11:16, conforme horário oficial de Brasília, com fundamento no art. 8º, caput, da Portaria SEI nº 002, de 25 de outubro de 2018.

Documento assinado eletronicamente por **LEANDRO BALBY MARINHO**, **PROFESSOR 3 GRAU**, em 12/04/2021, às 10:39, conforme horário oficial de Brasília, com fundamento no art. 8º, caput, da Portaria SEI nº 002, de 25 de outubro de 2018.

Documento assinado eletronicamente por **Paulo Salgado Gomes de Mattos Neto**, **Usuário Externo**, em 12/04/2021, às 16:29, conforme horário oficial de Brasília, com fundamento no art. 8º, caput, da Portaria SEI nº 002, de 25 de outubro de 2018.

Documento assinado eletronicamente por **Diego Furtado Silva**, **Usuário Externo**, em 20/04/2021, às 15:26, conforme horário oficial de Brasília, com fundamento no art. 8º, caput, da Portaria SEI nº 002, de 25 de outubro de

2018.

A autenticidade deste documento pode ser conferida no site https://sei.ufcg.edu.br/autenticidade, informando o código verificador **1392126** e o código CRC **F18919D5**.

Referência: Processo nº 23096.017374/2021-81 SEI nº 1392126

Dedicatória

Dedico este trabalho às memórias de **Bráulio Maia Júnior**, meu sogro, exemplo de bondade, coragem e simplicidade. **Eduardo Marcelo** e **Felipe Adelino**, jovens amigos, pais de família, que foram levados pela pandemia, mas que deixaram um grande legado de amizade, alegria e amor ao futebol.

Agradecimentos

É o fim da caminhada. Há 5 anos, quando comecei este doutorado, tinha poucas certezas, mas duas delas se confirmam hoje. A primeira é de que conseguiria chegar até aqui. A segunda é de que chegaria muito grato. Assim, chega a hora de agradecer.

Começarei agradecendo aos meus orientadores. Eu posso dizer que tive muita sorte quando chamei **Carlos** para conversar. Levei um tema novo, desafiador, fora da zona de conforto e ele decidiu encarar esse desafio. Desse dia em diante, Carlos preencheu todas as definições de "orientador" que um dicionário pode apresentar. Um guia, um mentor, um modelo. Organizado, colaborativo e responsável. Sempre se propondo a aprender e a ensinar. Para completar minha sorte, Carlos convidou **Leandro** para ser meu co-orientador. Leandro trouxe todo seu vasto conhecimento técnico para apontar os caminhos que eu poderia explorar. Um orientador prestativo e mão-na-massa. Aos dois, toda minha gratidão pela parceria e admiração pelas pessoas que são.

Se na Universidade eu estive bem assessorado, dentro de casa não foi diferente. **Mikaela** foi minha base. Lembro que quando chegamos na Alemanha para meu doutorado sanduíche, fomos ao Oktoberfest em Munique. Lá, em uma das mesas coletivas, conhecemos um casal Holandês e gastamos algumas horas de conversa. Ao conhecer um pouco da nossa história, a holandesa me chamou em particular e falou admirada: *"Você sabia que ela (Mikaela) está abrindo mão da vida dela para estar aqui com você?"*. Sim, eu sempre soube. E não apenas pelo ano que ficamos na Alemanha, mas por todos esses cinco anos de trabalho. Por muitas e muitas vezes, Mikaela teve que abrir mão de alguma coisa para que eu pudesse tocar o barco. Eu definitivamente não teria conseguido sem ela. Só estou tranquilo, pois terei o resto da vida para retribuir esse amor traduzido em abdicação e companheirismo.

Mas a caminhada de um doutor não se restringe aos anos de doutorado. **Lamartine e Marizélia**, meus pais, passaram anos pavimentando essa via para que eu pudesse atravessar de forma segura. Cuidadosos e atenciosos, sempre empreenderam todo esforço para me proporcionar uma educação de qualidade. Não tenho dúvidas de que minha vitória é a vitória deles. É também a vitória da minha parceira de vida e irmã, **Raíssa**. É a alegria de todos meus tios, primos, avós, sobrinhos e cunhados, sogro e sogra, que ao longo dessa caminhada

vibraram com as minhas conquistas. A todo meu **núcleo familiar**, meu caloroso abraço e beijo carinhoso.

Aos amigos de **Bodoca**, do **Treze**, do **Racha**, da **GAF**, dos **Casados**, do **LQD**, do **LSD** e da vida, meu muito obrigado por todas as energias positivas enviadas ao longo dessa jornada. Peço desculpas pelas ausências no momento de aperto e agradeço pelas trocas de mensagens diárias que por muitas vezes deixaram os dias mais leves. De agora em diante, teremos mais tempo para cervejinha e para celebrar as coisas "pequenas" que fazem a vida valer a pena.

Reservo um parágrafo especial para agradecer algumas pessoas que fizeram parte de uma das maiores aventuras da minha vida: o doutorado sanduíche na Alemanha. Começo agradecendo a **Universidade de Hildesheim** e ao **prof. Lars**, que abriram as portas para mim e me deram toda a estrutura para que eu pudesse desempenhar meu trabalho com qualidade. Agradeço a todos meus **companheiros do laboratório ISMLL**, pelas inúmeras trocas de conhecimento sobre ciência, sobre a vida e sobre o mundo. Foi incrível. Em particular, agradeço ao brazuca **Rafael**, um cara especial, de grande coração, que foi sempre um apoio amigo em terras germânicas. Ao meu parceiro de escritório, **Lukas**, um alemão prestativo e bem-humorado, sempre pronto para uma boa conversa sobre futebol. Ao casal albanês, **Josif e Linda**, amigos maravilhosos e acolhedores que proporcionaram momentos inesquecíveis para mim e para Mikaela. E por fim, agradeço aos nossos "europeus tupiniquins" **Catão**, **Lorena**, **Maria Antônia**, **Marcelo** e **Yuska**, que foram um pouco da nossa casa e da nossa família. Todas as nossas viagens e momentos juntos foram combustível indispensável para que eu pudesse concluir essa missão. *Vielen Dank, meine Freunde!*

Agradeço a todos funcionários e docentes da **Universidade Federal de Campina Grande**, em especial aos da **COPIN**, que sempre foram muito solícitos. Da mesma forma, agradeço ao **IFPB - Campus Campina Grande**, principalmente aos meus colegas **docentes da COAIN** que deram todo suporte para que eu tirasse licença durante quatros anos e realizasse esta capacitação importante para nosso grupo, nossos alunos e para nossa instituição.

Por fim, finalizo fazendo um sincero agradecimento a **mim** mesmo. Pode soar pedante, mas a verdade é que mesmo com todo apoio e incentivo, só nós mesmos podemos domar nossos leões. E não foram poucos. Acadêmicos e pessoais. Não houve um só dia nesses cinco anos que eu não estive de alguma forma conectado com este trabalho. Por muitas vezes me senti frustrado e cansado, com resultados que demoram a aparecer. Mas no final,

o que sobra é uma enorme satisfação. Me orgulho muito da minha trajetória. Das minhas renúncias. Da minha dedicação. Hoje, apesar de estar publicando este trabalho científico, não tenho dúvidas de que o maior "produto" deste doutorado é a pessoa que me tornei. Que ao longo da minha vida, eu possa repartir esse aprendizado com meus alunos, amigos e familiares. Principalmente com minha pequena **Isa**, o presente que veio pra deixar o último ano de doutorado ainda mais louco e intenso, mas também mais feliz e prazeroso. Pra você minha filha, todo o melhor de mim, sempre e para sempre.

Resumo

No futebol, a predição de resultados é historicamente uma tarefa desafiadora devido à natureza estocástica do esporte e à complexidade dos inúmeros fatores que influenciam o desenrolar de uma partida. Na última década, com a evolução nas técnicas de aquisição, armazenamento e processamento de grandes volumes de dados, surgiram diversas fontes de dados online (textuais, tabulares, etc.) com informações a respeito de partidas disputadas (escalações, placares, estatísticas, etc.), além de dados referentes às cotações dos mercados de aposta para as partidas. Nesse cenário, técnicas de aprendizagem de máquina podem ser uma alternativa viável para a descoberta de padrões nos dados históricos e consequentemente para a predição de resultados. Este trabalho explora diferentes abordagens e técnicas de aprendizagem, que vão desde classificadores simples a redes neurais complexas, para realização de predições de resultados tanto antes do início das partidas, como também durante (em tempo real). Para a predição pré-jogo, este trabalho ataca um alvo de predição ainda pouco explorado pela literatura: "ambas as equipes vão marcar gols?". Esse tipo de predição, além de ter sido pouco explorada pela literatura, tem despertando um interesse crescente nos últimos anos, devido ao crescimento do mercado de apostas esportivas. Para a predição durante a partida, este trabalho visa estimar o resultado final da disputa, atualizando a predição minuto-a-minuto. Um conjunto de experimentos foi realizado para avaliar o desempenho dos diferentes modelos em termos de acurácia e lucratividade no mercado de apostas. Os resultados obtidos são importantes em vários aspectos, que vão desde a avaliação dos fatores que influenciam o resultado de uma partida até a análise da eficiência de técnicas de aprendizagem de máquina no mercado de apostas esportivas.

Palavras-chave: Futebol; Modelagem Preditiva; Apostas Esportivas; Predição em Tempo Real; Lucratividade; Eficiência de Mercado.

Abstract

In soccer, predicting soccer results is a challenging task due to the sport's stochastic nature and the complexity of the innumerable factors that influence the match's result. In the last decade, with the evolution in the techniques of acquisition, storage, and processing of large volumes of data, several online data sources (e.g., textual and tabular) appeared. These sources contain information on played matches (e.g., lineups, scores, and scouts) and information on betting market movements. In such a scenario, data mining and machine learning techniques can be a viable alternative for discovering patterns in historical data and, consequently, for the prediction of results. This work explores different machine learning techniques, ranging from simple classifiers to complex neural networks, to make predictions both before the start of matches and during the match (in real-time), adjusting the probabilities of results as soon as new events occur (e.g., goals and red cards). In the case of pre-game prediction, this work addresses a problem that is still little explored in the literature: "will both teams score goals?". This type of prediction has aroused growing interest in recent years, due to the growth of the sports betting market. In the case of prediction during the match, the objective is to predict the final result of the game. A set of experiments was carried out to evaluate different techniques in terms of accuracy and profitability in the betting market. The results obtained are important in several aspects, ranging from evaluating the factors that influence the outcome of a match to the analysis of the efficiency of machine learning techniques in the sports betting market.

Palavras-chave: Soccer; Predictive Modelling; Sports Betting; Real-Time prediction; Profitability; Market efficiency.

Lista de Figuras

Lista de Tabelas

Lista de Símbolos

RPS - *Ranked Probability Score*

ACC - *Acurácia*

BRS - *Brier Score*

LUC - *Lucratividade*

ROI - *Return of Investment*

AI - *Aposta Ingênua*

VE - *Valor Esperado*

AP - *Aposta Proporcional*

MU - *Modelo único*

MM - *Múltiplos Modelos*

MTU - *Modelo temporal único*

MTM - *Modelo temporais múltiplos*

Conteúdo

Capítulo 1

Introdução

O futebol é considerado atualmente o esporte mais popular do mundo. Segundo a FIFA[1] - Federação Internacional de Futebol [6], são mais de 250 milhões de praticantes, dos quais 40 milhões são profissionais distribuídos por quase 300 mil clubes de 207 países. Esses números demonstram o interesse mundial pelo esporte que influencia diariamente a vida de aproximadamente 3,5 bilhões de fãs.

Uma das principais características que impulsionam a popularidade desse esporte é sua forte natureza estocástica. O futebol é um esporte dinâmico em que diversos fatores influenciam o desenrolar do jogo e, não raramente, uma equipe de qualidade inferior acaba vencendo uma equipe de qualidade superior.

Essa dinâmica que mistura talento e acaso é historicamente a motivação para uma série de estudos científicos. Desde meados dos anos 50, quando Moroney [7] estudou a distribuição dos gols marcados em um campeonato, cientistas vêm buscando compreender o papel da previsibilidade no futebol. Nesse cenário, uma tarefa permanece sendo desafiadora até mesmo para os especialistas: **a predição de resultados de partidas**. Este é o tema central deste trabalho, cuja proposta é utilizar técnicas de aprendizagem de máquina para realizar a predição de resultados, antes e durante as partidas, como também avaliar o desempenho dessas predições no mercado de apostas.

Este capítulo apresenta a motivação e os objetivos deste trabalho, contextualizando quais aspectos do problema ainda são pouco explorados pela literatura e qual a relevância esportiva, científica, social e econômica desta pesquisa.

[1] Em FIFA - Fédération Internationale de Football Association (em francês)

1.1 Motivação

A busca por compreender melhor os fatores que influenciam o resultado de uma partida de futebol é pauta para diversos grupos de interesse como: torcedores, apostadores, imprensa esportiva, cientistas e desportistas em geral. No campo científico, por exemplo, Anderson & Sally [8] compilaram diversos trabalhos para mostrar que as estatísticas do jogo podem ajudar a compreender o futebol em sua essência e demonstraram como a ciência pode contrariar algumas crendices populares sobre esse esporte como, por exemplo, a ideia de que mais escanteios representam mais gols ou de que a troca de treinador pode mudar os rumos de uma equipe no campeonato.

Apesar de vários aspectos do jogo serem objetos de estudo, a pontuação final de uma disputa é determinada por um único fator: o gol. Diferente de outros esportes em que a pontuação é constante, o gol não é um evento frequente. Isso faz com que, no futebol, o empate seja mais comum do que em outros esportes. Essa combinação de "gols pouco frequentes"e "empates frequentes"fazem do futebol um dos esportes coletivos mais imprevisíveis [9].

Assim, pode-se dizer que a tarefa de prever resultados no futebol passa por encontrar estratégias que controlem ou minimizem essa imprevisibilidade inerente ao esporte. Nas últimas décadas, pesquisas mostraram que o modo como as pessoas fazem avaliações, em situações que envolvem o acaso, costumam ser gravemente deficientes [10]. Portanto, técnicas computacionais podem ser alternativas viáveis para atuar nesse domínio.

Na área da Computação, a predição de resultados no futebol tem a mesma natureza de outros problemas de predição: "a inferência sobre grandezas desconhecidas a partir de valores conhecidos e observados"[11]. Especificamente no futebol, esses valores observados se avolumaram na última década, com a evolução das técnicas de aquisição, armazenamento e processamento de dados, sendo contínuo o crescimento de dados sobre partidas, jogadores, equipes e campeonatos.

Essa abundância de dados tem fomentado o desenvolvimento de diversos modelos de predição para futebol. Na plataforma *Scopus*[2], por exemplo, ao ser consultado o termo *"soccer prediction"* ou *"football prediction"* são encontrados 129 trabalhos, sendo 76 deles publicados nos últimos cinco anos. Esses números mostram o crescente interesse pelo assunto.

[2]Scopus.com - Um dos maiores indexadores de estudos científicos do mundo.

Além disso, algumas competições para criação de modelos de predição no futebol tem movimentado a comunidade científica (Ex. [12]), como também chamadas para edições especiais em periódicos importantes (Ex. [13]).

Nesse contexto, uma grande parte dos trabalhos que fazem predição de resultados têm utilizado técnicas de aprendizagem de máquina para extrair automaticamente conhecimento a partir de dados históricos, sem intervenção humana. Entretanto, mesmo tendo a predição resultados como tarefa básica, os propósitos gerais desses estudos são distintos e podem ser caracterizados de acordo com três perspectivas: esportiva, técnica e econômica.

Na perspectiva *esportiva*, os estudos têm objetivos focados na descoberta de conhecimento, buscando prever resultados, simular campeonatos ou "lançar luz"sobre os fatores que influenciam os resultados das partidas. Na perspectiva *técnica*, os objetivos dos estudos estão mais voltados para os métodos de predição em si, buscando adaptar as técnicas para o domínio e analisar quais delas apresentam melhor desempenho preditivo. Por fim, na perspectiva *econômica*, os objetivos dos trabalhos estão direcionados para avaliar o desempenho das predições no mercado de apostas esportivas.

Essa dissociação de propósitos faz com que a relevância de alguns estudos seja limitada. Trabalhos recentes como [14] e [15] argumentam que muitas pesquisas não comparam as predições de seus modelos com as predições do mercado de apostas ou predições de especialistas. Ambos os trabalhos apontam que as predições do mercado de apostas são o "limite superior"em termos que qualidade de predição, pois representam essencialmente o conhecimento coletivo dos especialistas das casas de apostas e dos próprios apostadores. Em outras palavras, a definição das cotações das casas de apostas "encapsula"o conhecimento de várias pessoas sobre o evento. Dessa forma, os estudos deveriam comparar as predições de seus modelos com as predições do mercado para entender até que ponto o modelo é útil e quais são suas limitações.

Na prática, se os modelos criados forem capazes de realizar melhores predições que o mercado, há indícios de que eles poderiam gerar lucro no mercado de apostas. Essa perspectiva tem recebido enorme atenção nos últimos anos devido ao vertiginoso crescimento do mercado de apostas esportivas em escala global.

Nesse mercado, os clientes estão na constante busca por "bater as casas de apostas", ou seja, obter lucros fazendo previsões. Além de ser uma tarefa que envolve o conhecimento

sobre um esporte popular, também é cercada de divertimento, adrenalina e excitação pelos graus de incerteza e risco inerentes [16].

As apostas esportivas fazem parte da cultura de alguns países, como a Inglaterra, onde historicamente é comum que as pessoas façam suas apostas diariamente nas diversas casas de apostas físicas espalhadas pelo país. Em contrapartida, em países como o Brasil, a abertura de casas de apostas ainda é uma atividade proibida.

Contudo, nos últimos anos, a democratização da Internet está modificando esse cenário. Com o advento das casas de apostas online, pessoas de vários lugares do mundo passaram a realizar apostas em plataformas hospedadas em países onde a atividade é legalizada, o que também acabou impulsionando o mercado informal em países sem regulamentação [17].

Devido a essa informalidade, estimar o valor dessa indústria é uma tarefa difícil, mas acredita-se que ela esteja movimentando valores entre 700 bilhões e 1 trilhão de dólares por ano [18], dos quais cerca de 70% são oriundos de apostas em partidas de futebol. Até mesmo no Brasil, onde o mercado ainda não é regulamentado, já há uma movimentação anual de 6 bilhões de reais [19].

Essa nova tendência mundial está fazendo com que muitos países revejam as leis que regulamentam a atividade de apostas esportivas. A Suprema Corte dos Estados Unidos, por exemplo, derrubou em 2018 a proibição de apostas esportivas no país e alguns estados ianques já começaram a legalizar a operação do mercado. Assim, a perspectiva para os próximos anos é de contínua expansão global, um crescimento que acaba também por impulsionar uma grande questão para os apostadores, economistas e cientistas: *"É possível lucrar no mercado de apostas?"*

Nesse contexto, pode-se concluir que o estudo da predição de resultados é motivado, pelo menos, por três grupos de interesse: o grupo *social e esportivo*, como torcedores, comentaristas, desportistas, que desejam explorar as predições como elementos para debate sobre o jogo; o grupo *técnico*, que se concentra na avaliação dos métodos para resolver um problema difícil num domínio popular; e o grupo *econômico*, que analisa a eficiência de modelos de predição no mercado de apostas.

1.2 Contextualização

Analisando de forma geral as modelagens preditivas em futebol nos trabalhos encontrados na literatura, percebe-se uma divisão natural entre trabalhos que utilizam modelos estatísticos e outros que utilizam aprendizagem de máquina e modelos probabilísticos gráficos. Os trabalhos que usam uma abordagem estatística incluem geralmente modelos de regressão probit ordenado, modelos Poisson ou variações de sistemas de avaliação (*rating*) como *Elo Rating* [20] ou *Pi-Rating* [21]. Os trabalhos que usam abordagem de aprendizagem de máquina apresentam uma diversidade de métodos como algoritmos genéticos, modelos Bayesianos e de Markov, redes neurais, árvores de decisão, *gradient boosting*, entre outros.

Na área econômica, a predição de resultados em futebol está fortemente relacionada com o mercado de apostas esportivas e tem sido analisada pelo prisma da *hipótese do mercado eficiente*. Por definição geral, esta hipótese afirma que mercados financeiros são "eficientes à informação", ou seja, uma pessoa (agente) não consegue alcançar continuamente retornos maiores que a média do mercado, considerando as informações públicas disponíveis no momento em que os investimentos são realizados [22]. No caso das apostas esportivas, essa média do mercado seria "zero", dado que um aposta é um jogo de "soma-zero", no qual o ganho de um lado (o jogador) representa necessariamente a perda para o outro lado (outro jogador/casa de apostas).

Assim, no âmbito das apostas esportivas, pode-se inferir que, se um apostador consegue ter estratégias que geram rentabilidades positivas contínuas, então o mercado é ineficiente; ou seja, as probabilidades calculadas pelo mercado não estariam representando perfeitamente as chances das equipes nas partidas. Dessa forma, vários estudos buscam testar a eficiência do mercado de apostas esportivas e um dos caminhos possíveis para isso é avaliando se modelos automáticos de predição conseguem ser lucrativos.

1.3 Problematização

De uma forma geral, este trabalho realiza predições de resultados antes (pré-jogo) e durante as partidas, e analisa o desempenho dessas predições no mercado de apostas esportivas. Dessa forma, as contribuições deste trabalho podem ser categorizadas a partir de três pers-

pectivas: (i) a predição de resultados no pré-jogo, (ii) a predição de resultados em tempo real; e (iii) o desempenho desses modelos preditivos no mercado de apostas (ver Figura 1.1).

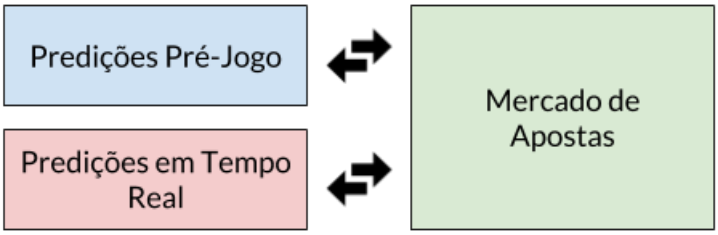


Figura 1.1: Visão geral dos domínios abordados neste trabalho.

A **predição de resultados pré-jogo** já é bastante discutida na literatura. A grande maioria dos trabalhos fazem predição apenas para definir o resultado de uma partida. Com o crescimento do mercado de apostas, outros alvos de predição ganharam importância como, por exemplo, prever a quantidade de gols de uma partida ou se ambas as equipes vão marcar gols em uma partida. Este trabalho ataca um problema não explorado anteriormente: **prever se ambas as equipes irão marcar gols em uma partida**. No mercado de apostas essa predição é conhecida como *"both teams to score"* (BTTS) ou "ambos marcam". O BTTS é um dos ramos mais populares nas apostas em futebol [23] e tem despertado um interesse crescente nos últimos anos (ver Figura 1.2).

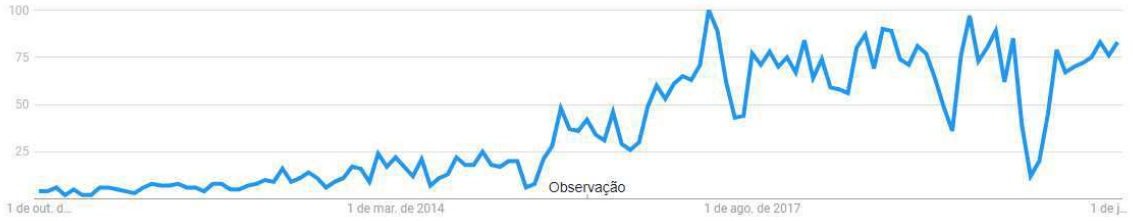


Figura 1.2: Interesse no mercado BTTS (ambos marcam) entre 2010-2020 de acordo com o Google Trends

Além de explorar um novo alvo de predição, este trabalho mitiga outras limitações encontradas em outros estudos, tais como:

- *Limitação do conjunto de dados*: muitos trabalhos utilizam um conjunto restrito de dados observados ou consideram dados de apenas um campeonato. A predição baseada em dados de um único campeonato limita a generalização dos modelos, como

mostrado e [14]. O presente trabalho apresenta a coleta e construção de um conjunto de dados com informações de seis temporadas de nove ligas nacionais diferentes;

- *Limitação das técnicas de predição*: muitos trabalhos utilizam uma única técnica de predição. Este trabalho avalia um amplo conjunto de técnicas diferentes e verifica qual a mais adequada para o problema.

Embora a predição pré-jogo seja bastante explorada, as partidas de futebol geralmente não progridem como esperado inicialmente. As pessoas, em particular os fãs de futebol, muitas vezes conseguem, ao assistir uma partida, distinguir se uma determinada previsão inicial começa a perder força. Assim, mesmo quando o placar não reflete, em algum momento do jogo, o desempenho das equipes, os humanos podem ajustar as previsões sobre o resultado final mais provável. A variabilidade no desempenho de uma equipe durante uma partida pode ser geralmente percebida por meio das estatísticas do jogo (como posse de bola, número de chutes a gol e ataques perigosos). Assim, se essas informações estiverem disponíveis durante a partida, podem ser úteis para alimentar classificadores de aprendizagem de máquina, de forma que os mesmos possam aprender a fazer predições corretas em tempo real.

Nesse contexto, este trabalho ataca o problema da **predição de resultados durante as partidas**, avaliando diferentes técnicas de aprendizagem de máquina que vão desde classficadores tradicionais até redes neurais complexas. Para esse fim, a predição do resultado final da partida deve ser ajustada minuto-a-minuto. Para cada minuto, uma partida pode ser representada por duas estruturas diferentes, como: *estado do jogo*, uma representação da frequência cumulativa de vários eventos até o minuto correspondente ou a *série temporal do jogo*, uma representação da sequência de eventos ocorridos desde o início da partida até o minuto correspondente. No *estado do jogo*, para cada minuto, há um resumo da partida; Na *série temporal do jogo*, além do resumo há também os detalhes de quando cada evento ocorreu. A figura 1.3 mostra as séries temporais formadas por diferentes eventos ao longo do jogo Real Madrid e Liverpool, na final da Champions League 17/18.

Por fim, analisar **o desempenho desses classificadores no mercado de apostas** também traz contribuições significativas para o problema. Como citado anteriormente, apenas uma parcela dos trabalhos comparam seus modelos com predições de especialistas, o que

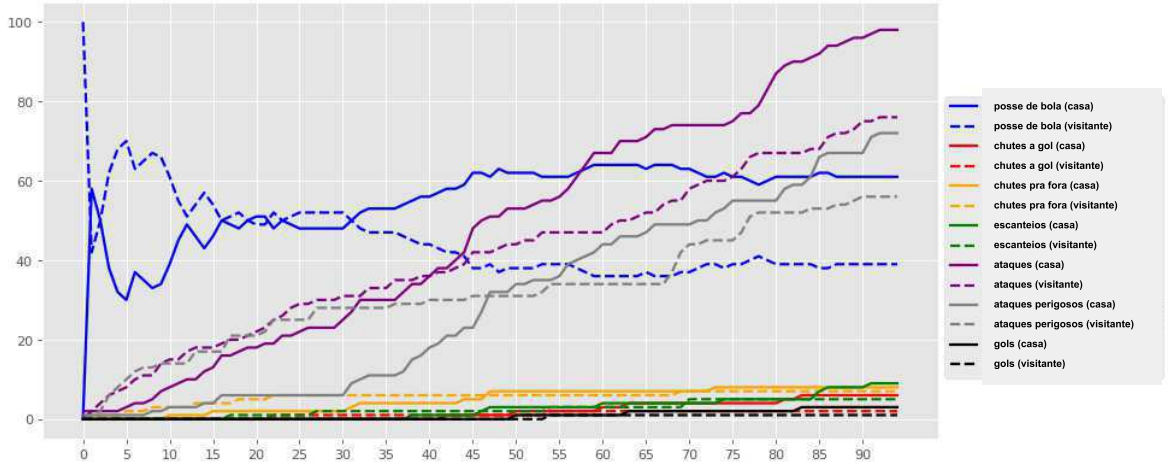


Figura 1.3: Séries temporais de diferentes eventos no jogo Real Madrid vs. Liverpool - Final da Champions League 17/18

pode prejudicar a avaliação do desempenho do modelo. Neste trabalho, as predições dos especialistas serão representadas pelas predições do mercado de apostas esportivas. Dessa forma, é possível comparar as predições dos modelos criados com as predições do mercado de apostas, em termos de acurácia e lucratividade.

1.4 Objetivos

Dado o crescimento do volume de dados disponíveis sobre futebol e constatado o progressivo interesse da comunidade científica pelo tema, este trabalho tem como objetivo geral explorar e avaliar técnicas de aprendizagem de máquina para predição de resultados de futebol antes e durante as partidas. As predições dos modelos são comparadas com as predições do mercado de apostas em termos de acurácia e lucratividade, como também, comparadas com as predições de modelos presentes na literatura.

Assim, a pergunta geral de pesquisa deste trabalho consiste em avaliar se, a partir de dados históricos de partidas futebol, é possível superar as predições do mercado em termos de acurácia e lucratividade utilizando técnicas de aprendizagem de máquina. Nesse contexto, os objetivos específicos deste trabalho são os seguintes:

- Desenvolver indexadores automáticos (*crawlers*) para coletar dados de fontes da Internet, visando a criação de uma base de dados estruturada e integrada com informações

sobre partidas de futebol (campeonatos, resultados, placares, estatísticas, etc.) e cotações do mercado de apostas para essas partidas;

- Transformar os dados coletados em um conjunto estruturado de atributos relavantes para o treinamento das técnicas de aprendizagem de máquina (engenharia de atributos);
- Investigar e avaliar técnicas de aprendizagem de máquina para prever se ambas as equipes vão marcar, antes das partidas começarem (pré-jogo);
- Investigar e avaliar técnicas de aprendizagem de máquina para prever o resultado final de forma contínua, ou seja, no decorrer de uma partida;

1.5 Contribuições

Este trabalho apresenta uma série de inovações e contribuições. Dentre as mais importantes, pode-se destacar:

- Para a predição pré-jogo é explorado um alvo de predição ainda pouco discutido na literatura: a predição de "ambos marcam". Esse é um problema que tem despertado grande interesse do mercado de apostas esportivas, nos últimos anos;
- É apresentado um rico conjunto de dados, contendo informações históricas de partidas de 9 (nove) campeonatos nacionais ao longo de 6 (seis) temporadas. Com esse conjunto de dados, foi realizada uma cuidadosa engenharia de atributos, em que o autor do trabalho aplicou seu conhecimento sobre o domínio para gerar um conjunto de atributos informativos para o processo de aprendizagem de máquina;
- É apresentado um dos primeiros *benchmarks* para o problema de "ambos marcam", avaliando diferentes tipos de classificadores, como baseados em Poisson, baseados em aprendizagem de máquina e baseados no mercado de apostas. Nesse contexto, são comparados modelos que tiveram sucesso em trabalhos anteriores;
- É realizada uma série de experimentos que apontam que a predição de "ambos marcam"é um problema bastante difícil e que os classificadores baseados na opinião do mercado de apostas são difíceis de serem superados;

- São experimentadas diferentes estratégias de apostas e algumas delas demonstram ser capazes de superar o mercado e obter um lucro sistemático ao longo de duas temporadas, dando indícios de que o mercado não consegue ser completamente eficiente para a predição de "ambos marcam";

- Para a predição de resultados durante a partida, este trabalho apresenta a construção de um rico conjuntos de dados estruturado que inclui eventos de partidas e *odds* do mercado de apostas, minuto-a-minuto. A partir desse *conjunto de dados*, são extraídas séries temporais multivariadas. Esse tipo de dado ainda não havia sido explorado por trabalhos anteriores.

- Para a predição durante a partida, também é apresentado um dos primeiros *benchmarks* para o problema. Foram experimentadas diferentes abordagens, como classificadores capazes de prever qualquer minuto do jogo e conjunto de classificadores treinados individualmente para cada minuto da partida. Para cada abordagem foram avaliados diferentes tipos de técnicas de aprendizagem de máquina;

- Além de avaliar diferentes arquiteturas de redes neurais que tiveram bons desempenhos com classificação de séries temporais multivariadas, este trabalho buscou encontrar uma arquitetura específica para o problema. Para isso, foi usada uma combinação de ajuste manual e busca em *grid gridsearch*. Como resultado, é apresentado uma arquitura composta por uma camada convolucional (CNN) seguida de um LSTM (Long Short Term Memory) Bi-Direcional, que apresentou os melhores resultados, em termos de predição;

- Foram realizados diversos experimentos, que mostraram que os melhores classificadores precisam de cerca de 14 minutos, em média, para fazer uma predição tão boa quanto o mercado de apostas faz antes da partida iniciar. Além disso, é possível observar que as redes neurais mais complexas conseguiram extrair valor da séries temporais, refinando a predição, ainda que isso só ocorra em um momento específico do segundo tempo do jogo.

1.6 Método de Pesquisa

Visando a proposição de soluções para os problemas e desafios abordados, o método utilizado divide esta pesquisa em duas atividades: a predição antes das partidas iniciarem e a predição no decorrer das partidas.

A primeira atividade é semelhante ao que é realizado pela maioria dos trabalhos relacionados, nos quais a predição é feita antes de a partida iniciar, utilizando dados históricos sobre as equipes envolvidas.

A segunda atividade está focada na predição após a partida começar, no qual as probabilidades são atualizadas continuamente até o fim da disputa. Para esse fim, as predições devem levar em consideração eventos importantes ocorridos durante as partidas, como ocorrência de gols, chutes a gol, atribuição de cartões e principalmente o decorrer do tempo.

Em ambos os casos, o desenvolvimento das atividades deve seguir um processo de descoberta de conhecimento iterativo e interativo com os seguintes passos:

1. Coleta de dados automática a partir de fontes da Internet;
2. Análise exploratória dos dados coletados para identificação de padrões preliminares;
3. Pré-Processamento e seleção de dados relevantes para "alimentar"as técnicas de aprendizagem de máquina (*Feature Engineering*);
4. Modelagem preditiva usando técnicas de aprendizagem de máquina;
5. Avaliação dos resultados obtidos a partir de métricas de acurácia e lucratividade;
6. Análise crítica do desempenho e formulação de novas hipóteses para recomeçar o ciclo.

1.7 Estrutura do Documento

O restante deste documento está estruturado da seguinte forma. No próximo capítulo, serão apresentados os fundamentos teóricos necessários para o melhor entendimento deste trabalho. No Capítulo 3, serão discutidas as características dos trabalhos relacionados ao tema.

No Capítulo 4, serão detalhadas as etapas, do processo de modelagem, realizadas para a predição pré-jogo para "ambos marcam", enquanto que, no Capítulo 5, para predições durante as partidas. Por fim, no Capítulo 6, serão apresentadas as conclusões e as propostas para trabalhos futuros.

Capítulo 2

Fundamentação Teórica

Neste capítulo, serão abordados os fundamentos teóricos necessários para melhor compreensão deste documento. Primeiramente, serão discutidos fundamentos das áreas de modelagem preditiva e aprendizagem de máquina, bem como destacados princípios importantes encontrados nos trabalhos de predição de resultados de futebol. Em seguida, serão discutidos os conceitos básicos dos mercados de apostas, fundamentais para melhor compreensão de parte do conjunto de dados a ser usado nos experimentos deste trabalho.

2.1 Técnicas de Modelagem Preditiva

Na aprendizagem de máquina preditiva, as técnicas existentes utilizam dados rotulados para indução de modelos de *classificação* ou de *regressão*. O conjunto de dados rotulados é tipicamente denotado como conjunto de treinamento. Os rótulos do conjunto de treinamento correspondem a classes ou valores obtidos a partir de alguma função desconhecida. Dessa forma, as técnicas de aprendizagem de máquina buscam produzir um modelo capaz de generalizar as informações contidas em um conjunto de treinamento, com a finalidade de obter rótulos de objetos desconhecidos.

Formalmente, um conjunto de treinamento pode ser definido como uma coleção de entradas e saídas $\{(x_i, y_i)\}_{i=1}^{n}$, no qual cada tupla, x_i é um vetor que representa um coleção de k características, e y_i indica o rótulo que correspondente a x_i. Cada y_i foi gerado por uma função desconhecida de $y = f(x)$. A tarefa da aprendizagem de máquina é descobrir uma função que se aproxima da verdadeira f. No casos em que o valor de y_i for um conjunto fi-

nito de valores, trata-se de uma tarefa de *classificação*; quando tais valores forem contínuos, trata-se de uma tarefa de *regressão* [24].

Em alguns cenários, o vetor x_i pode conter um conjunto de séries temporais. Nesse contexto, o problema de predição pode ser considerado como um problema de classificação de séries temporais multivariadas (*mutivariate time series classification - MTSC*). MTSC é um caso particular de classificação, no qual para cada tempo t, o vetor x_i contém N diferentes séries temporais que occorrem sincronamente e têm tamanho t, ou seja, $x_i = (F^1, F^2, \ldots, F^N)$: $F^i \in \mathbb{R}^t$.

Dado que existem diversas técnicas para execução das tarefas de classificação e regressão. Esta seção apresenta uma descrição das técnicas relevantes para este trabalho.

2.1.1 *Gaussian Naive Bayes*

Gaussian Naive Bayes é um classificador probabilístico simples, baseado na aplicação do Teorema de Bayes [25]. Esse classificador é denominado "ingênuo"(*Naive*) pois pressupõe a independência entre os atributos, ou seja, desconsidera completamente a existência de correlação entre as variáveis aleatórias de entrada.

A principal vantagem desse classificador é que, devido a sua simplicidade, requer apenas uma pequena quantidade de dados para estimar as médias e variâncias necessárias para a classificação. Dessa forma, sua aplicação também é extremamente rápida.

Para lidar com dados contínuos, o classificador assume que os valores associados a cada atributo são independentes e quando agrupados estão distribuídos de acordo com uma *função de densidade de probabilidade gaussiana* [26]. Formalmente, para algum valor v do vetor de atributos F, a probabilidade da distribuição de v, dada uma classe c é:

$$P(F = v|c) = \frac{1}{\sqrt{2\pi\sigma_c^2}} exp \frac{(-v - \mu_c)^2}{2\sigma_c^2} \tag{2.1}$$

no qual σ_c é o desvio padrão da distribuição e μ_c é a média da distribuição.

2.1.2 Regressão Logística

A regressão logística é uma técnica estatística que visa produzir, a partir de um conjunto de observações, um modelo que permita predizer variáveis categóricas (valores discretos), fre-

quentemente binárias, em função de uma ou mais várias independentes. A partir do modelo gerado, é possível calcular ou prever a probabilidade um determinado evento ocorrer em uma uma nova observação.

Formalmente, considere Y uma variável binária. A função de probabilidade $P(x) = P(Y = 1|X = x)$ pode ser usada para prever a probabilidade de $Y = 1$, e x representa um conjunto de variáveis independentes. Para representar a relação entre a probabilidade $P(x)$ e as vairáveis independentes x, considere uma função logit ϕ:

$$\phi(x) = \frac{P(x)}{1 - P(x)} = \beta_o + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n \tag{2.2}$$

no qual $\beta_0, \beta_1, ..., \beta_n$ são os coeficientes de regressão. Por fim, a função logit ainda está em uma escala irrestrita, então uma função sigmóide é usada para restringir a saída para um valor entre 0 e 1 (ver Equação 2.3)

$$P(x) = \frac{1}{1 + \epsilon^{-(\phi(x))}} \tag{2.3}$$

2.1.3 *Gradient Boosting*

Entre as técnicas que combinam vários modelos (*ensemble methods*) em um único preditor estão as técnicas de *bagging* e *boosting*. Na técnicas de *bagging*, os modelos são construídos de forma independente e combinados a partir de alguma estratégia. No *boosting*, os modelos não são construídos de forma independente, mas sim, de forma sequencial. A técnica de *boosting* é baseada na lógica de que um preditor subsequente aprende com os erros dos preditores anteriores. Dessa forma, o *Gradient Boosting* (GB) é uma técnica que combina vários modelos fracos em um único preditor, utilizando a estratégia de *boosting*. Em outras palavras, GB é um método iterativo que faz a combinação de funções parametrizadas simples com desempenho limitado para produção de uma regra de predição mais precisa.

2.1.4 Redes Neurais Artificiais

Redes neurais artificiais (RNA) é um modelo computacional baseado na arquitetura das redes neurais biológicas. Uma RNA consiste em um grupo de unidades de computação interconectadas (neurônios artificiais) que transformam um conjunto de entradas em uma saída

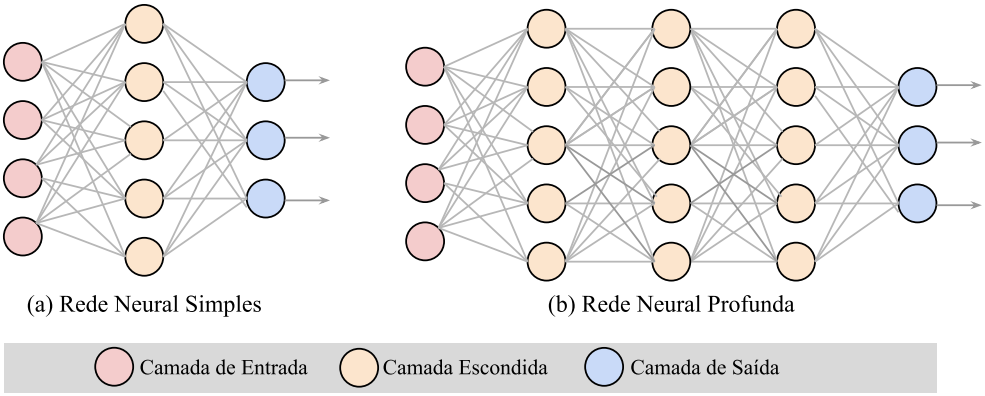


Figura 2.1: Exemplo de Arquitetura de Redes Neurais

desejada (Ver Figura 2.1). Na RNA, os pesos associados aos componentes interconectados são ajustados continuamente durante a etapa de treinamento, de forma que permitam a rede atingir os melhores níveis de predição. Assim, a saída da RNA geralmente depende dos recursos de entrada e dos pesos ajustados em suas conexões.

Com os avanços da tecnologia de hardware na última década, tornou-se possível a criação de redes neurais mais profundas e complexas. Nesse contexto, surge o conceito de aprendizagem profunda (*Deep Learning*). A aprendizagem profunda é responsável por avanços recentes em visão computacional, reconhecimento de fala, processamento de linguagem natural e reconhecimento de áudio [27]. A seguir, são detalhadas algumas das arquiteturas de aprendizagem profunda discutidas no presente trabalho.

Fully Convolutional Neural Network (FCN)

FCN tem mostrado qualidade e eficiência para classificação de séries temporais multivariadas [28]. Em muitos desses problemas, a FCN é executada como um extratora de atributos. O bloco básico da arquitetura é uma camada convolucional (*Convolutional Layer*) seguida por uma camada de normalização em lote (*batch normalization*) e uma camada de ativação ReLU. A operação de convolução é realizada por três núcleos 1-D com os tamanhos {8, 5, 3}. No presente trabalho, a FCN segue o padrão de construção dado por Wang et. al. [29], no qual três blocos de convolução são empilhados, usando {128,256,128} filtros em cada bloco, respectivamente. Após os blocos de convolução, os recursos são alimentados em uma camada de global pooling média (*global average pooling layer*) em vez de uma camada to-

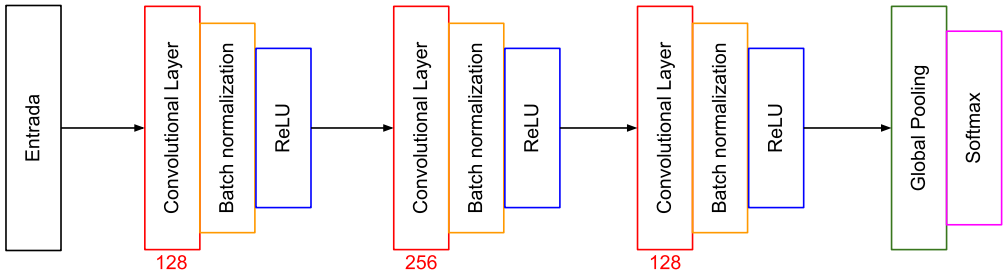


Figura 2.2: Arquitetura de uma Fully Convolutional Neural Network (FCN)

talmente conectada, o que reduz em grande parte o número de pesos. Sua saída final é dada por uma camada *softmax* (Ver Figura 2.2).

InceptionTime

Embora inicialmente aplicada a problemas de classificação de imagens, *InceptionTime* tem tido resultados muito primissores em uma ampla variedade de problemas que envolvem séries temporais [1].

A arquitetura dessa rede contém dois blocos residuais diferentes (Ver Figura 2.3). Cada bloco é composto por três módulos conhecidos como *Inception* [30], em vez de camadas convolucionais tradicionais. A ideia central do módulo *Inception* é aplicar vários filtros simultaneamente a uma série temporal de entrada. Esse módulo inclui filtros de vários tamanhos que permitem que a rede extraia automaticamente recursos relevantes de séries temporais longas e curtas. Seguindo esses blocos residuais, emprega-se uma camada *Global Average Pooling*. Por fim, aplica-se uma camada *softmax* totalmente conectada, com um número de neurônios igual ao número de classes do conjunto de dados.

2.1.5 RNN-LSTM e RNN-BiLSTM

A Rede neural recorrente (RNN) é uma arquitetura muito útil no caso de entradas de tamanhos variados, como séries temporais, e consequentemente para problemas como tradução automática, reconhecimento automático de voz e reconhecimento automático de padrões. Entretanto, ela tem seu uso limitado, porque pode sofrer do problema de dissipação do gradiente (*vanishing gradient problem*) [31] que faz com que a rede pare de aprender durante o treinamento.

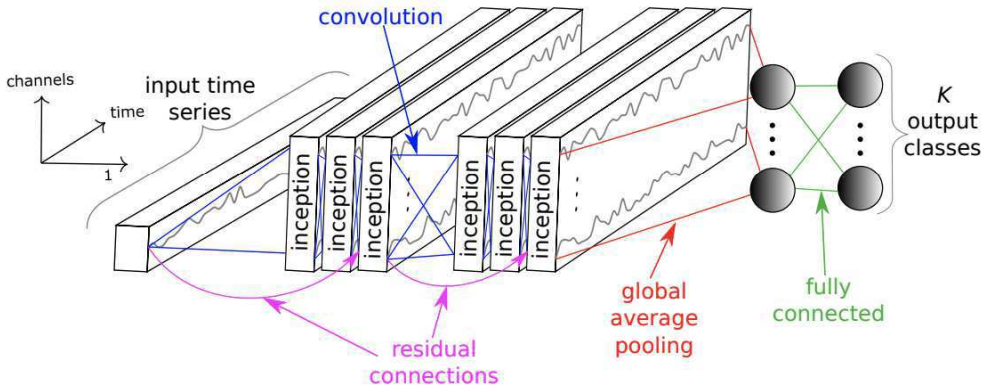


Figura 2.3: Arquitetura de uma InceptionTime (Extraído de [1])

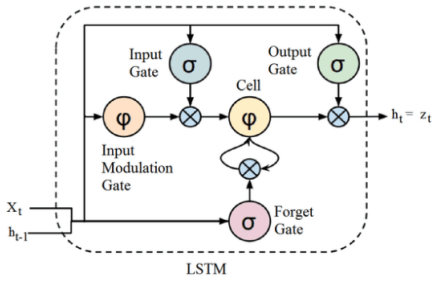


Figura 2.4: Arquitetura de uma célula LSTM

Uma *Long Short-Term Memory Neural Network* (LSTM) é uma extensão de uma rede neural recorrente, proposta por Hochreiter e Schmidhuber [32]. Ela é capaz de resolver o problema de dissipação do gradiente graças à sua memória, que permite ler, escrever e apagar os dados através de três portas: uma porta que permite/bloqueia atualizações (*input gate*); uma segunda porta que desabilita um neurônio se não for importante, com base nos pesos aprendidos pelo algoritmo (*Forget Gate*); e a terceira porta que controla o estado do neurônio na saída (*output gate*) (Ver Figura 2.4).

A BiLSTM é uma variação da LSTM que processa os dados em duas direções, pois trabalha com duas camadas ocultas (Ver Figura 2.5). A BiLSTM provou bons resultados principalmente no processamento de linguagem natural [33].

2.1.6 Cadeias de Markov

Uma cadeia de *Markov* é um caso particular de processo estocástico com estados discretos, no qual a distribuição de probabilidade do próximo estado depende apenas do estado atual e

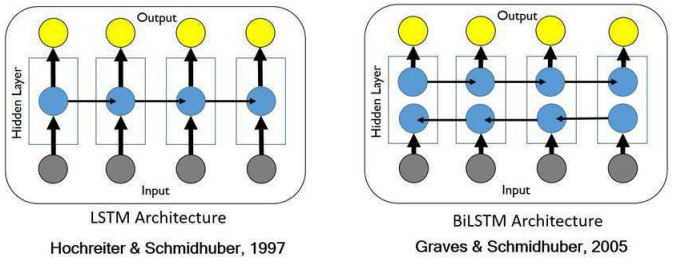


Figura 2.5: Arquitetura de uma rede LSTM e uma rede BiLSTM (extraído de [2])

não da sequência de estados passados. Formalmente, uma cadeia de Markov tem um grupo de estados $S = \{s_1, s_2, ..., s_n\}$. O processo se inicia em algum desses estados e, ao longo tempo, é possível se mover de um estado para outro. A probabilidade de haver transição de um estado s_i para o estado s_j é denotada por p_{ij}. Essas probabilidades não dependem de quais estados o processo esteve antes, mas apenas do estado atual. A cada mudança de estado, as probabilidades de transição são atualizadas.

As cadeias de Markov são frequentemente descritas por um grafo, onde as arestas do grafo são rotuladas pelas probabilidades de ir de um estado, no tempo n, para outros estados no tempo $n + 1$. A Figura 2.6 mostra um exemplo de cadeia de Markov com três estados e os valores das probabilidades de transição entre eles.

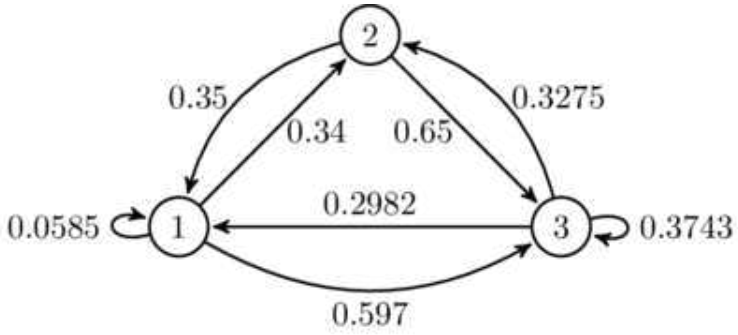


Figura 2.6: Caracterização de uma cadeia de Markov (extraída de [3])

2.2 Avaliação de Desempenho de Modelos

As técnicas de aprendizagem de máquina constroem seus modelos baseados em dados históricos. Entretanto, para avaliar o desempenho do modelo, é necessário aplicá-lo em um

conjunto de dados não conhecido previamente. Assim, é necessário dividir o conjunto de dados em duas partes: conjunto de treinamento e conjunto de teste.

O *conjunto de treinamento* contém os dados históricos utilizados para construção do modelo. O *conjunto de testes* é usado para avaliar o desempenho real do modelo segundo alguma métrica de avaliação.

Durante a fase de treinamento, é comum que as técnicas utilizem um subconjunto dos dados para validação. Esse subconjunto é usado para fornecer uma avaliação imparcial da aprendizagem, enquanto os parâmetros do modelo são estimados (otimizados).

Existem algumas técnicas para realizar a divisão desse conjunto de dados. Neste trabalho, denominamos essas técnicas como *Métodos de Avaliação de Desempenho*. Nesta seção, serão detalhados os *métodos de avaliação de desempenho* e as *métricas de avaliação de qualidade* relevantes para compreensão deste documento.

2.2.1 Métodos de Avaliação de Desempenho

Nesta seção serão apresentados os seguintes métodos de avaliação de desempenho: *holdout*, *cross-validation*, *growing window* e *sliding window*.

HoldOut

O método *holdout* consiste na divisão do conjunto de dados em dois subconjuntos mutuamente exclusivos, sendo um para treinamento e outro para teste. Não há uma padrão para a divisão, mas é comum que seja considerado 2/3 dos dados para treinamento e o 1/3 restante para teste (ver Figura 2.7).

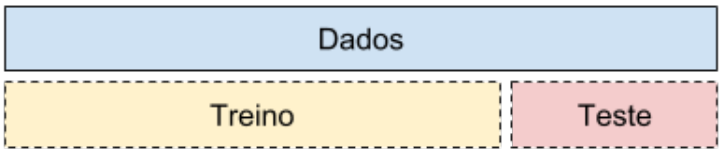


Figura 2.7: Divisão de conjunto de dados de acordo com o método *holdout*.

Cross-Validation

O método *cross-validation* ou *validação cruzada* consiste na divisão do conjunto de dados em k subconjuntos de tamanhos aproximadamente iguais. A cada iteração, uma das partições é usada para teste, enquanto as demais são utilizadas para treino. Dessa forma, esse processo é realizado k vezes e, a cada ciclo, uma partição diferente é usada para teste. O desempenho final é calculado pela média dos desempenhos observados em cada iteração (ver Figura 2.8).

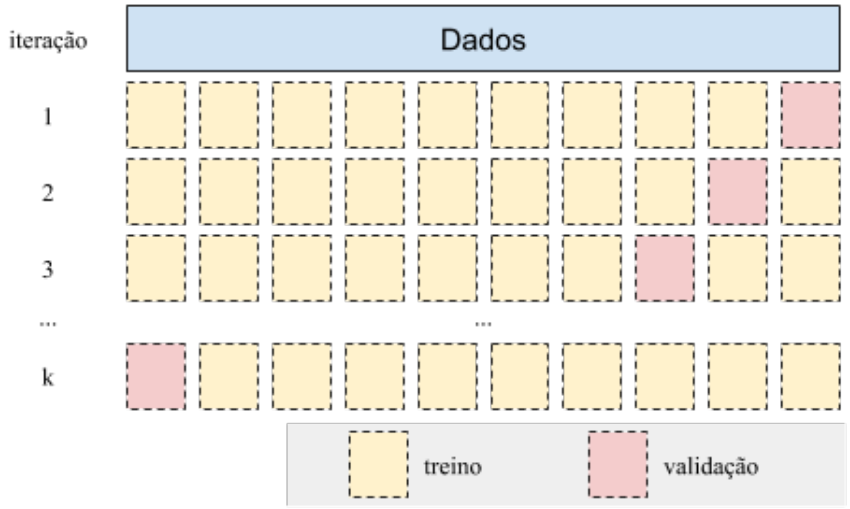


Figura 2.8: Divisão de conjunto de dados segundo o método *cross-validation*.

Growing Window

O método *growing window* ou *landmark window* é um método utilizado para conjuntos de dados com características temporais. Nesse método, o conjunto de treinamento vai crescendo à medida que se faz as predições com os dados de teste. Dessa forma, após cada iteração, os dados que foram usados para testes são adicionados ao conjunto de treinamento da próxima iteração (ver Figura 2.9).

Sliding Window

O método *sliding window* ou janela deslizante também é um método utilizado para conjuntos de dados com características temporais. A diferença para o "Growing Window"é que, cada vez que novos dados são adicionados ao conjunto de testes, um subconjunto de tama-

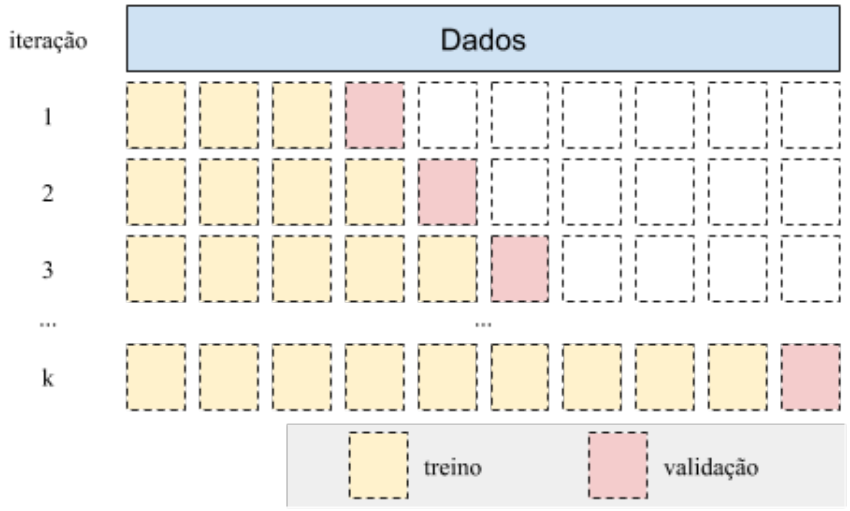


Figura 2.9: Divisão de conjunto de dados de acordo com o método *growing window*.

nho semelhante, com dados mais antigos, é descartado. Isso faz com que o conjunto de treinamento considere apenas uma "janela de dados mais recentes"em todas as iterações (ver Figura 2.10).

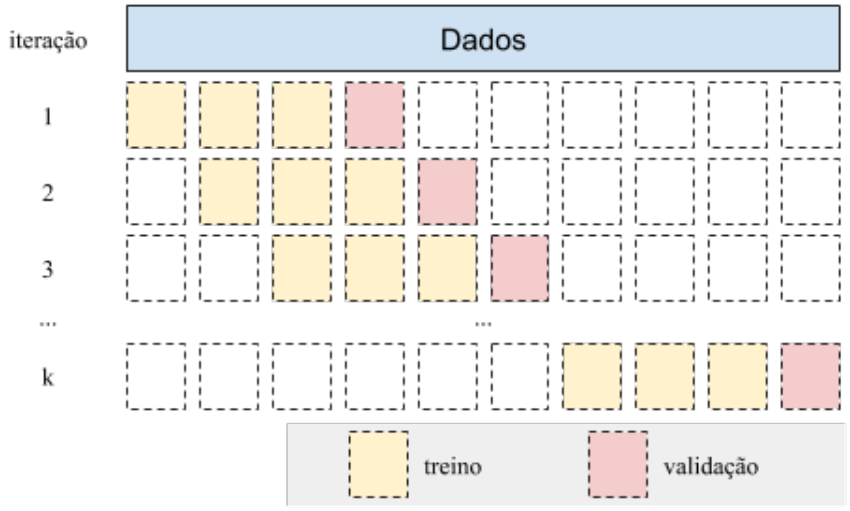


Figura 2.10: Divisão de conjunto de dados através do método *sliding window*.

2.2.2 Métricas de Avaliação de Desempenho

Nesta seção, serão apresentadas as métricas de avaliação de desempenho usadas neste trabalho.

Acurácia e Taxa de Erro

Em predições que envolvem a tarefa de classificação, duas métricas são bastante utilizadas: *acurácia* e *taxa de erro*. A acurácia tem como função identificar a quantidade de predições corretas em relação ao número total de predições. Formalmente:

$$\text{acurácia} = \frac{\text{número de predições certas}}{\text{número total de predições}} \tag{2.4}$$

De forma equivalente, a taxa de erro tem como função identificar a quantidade de predições erradas em relação ao número total de predições. Formalmente:

$$\text{taxa de erro} = \frac{\text{número de predições erradas}}{\text{número total de predições}} \tag{2.5}$$

Dessa forma, os métodos de classificação buscam criar modelos que tenham maior acurácia ou, equivalentemente, menor taxa de erro.

Brier Score

A acurácia é a porcentagem de previsões corretas, enquanto o *Brier Score* mede o desempenho do modelo na previsão da probabilidade de cada classe. Neste trabalho, todas as probabilidades são expressas como porcentagens. Formalmente, considerando N o número de jogos, m uma instância de um jogo, $\hat{y}_m$ a probabilidade prevista pelo classificador para *ambas marcam* e o_m o resultado real (0 se não acontecer; 1, caso contrário), pode-se definir o *Brier Score* formalmente como:

$$BRS = \frac{1}{N}\sum_{m=1}^{N}(\hat{y}_m - o_m)^2 \tag{2.6}$$

Para a questão de pesquisa PP3, os modelos foram avaliados em termos de lucratividade (*LUC*) e Retorno do Investimento (*RoI*). A lucratividade mede quanto dinheiro se ganha ou perde (saldo) depois de realizar uma aposta com base em uma determinada estratégia. Em outras palavras, se uma aposta acerta determinado resultado, um lucro é obtido com base nas probabilidades oferecidas pela casa de apostas. Por outro lado, se a aposta falha, perde-se todo o valor apostado. Assim, *lucratividade* pode ser definida como a soma desses ganhos e perdas, considerando um conjunto de jogos, enquanto o *RoI* mede o ganho ou a perda obtida em um conjunto de apostas em relação à quantidade de dinheiro apostado.

Ranked Probability Score

O RPS (*Ranked Probability Score*) é uma métrica formulada por Epstein [34] que passou a ser adotada para avaliação de modelos de predição de futebol a partir de [35].

A equação do RPS é definida por:

$$RPS(p_1, ...p_{r-1}, a_1, ...a_{r-1}) = \frac{1}{r-1} \sum_{i-1}^{r-1} (\sum_{j=1}^{1} (p_j - a_j))^2 \tag{2.7}$$

em que r é o número de resultados possíveis em uma partida (r=3 no caso de previsões de resultados de futebol), p_j é a probabilidade prevista para um determinado resultado j, ou seja, $p \in [0, 1]$ para $j = 1, 2, ..r$, e a_j indica se o resultado j ocorreu, sendo $j = 1$ se ocorreu e $j = 0$ caso contrário.

2.3 Apostas Esportivas

O mercado de apostas esportivas pode ser dividido em dois segmentos principais: Casas de Apostas e Bolsas de Apostas. O segmento de Casa de Apostas é o modelo mais tradicional, no qual um apostador palpita sobre um determinado evento e, se obtiver sucesso, a Casa de Apostas paga o prêmio equivalente à aposta realizada. Caso o palpite não tenha sucesso, o apostador perde o valor investido para a Casa de Apostas. São exemplos de Casas de Apostas a Bet365[1], Sportingbet[2] e 188bet[3]. Já no segmento de Bolsa de Apostas, que tem tido grande crescimento nos últimos anos graças à democratização da Internet, os apostadores "casam"apostas diretamente entre si através de uma plataforma online que serve como mediadora dos acordos. Dessa forma, enquanto nas Casas de Apostas um apostador aposta contra "a casa", na Bolsa de Apostas, um apostador aposta contra outro apostador. São exemplos de Bolsas de Apostas, a Betfair[4] e Betdaq[5].

A seguir, serão apresentados os conceitos básicos relacionados às apostas esportivas e discutido o funcionamento das Casas de Apostas e Bolsas de Apostas, através de exemplos para melhor compreensão do domínio.

[1] `https://www.bet365.com`
[2] `https://www.sportingbet.com`
[3] `https://www.188bet.com`
[4] `https://www.betfair.com`
[5] `https://www.betdaq.com`

2.3.1 O conceito de *Odds*

Em uma aposta esportiva, um apostador investe um determinado valor em um certo resultado e, caso o resultado ocorra, recebe um prêmio pelo acerto. O valor desse prêmio é definido por uma cotação conhecida como *odd*. O conceito de *odd* pode ser visto sob diferentes perspectivas. Inicialmente, considere *odd* como a cotação que determina quanto um apostador poderá lucrar ao apostar em determinado resultado. Essas cotações podem ser apresentadas em diferentes formatos: Odd *Decimal* (europeia), *Odd Fracionária* (inglesa) ou *Odd Americana*. Neste trabalho, serão usadas as *odds decimais* por serem as mais difundidas mundialmente.

A *odd decimal* é por norma um valor em formato de número inteiro ou decimal maior que 1 (um). De forma geral, para calcular qual é a *odd* de um determinado resultado, divide-se 1 (um) pela probabilidade daquele resultado ocorrer. Dessa forma, considerando $prob$ a probabilidade de um determinado evento ocorrer, a sua *odd* pode ser calculada pela Equção 2.8.

$$odd = \frac{1}{prob} \tag{2.8}$$

Para determinar quanto um apostador pode ganhar apostando em um certo resultado, além do valor da *odd*, é necessário levar em consideração qual foi a quantia apostada. Dessa forma, considerando v, o valor apostado, pode-se definir o possível lucro w como:

$$w = v * (odd - 1) \tag{2.9}$$

Para melhor entendimento, analise o exemplo da final da Copa do Mundo de 2018, entre França e Croácia. Na Casa de Apostas Bet365 [36], uma das mais populares do mundo, as *odds* para o resultado final antes da partida eram:

- França: $1,90$
- Empate: $3,30$
- Croácia: $5,00$

Nesse cenário, caso um apostador realizasse uma aposta de \$10 em um dos resultados, os possíveis lucros seriam:

- Aposta na França: $\$10 * (1.9 - 1) = \$9,00$
- Aposta no Empate: $\$10 * (3.30 - 1) = \$22,30$
- Aposta na Croácia: $\$10 * (5.00 - 1) = \$40,00$

É importante observar que, quanto menor o valor da *odd* (mais próximo de 1), maior é a probabilidade de aquele evento ocorrer e, consequentemente, menor é o possível lucro daquela aposta. Da mesma forma, quanto maior o valor da *odd*, menor é a probabilidade do evento ocorrer e consequentemente maior é o possível lucro para uma aposta.

2.3.2 *Overround* e Balanceamento de *Odds*

Na seção anterior, foi observado que a cotação das *odds* além de representar o valor a ser pago a um apostador, também representa implicitamente as probabilidades de um determinado evento ocorrer. Sabe-se que, em uma divisão justa, a soma das probabilidades de todos os resultados possíveis deve totalizar 100%. As Casas de Apostas, no entanto, não oferecem *odds* justas, pois precisam obter uma margem de lucro. Para melhor entendimento, considere novamente o caso da final da Copa do Mundo de 2018. Reescrevendo a Equação 2.8, pode-se calcular as probabilidades que geraram as *odds* da seguinte forma:

$$p = \frac{1}{odd} \tag{2.10}$$

Sendo assim, calculando as probabilidades para cada resultado do exemplo anterior, têm-se:

- França: $\frac{1}{1.90} \approx 0.5263 \approx 52,63\%$
- Empate: $\frac{1}{3.30} \approx 0.3030 \approx 30,30\%$
- Croácia: $\frac{1}{5.00} = 0.20 = 20\%$

Observa-se que a soma das probabilidades calculadas totaliza aproximadamente $102,93\%$. Essa soma é conhecida como ***booksum***, enquanto a diferença (para $100,00\%$) de $2,93\%$ é conhecida como margem de lucro ou ***overround***.

Para entender essa margem, considere que a Casa de Apostas não está unicamente interessada em acertar o resultado de um evento. Em vez disso, ela espera que as proporções dos volumes recebidos em apostas para cada resultado estejam equilibrados com as probabilidades das *odds* oferecidas. Dessa forma, para o exemplo acima, a Casa de Apostas espera que, para cada $\$102.93$ recebidos em apostas, $\$52.63$ esteja direcionado para a vitória da França, $\$30.30$ para o empate e $\$20$ para a vitória da Croácia. Nesse cenário ideal, ao final do evento, a Casa de Apostas poderia registrar os seguintes resultados:

- Vitória da França: a Casa de Apostas pagaria $\$1.9$ para cada $\$1$ apostado. Sendo assim, como ela recebeu $\$52.63$ em apostas neste resultado, pagaria $\$99.99$ em apostas e ficaria com um lucro de $\$2.94$ do total recebido ($\$102.93$);

- Empate: a Casa de Apostas pagaria $\$3.3$ para cada $\$1$ apostado. Sendo assim, como ela recebeu $\$30.30$ em apostas neste resultado, pagaria $\$99.99$ em apostas e ficaria com um lucro de $\$2.94$ do total recebido ($\$102.93$);

- Vitória da Croácia: a Casa de Apostas pagaria $\$5$ para cada $\$1$ apostado. Sendo assim, como ela recebeu $\$20$ em apostas neste resultado, pagaria $\$100$ em apostas e ficaria com um lucro de $\$2.93$ do total recebido ($\$102.93$).

Percebe-se que, independente do resultado final, o *overround* garante que a Casa de Apostas obtenha lucro em todos os cenários. Caso as proporções recebidas em apostas comecem a ficar desbalanceadas, ou seja, tornando-se diferentes das probabilidades das *odds*, a Casa de Apostas pode reajustar os valores das *odds* de forma que retratem a nova tendência dos apostadores. Esse movimento é conhecido como ***"balanceamento de odds"***.

Em resumo, as Casas de Apostas têm especialistas que realizam a definição das *odds* iniciais para cada evento. A partir do momento que as apostas iniciam, as Casas de Apostas podem usar a tendência dos apostadores para fazer o *balanceamento das odds* e manter sua margem de lucro. Por exemplo, se no cenário apresentado, muitos apostadores começassem a apostar na Croácia, fazendo com que o montante de valor apostado variasse de 20% para 25%, a *odd* da Croácia seria ajustada de 5.00 para 4.00 e, consequentemente, as *odds* de França e Empate também seriam alteradas para se adequar a essa nova realidade. Dessa forma, percebe-se que o balanceamento das *odds* gera um novo efeito: as *odds* das Casas de Apostas podem representar a opinião média do mercado.

2.3.3 Determinando probabilidades a partir das odds

Como observado na seção anterior, para determinar as probabilidades implícitas das *odds* é necessário eliminar o *overround*, de forma que a soma das probabilidades resultem em 100%.

A maioria dos estudos usam **normalização básica**, dividindo a inversa da odds pelo *booksum*. Essa se tornou a forma padrão de se referenciar probabilidades extraídas de odds. Formalmente, considere $o = \{o_1, o_2, o_3\}$ as cotações para os possíveis resultados de uma partida. Considere $k = \{k_1, k_2, k_3\}$ a inversa dessas cotações, ou seja, $k_i = \frac{1}{o_i}$. Considere $\beta = \sum_{i=1}^{3} k_i$, o *booksum*. Dividindo cada k_i pelo *booksum*, $p_i = \frac{k_i}{\beta}$, é possível obter um conjunto de probabilidades em que a soma é exatamente 1 e que pode ser interpretado com as probabilidades para cada resultado.

Apesar do amplo uso e da simplicidade, não é possível garantir que as casas de apostas usam dessa fórmula para distribuir seu *overround*. Smith et al. [37] utilizou um modelo teórico de como as casas de apostas definem suas probabilidades, a partir de uma ideia originalmente proposta por Shin [38]. O **método de Shin** pode ser usado para fazer engenharia reversa das crenças probabilísticas a partir da cotação das *odds*. Ele assume que as casas de apostas determinam as *odds* visando maximizar o lucro esperado na presença de apostadores desinformados e operadores com informações privilegiadas. Os operadores são aqueles que, devido a terem melhores informações, presume-se que "já"sabem o resultado de um determinado evento, antes que o evento ocorra. Formalmente, a contribuição geral no volume global de apostas é quantificada pela porcentagem z. Assim, o modelo de Shin para trabalhar explicitamente a expressão para as probabilidades de apostas, é dado por:

$$\pi(z)_i = \frac{\sqrt{z^2 + 4(1-z)\frac{o_i^2}{\sum_i o_i}} - z}{2(1-z)}$$

A literatura corrente, denota essas probabilidades por *probabilidades de Shin*. Para detalhes matemáticos sobre o método de Shin, leia [39]. Apesar de ser menos utilizado, alguns trabalhos demonstraram que o modelo de Shin é superior a normalização básica.

2.3.4 Casas de Apostas x Bolsas de Apostas

Uma vez compreendidos os fundamentos básicos de uma aposta esportiva, pode-se entender melhor as diferenças entre Casas de Apostas e Bolsa de Apostas. Nas Casas de Apostas tradicionais, o tipo de aposta mais comum é conhecido como *punting*. Essa é a aposta mais popular, em que o apostador espera até o fim do evento para saber se o palpite obteve sucesso. Nas Bolsas de Apostas, além da aposta *punting*, é comum ocorrer o *trading*, modalidade na qual os apostadores não precisam esperar até o fim do evento e podem tentar encerrar suas apostas a qualquer momento.

3 selections	Back all			Lay all		
France	1.95 $25729	1.96 $118868	1.97 $25232	1.98 $69994	1.99 $36490	2 $71006
Croatia	4.8 $25055	4.9 $53171	5 $18527	5.1 $20945	5.2 $37917	5.3 $48704
The Draw	3.25 $171935	3.3 $172956	3.35 $99948	3.4 $172105	3.45 $145101	3.5 $64856

Figura 2.11: *Odds* da Betfair antes da partida final entre França e Croácia (extraída de [4])

Na Figura 2.11, é possível observar as *odds* oferecidas para a final da Copa do Mundo 2018 na plataforma Betfair [4], o maior site de Bolsa de Apostas do mundo. Assim como nas Casas de Apostas, a Bolsa de Apostas tem a opção de apostas "a favor"de um determinado resultado (representados pela coluna "*Back All*"). A diferença, nesse caso, é que na Bolsa de Apostas praticamente não existe *overround*:

- França: $\frac{1}{1.97} \approx 0.5076 \approx 50,76\%$
- Empate: $\frac{1}{3.35} \approx 0.2985 \approx 30,30\%$
- Croácia: $\frac{1}{5.00} = 0.2000 = 20\%$

A soma das três probabilidades acima totaliza o valor de $100,01\%$, ou seja, como os apostadores estão "casando"apostas diretamente (sem *overround*), as cotações das Bolsas de Apostas geralmente são um pouco melhores do que as das Casas de Apostas. Em contrapartida, de todas as apostas vencedoras, a Bolsa de Apostas cobra uma comissão sobre os ganhos líquidos. Para apostadores brasileiros, a comissão varia de 2,6% até 6.5%, dependendo da assiduidade do apostador. Caso a aposta seja perdedora, não é cobrada taxa. Outra

diferença das Bolsas de Apostas é a possibilidade de apostar contra um resultado (a coluna "*Lay*"). Ao apostar contra a França, por exemplo, um apostador terá sua aposta vencedora tanto se der empate como também se a Croácia for vencedora.

Essa dinâmica de apostas "a favor"(*back*) e "contra"(*lay*) é que permite o casamento das apostas entre os apostadores. Na Bolsas de Apostas, só é possível investir a favor de um determinado resultado se, do outro lado, houver um apostador fazendo o mesmo contra o resultado escolhido e vice-versa. Para isso ser possível, a cotação para as apostas *lay* não segue o mesmo padrão da cotação para as apostas *back*.

Por exemplo, se a *odd* da vitória na Croácia determina que, para cada $\$1$ investido, o apostador pode lucrar $\$4$ (recebe $\$5$, mas $\$1$ foi o próprio investimento), do outro lado, um outro apostador precisa apostar $\$4$ contra a Croácia para lucrar $\$1$. Dessa forma, a cotação para *lay* representa o quanto deve ser apostado para alcançar um determinado lucro. No mercado de apostas, esse valor também é conhecido como a "responsabilidade"do apostador. Formalmente, considerando q o valor da *odd lay* e u o lucro desejado, a responsabilidade r do apostador pode ser calculada por:

$$r = (q - 1) * u \tag{2.11}$$

Por exemplo, se um apostador deseja lucrar \$4 apostando contra a Croácia, o valor que ele precisa apostar (sua responsabilidade) seria de $r = (5.1 - 1) * 4 = \$16.40$

A plataforma buscará fazer o casamento dessa aposta contra alguém que apostou \$4 a favor. Caso não encontre imediatamente, a aposta fica aberta até que algum outro apostador deseje fazer o casamento. Essa dinâmica faz com que o mercado de bolsas esportivas se comporte de forma semelhante ao mercado de bolsa de valores. Em resumo, o apostador é quem determina a cotação da *odd*, assim como um corretor da bolsa determina o valor de uma ação a ser vendida. A plataforma fica encarregada de criar uma "fila"de apostadores, priorizando os que fazem a melhor oferta. Quando um outro apostador "casa a aposta", é como se na Bolsa de Valores alguém tivesse comprado a ação ofertada.

Essa dinâmica, semelhante ao mercado financeiro, permite que os apostadores possam encerrar suas apostas a qualquer momento. Esse movimento é conhecido como "*Cash out*". Considere que na Final da Copa do Mundo de 2018, um apostador tenha apostado na França com *odd* de 1.97. Após a França fazer um gol, seu favoritismo aumentou e a *odd* passou

para 1.4. Na plataforma, o apostador pode tentar realizar o *cash-out* e lucrar com a variação negativa da *odd*. É como se uma pessoa comprasse uma ação e ela tivesse valorizado com o tempo. A plataforma computa automaticamente qual o lucro que o apostador obteve com a variação e, ao encerrar a aposta, o apostador já garantiu o lucro no final do evento, mesmo que a França depois não vença a partida. Por outro lado, se a França iniciar a partida perdendo e sua *odd* aumentar, o apostador tem a opção de encerrar a aposta para tentar diminuir seu prejuízo.

Para melhor entendimento dessa dinâmica, considere o confronto entre Arsenal e Leicester, ocorrido em 10 de fevereiro de 2016, em jogo válido pela *Premier League* [40]. A partida teve os seguintes acontecimentos, em ordem cronológica:

- A partida teve início às 12h;
- Aos 45' do 1º tempo, o Leicester abriu o placar (1-0);
- O intervalo durou cerca de 15min e o segundo tempo começou às 13h05;
- Aos 10' do 2º tempo, o Leicester teve um jogador expulso;
- Aos 25' do 2º tempo, o Arsenal empatou o jogo (1-1);
- Aos 49' do 2º tempo, o Arsenal virou a partida (2-1).

A partir da Figura 2.12, pode-se relacionar os acontecimentos da partida às probabilidades implícitas das *odds* negociadas pelos apostadores na Betfair. Denotando as probabilidades implícitas simplesmente como "chances", pode-se alcançar as seguintes conclusões:

- Antes mesmo do jogo iniciar, as chances do Arsenal subiram em torno de 3%, saindo de 55% para aproximadamente 58%. Essa subida pode ser considerada como um sentimento subjetivo do mercado nas análises pré-jogo;
- No decorrer do primeiro tempo, as chances do Leicester tiveram oscilações brandas, mantendo-se próximas dos 20%. Entretanto, o gol marcado no final do 1º tempo fez as chances do Leicester subirem abruptamente para aproximadamente 45%, enquanto que, para o Arsenal, o gol sofrido fez com que as chances que estavam próximas dos 51% caíssem para 25%;

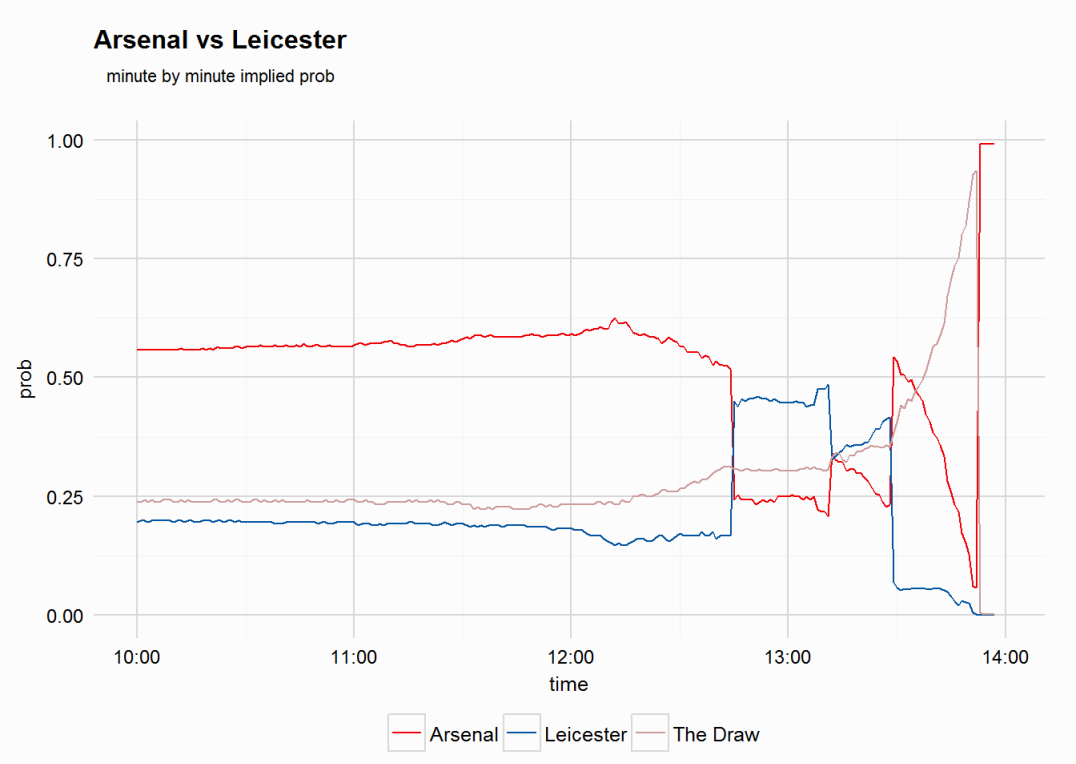


Figura 2.12: Probabilidades implícitas das *Odds* da Betfair para a partida entre Arsenal e Leicester válida pela Premier League 2016/2017 (extraída de [5])

- Durante o intervalo, houve pouca oscilação entre as probabilidades, indicando que não houve informações relevantes para alterar a percepção dos apostadores em relação às chances dos times;

- Após o início do 2º tempo, as chances do Leicester começaram a subir à medida que o tempo passava. Entretanto, aos 10', a expulsão de um jogador do Leicester voltou a equilibrar as chances. Naquele momento, todos os resultados passaram a ter uma probabilidade próxima dos $33,3\%$;

- Entre os 10' e os 25' do 2º tempo, o Leicester continuou jogando "a favor do tempo". Porém, com o gol de empate marcado pelo Arsenal aos 25', o cenário foi reconfigurado. O Arsenal voltou a ter 51% de chance de vitória, enquanto as chances do Leicester que eram de 40% caíram para próximo dos 10%, visto que nesse momento o Leicester já estava com um jogador a menos;

- Dos 25' até o último minuto de jogo, o Arsenal jogou "contra o tempo". À medida

que o tempo passava, suas chances de vitória iam reduzindo, enquanto as chances de empate aumentavam;

- O gol da vitória no último minuto, praticamente mudou e definiu as chances para todos os resultados possíveis do jogo. O Arsenal, que naquele momento tinha em torno de 10% de chances, passou a ter 99%. O Leicester que tinha aproximadamente 2% de chances caiu para praticamente 0%. Enquanto que o empate, que já estava quase garantido, passou de 90% de chances para aproximadamente 1%.

2.3.5 Tipos de Apostas

Atualmente, para o mercado de apostas em futebol, existem diversas opções de tipos de apostas. Os tipos que tem maior liquidez e que são citados neste trabalho são:

1. **Money Line** - Também conhecido como aposta simples, 1x2, ML, *odds result*. É um tipo de aposta de três vias no futebol. Nesse caso, o apostador deve escolher o resultado favorito, ou seja, vitória do time da casa, empate ou vitória do time visitante.

2. **Ambos Marcam** - Também conhecido como *both teams to score* ou BTTS. Nesse caso, o apostador faz uma escolha binária. Se ele acredita que ambas as equipes vão fazer pelo menos um gol na partida, então escolhe a opção "SIM"; caso contrário, escolhe a opção "NÃO".

3. **Gols Acima/Abaixo** - Também conhecido como mais/menos ou *over/under*. Esse é o tipo de aposta relacionado ao número total de gols de uma partida. A casa de apostas apresenta opções de apostas para diferentes limiares, como 0.5, 1.5, 2.5, etc. Nesse caso, o participante pode apostar se o número total de gols da partida vai ser maior ou menor que um determinado limiar. Por exemplo, se ele acredita que uma partida vai ter pelo menos 3 gols, então ele pode escolher a opção "over 2.5". Apesar de ser mais popular para "gols", também há mercado de over/under para outros eventos do jogo, como cartões e escanteios.

2.4 Considerações Finais

Neste capítulo, foram apresentados conceitos importantes sobre mineração de dados, aprendizagem de máquina e apostas esportivas. As definições das técnicas de mineração de dados e aprendizagem de máquina são importantes para uma melhor compreensão de como funcionam as técnicas experimentadas neste trabalho. Por sua vez, os conceitos sobre apostas esportivas são fundamentais para entender a natureza de alguns dos conjuntos de dados usados para modelagem e avaliação de modelos preditivos em futebol. No próximo capítulo, serão apresentados os trabalhos relacionados que foram referências para o desenvolvimento desta proposta de tese.

Capítulo 3

Trabalhos Relacionados

Neste capítulo, serão apresentados os trabalhos relacionados a este trabalho. Primeiramente, é feita uma caracterização dos trabalhos que realizam, especificamente, *modelagem preditiva de resultados de futebol*. Em seguida, são discutidos os trabalhos relacionados à *avaliação do mercado de apostas esportivas*.

3.1 Modelos Preditivos de Resultados de Futebol

Na literatura, há uma grande diversidade de trabalhos que envolvem futebol e predição. Este trabalho realizou uma revisão sistemática para estudar as principais pesquisas relacionadas ao domínio. Primeiramente, foi realizada uma busca avançada por palavras-chave na plataforma SCOPUS (Apêndice A.1). Tal busca produziu uma lista contendo 805 trabalhos. Em uma primeira análise, foram considerados o título e resumo dos trabalhos listados. Essa ação resultou na pré-seleção de 190 trabalhos. Os demais 615 foram desconsiderados por não terem relação com o escopo deste estudo.

Em uma segunda etapa, foram analisadas as seções de introdução e conclusão dos 190 trabalhos pré-selecionados. Esses estudos fazem parte da revisão desta pesquisa. Entretanto, 147 deles não serão utilizados para análise comparativa, pois não estão diretamente relacionados com a predição de resultados em futebol. Dentre os objetivos dos estudos eliminados nessa etapa, destacam-se:

- A modelagem preditiva para outros esportes;

- A modelagem para avaliação de desempenho de atletas;
- A modelagem para criação de *rankings* de times ou previsão da classificação final de um torneio;
- A modelagem para estudo de fatores específicos do jogo, como passes, chutes e posse de bola;
- A modelagem para a influência de determinados fatores sobre o resultado, como vantagem do time da casa, previsão do tempo e distância de viagens;
- A modelagem para avaliar a parte tática.

Por fim, analisou-se os 43 trabalhos selecionados. A partir da análise das referências desses trabalhos, outros 4 (quatro) foram adicionados à revisão, pois não estavam indexados na plataforma Scopus.

Dessa forma, nesta seção serão apresentadas as características de 47 (quarenta e sete) trabalhos selecionados, destacando seus pontos fortes e suas limitações. Esta caracterização será dividida em três dimensões: *"conjunto de dados utilizado"*, *"características da modelagem preditiva"* e *"formas de avaliação do modelo"*. Para cada dimensão, será apresentado um quadro comparativo, além de uma análise crítica do estado da arte. Ao final, será comparada a proposta deste trabalho com o estado da arte.

3.1.1 Conjuntos de Dados

Considerando a forte natureza estocástica do jogo, os resultados das partidas de futebol muitas vezes não refletem a diferença de qualidade entre as equipes. Dessa forma, alguns trabalhos têm explorado formas de selecionar e extrair variáveis relevantes para predição de resultados de futebol. Esta modelagem preditiva tem sido realizada a partir de dados de diferentes características. Este trabalho, esses conjuntos de dados estão classificados como:

- **Resultados de Partidas**: nome das equipes, quantidade de gols marcados pelas mesmas, identificação do campeonato, fase do campeonato e data da partida (Ex. [14; 15]);

- **Estatísticas (*Scouts*)**: principais estatísticas detalhadas de uma partida, como tempo de posse de bola, número de faltas, cartões, finalizações, escanteios, etc. (Ex. [41; 42]);
- **Resumos de Partidas**: dados sobre tempo dos gols, cartões, substituições, etc. (Ex. [43]);
- **Rankings/Ratings**: pontuações ou colocações das equipes em rankings populares como, por exemplo, Ranking da FIFA [6] ou Elo Rating [44] (Ex. [45; 46]);
- **Cotações de Casas de Apostas**: cotações no mercado de apostas antes e durante os jogos (Ex. [47; 48]);
- **Postagens em Mídias Sociais**: postagens de torcedores coletadas em redes sociais como o *Twitter* (Ex. [49; 50]);
- **Outros Dados**: dados diversos, como distância entre as cidades das equipes (Ex. [20]), avaliação de jogadores em *jogos de vídeo game* (Ex. [51]), número de jogadores lesionados (Ex.[52]), etc.

Os dados coletados, muitas vezes, passam por um pré-processamento, que envolve a extração de atributos, antes de serem utilizados na aprendizagem de máquina. Há uma variedade de estratégias nessa etapa, mas de uma forma geral os atributos derivados desse processo podem ser classificados em dois grupos: *atributos de desempenho* e *atributos de força*.

Os *atributos de desempenho* são geralmente extraídos para representar informações referentes à tabela de classificação dos campeonatos. A partir dos resultados das partidas, calcula-se, por exemplo, o número de pontos, gols marcados, gols sofridos, vitórias, empates e derrotas de cada time, a cada rodada. Para isso, alguns trabalhos, como em Tüfekci [53], usam os dados de todo o campeonato, enquanto outros, como em Hvattum [54], usam apenas dados de rodadas mais recentes. Além desses, há trabalhos que utilizam as duas estratégias ao mesmo tempo, como em Hubáček et al. [15].

Os *atributos de força* são gerados por modelos que transformam diferentes variáveis de entrada em um único valor. Esse valor representa a "força"de uma equipe. Alguns trabalhos, como Baio & Blangiardo [55], adotam a estratégia de separar esse atributo em dois:

"força ofensiva"e "força defensiva". Outros, como em Hubáček et al. [15], decompõem esse atributo em "força em casa"e "força como visitante".

Existem também *outros atributos* com características diferentes das citadas. Por exemplo, Tsakonas et al. [56] verifica o histórico do *confronto direto* entre duas equipes. Constantinou & Fenton [35] estima a *fadiga* de uma equipe a partir do intervalo de dias entre os jogos. Hubáček et al. [15] mensura a *importância da partida* a partir do posicionamento das equipes em um campeonato, entre outras estratégias pontuais.

Por fim, outra característica que também tem apresentado bastante variação é o tipo de campeonato utilizado nos estudos. Alguns trabalhos usam dados de ligas nacionais, como as da Inglaterra [20], Itália [55], Finlândia [57], Brasil [58], Turquia [53], Holanda [59] e Irã (Ex [60]), enquanto outros usam de campeonatos internacionais como Liga dos Campeões [52], UEFA Euro [61] e Copa do Mundo [48].

Entendidas as características dos conjuntos de dados usados, pode-se fazer uma análise comparativa dos estudos. A Tabela 3.1 caracteriza os trabalhos relacionados de acordo com os tipos de *dados* utilizados, os tipos de *atributos* modelados e os *campeonatos* usados como fonte.

Ao analisar a Tabela 3.1, percebe-se que todos os estudos utilizam dados de *resultados de partidas*. Muitas vezes, esses dados são usados para derivar atributos de *desempenho* e de *força*. Dentre as abordagens utilizadas para modelagem de *atributos de força* destacam-se: *Elo Rating*, *Pi-Rating* e *PageRank*.

O *Elo-Rating* foi criado inicialmente para avaliar jogadores de xadrez, mas posteriormente foi adaptado para medições de força no futebol. O *Pi-Rating* foi desenvolvido por Constantinou [21] e tem sido utilizado por trabalhos recentes (Ex. [15], [14]). O *PageRank* é um algoritmo utilizado, originalmente, para posicionar *websites* entre os resultados de busca. Lazova e Basnarkov [62] fizeram uma adaptação para usar o *PageRank* na classificação de força de times de futebol.

Considerando a predição pré-jogo, os resultados históricos são, com certeza, uma relevante fonte de informação para modelagem das forças das equipes. Entretanto, não é raro que um jogo não transcorra como previsto anteriormente. Nesse contexto, pode-se considerar que uma alternativa para refinar essa medição é a utilização das estatísticas do jogo (*scouts*). As estatísticas podem acrescentar detalhes não capturados pelos resultados das

partidas. Como observado, há poucos trabalhos explorando esse tipo de informação. Este trabalho pretende fazer uso desse conjunto de dados para o ajuste das predições durante as partidas. Além disso, a ampla maioria dos trabalhos não disponibilizam os conjunto de dados e os modelos para reprodução. Este trabalho, disponibiliza não só o conjunto de dados, como também todo o código do processo de *engenharia de atributos* que transforma dados em atributos.

Alguns trabalhos, como Constantinou [14], tem utilizado as cotações do mercado de apostas para avaliação dos modelos. Outros, como Odachowski & Grekow [63], usam essas cotações como dados de entrada para a própria modelagem preditiva. Como visto na Seção 2.3.1, essas cotações, na verdade, já representam predições para as partidas. Sendo assim, cada casa de apostas acaba sendo uma "especialista"do domínio e as técnicas de modelagem buscam aprender com elas. Essa é uma fonte de dados muito rica, dado que as cotações são supostamente baseadas em todas as informações disponíveis para o jogo. Este trabalho faz uso desse conjunto de dados tanto para modelagem preditiva, quanto para avaliação de modelos.

Por fim, pode ser observado que a grande maioria dos trabalhos utilizam dados de um único campeonato. Constantinou [14] mostram que essa estratégia limita a generalização do modelo, pois um modelo preditivo para um campeonato pode não ser adequado para outro. Este trabalho segue a estratégia de estudos mais recentes que têm buscado usar dados de vários campeonatos para ter uma melhor capacidade de generalização. Na *predição pré-jogo*, são usados dados de 9 (nove) ligas nacionais importantes, enquanto na *predição durante o jogo*, são considerados dados de mais de 50 campeonatos diferentes.

3.1.2 Modelos Preditivos

A predição de resultados de futebol (pré-jogo) é o objetivo comum da ampla maioria dos trabalhos relacionados. Entretanto, com o crescimento dos mercados de apostas esportivas, questões como *"Qual será o total de gols da partida?"*, ‘‘*Quantos escanteios acontecerão em um jogo?"*, *"Ambas as equipes vão marcar gols?"*, entre outras, passaram a ter semelhante relevância. Nesta seção, questões desse tipo serão denotadas como *predição para "outros alvos"*.

Além da tradicional predição pré-jogo, o mercado de apostas também está suscitando

um crescente interesse pelas predições durante as partidas (em tempo real). Esse tipo de predição visa "atualizar"as chances de cada equipe até o final da partida, à medida que novos eventos acontecem durante o decorrer da disputa. Neste trabalho, esse tipo de predição será identificado como *predição em tempo real*.

Para a modelagem preditiva tanto no pré-jogo, como em tempo real, diversas técnicas envolvendo estatística e aprendizagem de máquina são utilizadas. Essas técnicas podem ser classificadas como:

1. Regressão
    (a) Regressão Logística (Ex. [51])
    (b) Regressão Poisson (Ex [64]))
2. Abordagens Bayesianas
    (a) Naive Bayes (Ex. [65])
    (b) Redes Bayesianas (Ex. [14])
3. KNN (*K-Nearest Neighbors*) (Ex. [66])
4. Árvores de Decisão (Ex. [67])
5. Máquinas de Vetor de Suporte (SVM) (Ex. [68])
6. Redes Neurais (Ex. [41])
7. Algoritmos Genéticos (Ex. [69])
8. Combinação de Modelos (*Ensemble Methods*)
    (a) Floresta Randômica (Ex. [70])
    (b) *Gradient Boosting* (Ex. [15])
9. Outras Técnicas
    (a) Lógica Fuzzy (Ex. [57])
    (b) Raciocínio Baseado em Regras (Ex. [11]).

(c) Aprendizagem Preguiçosa (Ex. [42])

A Tabela 3.1 caracteriza os trabalhos relacionados de acordo com os tipos de *técnicas* utilizadas e os objetivos de *predição*. Percebe-se que a ampla maioria dos trabalhos realizam apenas predição pré-jogo. Alguns trabalhos que realizam predição durante o jogo, como [50; 49] são estudos baseados em *análise de sentimento* das redes sociais. Este trabalho pretende adotar uma estratégia semelhante a de [71], fazendo ajustes nas predições durante o jogo, baseados nos eventos que acontecem durante o jogo, como gols, cartões, ataques, chutes a gols, escanteios, etc.

Em relação à predição pré-jogo, o número de estudos para predições para "outros alvos"também é reduzido. Gomes et al. [42] apresentou um modelo para predição da quantidade de gols e de escanteios em uma partida. Boshnakov et al. [72] criou um modelo para avaliar se as partidas teriam mais ou menos que 2.5 gols. Owen [73] buscou prever se uma partida terminaria empatada sem gols. O presente trabalho além de focar na predição de resultados, pretende abordar um "outro alvo". Nesse caso, serão realizadas predições para a seguinte pergunta: "Ambas as equipes irão marcar?". Essa questão foi escolhida devido a sua popularidade no mercado de apostas esportivas e por não ter sido explorada por nenhum outro trabalho até o momento.

Na questão das técnicas de modelagem preditiva, percebe-se uma grande variação nas técnicas para modelagem de predição. Alguns trabalhos apresentam uma única técnica, como [68; 69; 60], enquanto outros privilegiam a comparação entre algumas delas, como [63; 53].

Neste trabalho, são aplicadas técnicas de aprendizagem de máquina que apresentaram resultados promissores em trabalhos anteriores, como *Gradient Boosting*, *Naive Bayes* e *Regressão Logística* para analisar qual(is) delas é(são) mais adequada(s) aos problemas propostos. Nesse caso específico, os classificadores baseados em Poisson são reconhecidamente bons baselines. Assim, os trabalhos de Maher [74], Dixon & Coles [75] e Rue & Salvesen [76] são usados como baselines.

No caso da predição em tempo real, a frequência acumulada de eventos ocorridos nas partidas (gols, chutes, escanteios, etc.) e as chances de cada equipe durante o jogo podem ser vistas como múltiplas séries temporais. Dessa forma, além de técnicas de aprendizagem de

máquina tradicionais, também são investigadas técnicas de redes neurais complexas, como LSTM, FCN e InceptionTime, apropriadas para classificação de séries temporais.

3.1.3 Avaliação dos Modelos

A grande maioria dos estudos avaliam seus modelos em termos de acurácia ou RPS - Ranked Probability Score, comparando suas predições com os resultados reais. Apesar desse procedimento padrão ser importante, as análises desses trabalhos não mostram se a técnica desenvolvida é superior a outras existentes.

É bem verdade que a comparação entre trabalhos é muitas vezes inviável, pois cada trabalho utiliza campeonatos e temporadas diferentes para predição. Constantinou [14] mostrou que não seria justo fazer comparações desse tipo, dado que cada campeonato tem um certo grau de preditibilidade. Dessa forma, trabalhos como o de Martins [43] que compara a acurácia do seu classificador com outros modelos, gerados a partir de dados de campeonatos distintos, têm relevância limitada.

Nesse cenário, é importante que os trabalhos façam comparações entre as técnicas a partir de um mesmo ambiente (conjunto de treino e teste), como feito, por exemplo, em Ulmer [77], Hucaljuk & Rakipović [52] e Joseph [78].

Além da comparação entre modelos, outra forma de avaliação importante é a comparação com o mercado de apostas. Hubáček [15] define as casas de apostas como um "limite superior"para a qualidade de predição, dado que o mercado tem diversos tipos de informação disponível sobre as partidas, como escalações, desempenho, importância da partida, moral das equipes no momento, entre outros. Dessa forma, um modelo que ultrapasse esse limite poderá ser capaz de gerar lucros sistemáticos no mercado de apostas.

Nesse cenário, comparar os modelos preditivos com as predições dos mercados de apostas pode indicar até que ponto o modelo é realmente útil, como também pode permitir a identificação de possíveis ineficiências do mercado. Trabalhos como Constantinou [14] e Hubáček [15] realizaram comparações com os mercados de apostas para avaliação de seus modelos.

Assim, a abordagem proposta neste trabalho faz comparações entre diferentes técnicas de modelagem sobre um mesmo ambiente (conjunto de treino e teste) e compara as pre-

dições desses modelos com as predições do mercado de apostas em termos de acurácia e lucratividade.

Na Tabela 3.2, pode-se observar quais estratégias de avaliação são adotadas pelos trabalhos relacionados. São raros os trabalhos que fazem comparações com outros existentes, prevalecendo assim a comparação entre técnicas dentro do mesmo trabalho. Percebe-se também uma preocupação dos trabalhos mais recentes em comparar seus modelos com as predições do mercado. Há ainda, trabalhos que não fizeram nenhum tipo de comparação com qualquer *baseline*, avaliando seu modelo apenas de forma argumentativa.

3.2 Avaliação do Mercado de Apostas Esportivas

O estudo de eficiência de mercados é bastante comum na área econômica. Por definição geral, a *hipótese do mercado eficiente* afirma que o mercado é eficiente se uma pessoa (agente) não consegue alcançar continuamente retornos maiores que a média do mercado, considerando as informações públicas disponíveis no momento em que os investimentos são realizados [22]. Analogamente, no âmbito das apostas esportivas, pode-se inferir que, se um apostador consegue ter lucros consistentes com as cotações oferecidas pelo mercado, então o mercado não é eficiente. Em outras palavras, sabe-se que as cotações disponibilizadas pelo mercado carregam as probabilidades implícitas para um evento, porém, em um cenário de mercado ineficiente, essas probabilidades não estariam representando adequadamente as chances das equipes nas partidas.

Para os economistas, os mercados de apostas esportivas são considerados, por diversas razões, um excelente domínio para a realização de testes de eficiência de mercado. Primeiramente, assim como os mercados financeiros, os de apostas esportivas também são caracterizados pela incerteza e pelo elevado número de agentes com diferentes níveis de conhecimento, com alto poder financeiro e com acesso a informações abundantes. Em segundo lugar, os mercados de apostas têm algumas características particulares que facilitam a avaliação de sua eficiência: a chegada da informação é extremamente clara e recebida por todos os agentes ao mesmo tempo, além dos resultados finais serem conhecidos em um limitado e curto espaço de tempo [96].

Ainda não há um consenso sobre a eficiência do mercado de apostas esportivas [96].

		Dados							Atributos			
Ref.	Trabalhos	Resultados	Estatísticas	Resumo	Rankings/Rating	Cotações	Mídias Sociais	Outros Dados	Desempenho	Força	Outros Atributos	Campeonatos
[76]	Rue e Salvesen, 2002	X				X				X		Inglês
[56]	Tsakonas et al., 2002	X							X	X	X	Ucraniano
[79]	Cheng et al., 2002	X							X	X	X	Italiano
[64]	Crowder et al., 2002	X								X		Inglês
[80]	Goddard e Asimakopoulos, 2004	X				X		X	X	X	X	Inglês
[81]	Forrest et al., 2005	X				X		X	X	X	X	Inglês
[57]	Rotshtein et al., 2005	X							X			Finladês
[82]	Goddard, 2005	X				X			X	X	X	Inglês
[78]	Joseph et al., 2006	X						X	X	X	X	Inglês
[83]	Aslan e Inceoglu, 2007	X							X	X	X	Italiano
[84]	Min, 2008	X						X	X	X	X	Copa do Mundo
[55]	Baio e Blangiardo, 2010	X								X		Italiano
[45]	Leitner et al., 2010	X			X					X		UEFA Euro
[20]	Hvattum et al., 2010	X			X	X				X		Inglês
[41]	Huang e Chang, 2010	X	X									Copa do Mundo
[58]	Alves et al., 2011	X								X		Brasileiro
[52]	Hucaljuk e Rakipović, 2011	X							X	X	X	Champions League
[35]	Constantinou, 2012	X				X		X	X	X	X	Inglês
[63]	Odachowski e Grekow, 2012	X				X						Indefinido/Variado
[69]	Cui et al., 2013	X				X			X	X	X	Inglês
[85]	Constantinou et al., 2013	X				X		X	X	X	X	Inglês
[60]	Arabzad et al., 2014	X							X		X	Iraniano
[68]	Igiri, 2015	X										Inglês
[49]	Le e Flammini, 2015	X				X	X				X	Indefinido/Variado
[48]	Dobravec, 2015	X				X				X		Copa do Mundo
[86]	Kopman e Lit, 2015	X				X				X	X	Inglês
[42]	Gomes et al., 2016	X	X			X			X	X		Indefinido/Variado
[53]	Tüfekci, 2016	X								X	X	Turco
[51]	Prasetio e Harlili, 2017	X	X					X		X		Inglês
[54]	Hvattum, 2017	X				X		X		X	X	4 Ligas Nacionais
[43]	Martins et al., 2017	X	X	X								4 Ligas Nacionais
[72]	Boshnakov et al., 2017	X	X			X			X	X		Inglês
[73]	Owen, 2017	X	X			X				X		Escocês
[14]	Constantinou, 2018	X				X			X	X	X	52 Ligas Nacionais
[15]	Hubáček et al., 2018	X				X			X	X	X	52 Ligas Nacionais
[87]	Cho et al., 2018	X						X	X			Champions League
[66]	Esme e Kiran, 2018	X				X		X	X	X	X	Turco
[50]	Brown et al., 2018	X				X	X					Inglês
[88]	Hervert-Escobar et al., 2018	X							X	X		Indefinido/Variado
[89]	Robberechts e Davis, 2018	X				X			X			Copa do Mundo
[90]	Schauberger e Groll, 2018	X				X		X	X	X	X	Copa do Mundo
[91]	Egidi et al., 2018	X				X			X		X	4 Ligas Nacionais
[92]	Koopman e Lit, 2019	X				X			X			10 Ligas Nacionais
[71]	Robberechts et al., 2019	X	X						X		X	20 Ligas Nacionais
[93]	Baboota e Kaur, 2019	X	X					X	X	X	X	Inglês
[94]	Stübinger et al, 2020	X						X			X	15 Ligas Nacionais
[95]	Arntze e Hvattum, 2020	X								X		Inglês
	Este trabalho	X	X	X		X			X			Variado

Tabela 3.1: Tabela Comparativa de Trabalhos Relacionados (Conjunto de Dados)

		Técnicas									Predição			Avaliação		
Ref.	Trabalhos	Regressão	Ab. Bayesianas	KNN	Árvores de Decisão	SVM	Redes Neurais	Alg. Genéticos	Ensemble	Outros	Resultados	Outros Alvos	Em Tempo Real	Entre Trabalhos	Entre Técnicas	Mercado de Apostas
[76]	Rue e Salvesen, 2002		X								X					X
[56]	Tsakonas et al., 2002						X	X		X	X				X	
[79]	Cheng et al., 2002						X			X	X				X	
[64]	Crowder et al., 2002	X	X								X			X		
[80]	Goddard e Asimakopoulos, 2004	X									X					X
[81]	Forrest et al., 2005	X									X					X
[57]	Rotshtein et al., 2005	X					X	X			X				X	
[82]	Goddard, 2005	X									X				X	X
[78]	Joseph et al., 2006		X	X	X						X				X	
[83]	Aslan e Inceoglu, 2007						X			X	X			X	X	
[84]	Min, 2008		X							X	X		X			
[55]	Baio e Blangiardo, 2010		X								X					
[45]	Leitner et al., 2010	X								X	X				X	
[20]	Hvattum et al., 2010	X								X	X					X
[41]	Huang e Chang, 2010						X				X					
[58]	Alves et al., 2011	X									X					
[52]	Hucaljuk e Rakipović, 2011		X	X			X		X	X	X				X	
[35]	Constantinou, 2012		X								X					X
[63]	Odachowski e Grekow, 2012		X		X					X	X					X
[69]	Cui et al., 2013						X	X			X				X	
[85]	Constantinou et al., 2013		X								X					X
[60]	Arabzad et al., 2014						X				X					
[68]	Igiri, 2015					X					X					
[49]	Le e Flammini, 2015		X							X	X		X			X
[48]	Dobravec, 2015		X							X	X				X	X
[86]	Kopman e Lit, 2015	X														X
[42]	Gomes et al., 2016		X		X	X				X		X			X	X
[53]	Tüfekci, 2016				X		X		X		X				X	
[51]	Prasetio e Harlili, 2017	X									X					
[54]	Hvattum, 2017	X										X			X	X
[43]	Martins et al., 2017		X		X	X	X			X	X			X	X	
[72]	Boshnakov et al., 2017	X										X				X
[73]	Owen, 2017		X									X				X
[14]	Constantinou, 2018		X								X					X
[15]	Hubáček et al., 2018								X	X	X				X	
[87]	Cho et al., 2018				X	X	X		X	X	X				X	
[66]	Esme e Kiran, 2018			X							X	X				X
[50]	Brown et al., 2018	X									X		X			X
[88]	Hervert-Escobar et al., 2018		X								X					
[89]	Robberechts e Davis, 2018	X							X	X	X				X	X
[90]	Schauberger e Groll, 2018	X							X		X				X	X
[91]	Egidi et al., 2018	X								X	X				X	X
[92]	Koopman e Lit, 2019	X									X				X	X
[71]	Robberechts et al., 2019	X	X						X				X		X	X
[93]	Baboota e Kaur, 2019		X						X		X				X	X
[94]	Stübinger et al, 2020	X				X			X	X	X				X	X
[95]	Arntze e Hvattum, 2020	X								X	X		X		X	X
	Este trabalho		X			X	X		X		X	X	X		X	X

Tabela 3.2: Tabela Comparativa de Trabalhos Relacionados (Predição e Avaliação)

Na predição pré-jogo, especificamente, trabalhos como Spann & Skiera [97] concluem que o mercado apresenta elevado grau de eficiência, a partir da análise de 837 jogos do Campeonato Alemão. Por outro lado, Kaunitz et al. [98] define o mercado como ineficiente, apresentando uma estratégia lucrativa que utiliza as divergências entre as próprias casas de apostas. Deutscher et al. [99] apresentou uma estratégia lucrativa, analisando o começo das temporadas de equipes recém-promovidas no Campeonato Alemão. A estratégia utilizada sugere que as casas de apostas subestimam a capacidade dessas novas equipes no início das temporadas. Wheatcroft [100] apresentou uma estratégia para o mercado over/under, que obteve um lucro médio de cerca de 0,8 por cento por aposta feita ao longo de doze anos, usando estatísticas de chutes a gols e escanteios (e não gols) como entradas.

Este trabalho, pretende analisar a eficiência do mercado, sob duas perspectivas, denotadas por Lobão & Rolla [96] como: eficiência estatística e eficiência econômica. Na *eficiência estatística*, pretende-se observar até que ponto as cotações do mercado constituem boas previsões para os resultados finais. Essa avaliação pode ser feita comparando a predição do mercado de apostas com as predições dos modelos preditivos. Na *eficiência econômica*, pretende-se identificar se o comportamento histórico das cotações apresenta alguma ineficiência, ou seja, se existem estratégias que conseguem gerar rentabilidades sistemáticas positivas.

Dessa forma, pode-se discutir as implicações deste trabalho sob duas perspectivas. A primeira é a perspectiva estatística. Se o mercado for eficiente estatisticamente, os modelos criados a partir de dados históricos das partidas não conseguirão superar as predições do mercado. A segunda é a perspectiva econômica. Se o mercado for eficiente economicamente, não é possível criar nenhuma estratégia lucrativa, com o auxílio dos modelos criados.

Essa segunda perspectiva tem um foco específico para o comportamento das cotações. Alguns trabalhos estão voltados para analisar como o mercado se ajusta às informações disponíveis sobre o jogo, discutindo a eficiência do mercado. Essas análises são tradicionalmente feitas para as predições antes do jogo, mas, com a recente popularização das bolsas de apostas, novos estudos estão surgindo para avaliar a eficiência do mercado durante as partidas.

Croxson & Reade [101], por exemplo, usaram dados das cotações da *Betfair* para testar a hipótese de eficiência. Esse trabalho foca nos gols marcados nos últimos cinco minutos

antes do fim do primeiro tempo. Dado que no intervalo de uma partida não há nenhum evento significante, esse seria o cenário ideal para verificar se o mercado levaria tempo para se ajustar a informação do gol. O estudo concluiu que o mercado reage de forma rápida e eficiente.

Dados da *Betfair* também foram usados por Choi & Hui [67] para testar a hipótese de que os apostadores tendem a reagir de forma moderada às informações que são esperadas e de forma exagerada às informações surpreendentes. Nesse trabalho, os autores analisaram o ajuste das cotações após o primeiro gol marcado em cada partida. O nível de "surpresa"é dado pela diferença entre as chances das duas equipes na partida. Assim, observando as cotações dois minutos após cada gol marcado, verificou-se que, quando o gol era "esperado", o número esperado de vitórias para os favoritos é maior do que o número observado; ao mesmo tempo que o número de vitórias esperado foi menor que o número observado, quando o primeiro gol foi uma surpresa.

Richard & Vecer [102] também aponta ineficiência do mercado. Eles avaliam especificamente estratégias de apostas durante partidas da Premier League. Eles demonstram como a maximização de uma função logaritmica pode ser útil para realização de *trades* durante as partidas, explorando potenciais discordâncias de opiniões do mercado.

Nesse contexto, este trabalho avalia, especificamente, se técnicas de aprendizagem são viáveis para aprender o comportamento das cotações das casas de apostas de acordo com os eventos ocorridos em uma partida. Encontrar reações inadequadas do mercado são gatilhos para inferir ineficiência de mercado e consequentemente pode servir de base para a adoção de estratégias lucrativas.

3.3 Considerações Finais

Neste capítulo foram discutidos os trabalhos relacionados a proposta deste trabalho. Foram apresentados os pontos fortes e as limitações da literatura, a partir de uma revisão sistemática contendo 47 trabalhos. Por fim, avaliou-se como a modelagem preditiva tem impactado os trabalhos sobre eficiência de mercado. Como resultado, foi detectada a carência de trabalhos na predição em tempo real. No próximo capítulo, serão apresentadas as etapas do processo de modelagem executadas por este trabalho.

Capítulo 4

Predição de Ambos Marcam

Neste capítulo, será detalhado o método proposto por este trabalho para prever, antes da partida iniciar, se ambas as equipes marcarão gols. No mercado de apostas, esse tipo de predição também é conhecido como *"ambas marcam"*. A predição de *ambas marcam* corresponde a uma classificação binária. Sendo assim, a variável-alvo pode assumir dois valores: *sim*, caso ambas as equipes marquem gols na partida, ou *não*, caso contŕario.

O capítulo está dividido como segue. Primeiramente, são apresentados os conjuntos de dados coletados para o problema. Em seguida, são descritas as etapas de pré-processamento de dados (*feature engineering*) e apresentado o conjunto de dados final usado para a modelagem de *predição de ambas marcam*. A seguir, são detalhados os classificadores e os experimentos executados, como também discutidos os resultados obtidos. Por fim, são apresentadas as conclusões e considerações finais sobre o capítulo.

4.1 Coleta de Dados

O conjunto de dados desta pesquisa é formado por dados coletados no *BetExplorer.com* [103], site que indexa as *odds* das principais casas de apostas, referentes às partidas de diversos campeonatos.

Foram coletados dados de partidas de 6 (seis) temporadas (2013-2019), de 9 (nove) ligas nacionais importantes mundialmente. Todas as ligas possuem formatos de disputa idênticos, nos quais as equipes se enfrentam em dois turnos (jogos de ida e volta). São elas: *Premier League* (Inglaterra A), *La Liga* (Espanha A), *Serie A* (Itália A), *Bundesliga A* (Alemanha A),

Ligue 1 (França A), *Primeira Liga* (Portugal A), *Eredivisie A* (Holanda A), *Brasileirão Série A* (Brasil A) e *Brasileirão Série B* (Brasil B).

Para cada campeonato, foram coletados os seguintes dados:

- **Partidas disputadas**: identificação do time da casa, identificação do time visitante, gols marcados pelo time da casa, gols marcados pelo time visitante, data do jogo, ano da temporada e nome do campeonato;
- **Cotações das casas de apostas para as partidas**: cotações de 19 casas de apostas para o mercado *Money Line* (Resultados de partidas) e *Both Teams to Score* (Ambas Marcam);

A Tabela 4.1 apresenta algumas estatísticas sobre o conjunto de dados coletado.

Campeonato	# Jogos	# Odds BTTS	# Odds ML
Brasil A	2.280	32.043	46.769
Brasil B	2.280	29.215	46.587
Portugal A	1.770	26.154	36.500
França A	2.280	34.406	47.754
Inglaterra A	2.280	36.130	47.938
Espanha A	2.280	34.618	47.806
Itália A	2.280	34.556	47.832
Alemanha A	1.836	27.914	38.484
Holanda A	1.836	26.647	37.654
Total	19.122	281.683	39.7324

Tabela 4.1: Estatísticas do Conjunto de Dados (# Jogos = Número de Jogos; # Odds BTS = Número de *odds* do Mercado *Both Team To Score*; # Odds ML = Número de *odds* do Mercado *Money Line*)

4.2 Pré-Processamento de Dados

Após a coleta de dados, a extração de atributos (*feature engineering*) é uma etapa indispensável para o sucesso de muitas técnicas de aprendizagem de máquina. Nesta seção, são

detalhadas as etapas do processo de criação de novos atributos para a tarefa de *predição de ambos marcam*.

4.2.1 Atributos de Desempenho

Em predições de futebol, o desempenho dos times nas partidas anteriores é considerado um dos fatores-chave para determinar o desempenho no próximo jogo. Uma das formas de medir o desempenho das equipes é a partir dos atributos da tabela de classificação de um campeonato no instante anterior à partida cujo resultado será predito. Esses dados podem ser extraídos a partir dos dados das *partidas disputadas*. Dessa forma, em um primeiro ciclo, foram extraídos os dados das tabelas de classificação de cada time, ao longo de uma temporada, após cada rodada. Os dados extraídos foram: *número de pontos*, *vitórias*, *empates*, *derrotas*, *gols marcados*, *gols sofridos* e *partidas realizadas*.

Considerando que estudos anteriores apontam que o desempenho das equipes é afetado pelo "fator casa", os locais das partidas também foram levados em consideração. Sendo assim, foram extraídas três visões das tabelas de classificação: uma considerando apenas os jogos como mandante, outra como visitante e outra considerando a classificação geral. Neste trabalho, esse novo grupo de atributos é chamado de **Atributos de Desempenho**.

Sabe-se que algumas técnicas de aprendizagem de máquina trabalham melhor quando recebem dados normalizados ou padronizados. Também é notório que, no futebol, devido a ajustes de calendário, alguns times podem ter disputado mais partidas que outros em algum momento do campeonato. Nesse contexto, considerar os atributos extraídos no formato original seria injusto, pois um time poderia ter pontuação maior que outro, apenas por ter disputado mais partidas. Dessa forma, os atributos previamente extraídos foram transformados em médias, ou seja, cada variável teve seu valor dividido pela quantidade de jogos disputados. Assim, o atributo "número de vitórias"foi transformado em "média de vitórias por partida", o atributo "gols marcados"em "média de gols marcados por partida", e assim sucessivamente.

4.2.2 Atributos Sequenciais

Embora os *atributos de desempenho* representem um conhecimento importante sobre o histórico das equipes, eles não incluem informações sobre a sequência dos eventos. Por exemplo, será que uma equipe, após marcar gol em uma partida, tem mais chance de continuar marcando na próxima? Ou será que essa chance diminui? Dado que o objetivo da predição é saber se ambas as equipes vão marcar, decidiu-se verificar se esses padrões sequenciais podem melhorar a predição dos classificadores. Para fazer essa modelagem, foram utilizadas cadeias de Markov (ver Seção 2.1.6).

Dois tipos de cadeia foram construídas para cada time: uma para avaliar os gols marcados e outra para avaliar os gols sofridos. Considerando que essa cadeias são bastante semelhantes, será apresentado o funcionamento de apenas uma. A outra cadeia funciona de forma análoga.

Considere, então, a construção da cadeia para modelar a sequência de partidas com gols marcados. Essa cadeia tem 2 (dois) estados possíveis: *s* (quando o time marca pelo menos um gol) e *n* (quando o time não marca gols). O desempenho no primeiro jogo do campeonato define o estado inicial da cadeia. A partir daí, a cada nova partida disputada, a cadeia atualiza as probabilidades de suas transições para refletir a nova informação recebida, ou seja, as probabilidades representam o aprendizado da cadeia, a cada rodada. A Figura 4.1 exemplifica o formato dessa cadeia, no qual:

- Os estados correspondem aos possíveis eventos (Sim ou Não);
- $P(s, s)$ é a probabilidade do time marcar pelo menos um gol, dado que isso aconteceu na partida anterior;
- $P(s, n)$ é a probabilidade do time não marcar gols, dado que ele tenha marcado pelo menos um gol na partida anterior;
- $P(n, n)$ é a probabilidade do time não marcar gols, dado que não tenha marcado gols na partida anterior;
- $P(n, s)$ é a probabilidade do time marcar pelo menos um gol, dado que não tenha marcado gols na partida anterior.

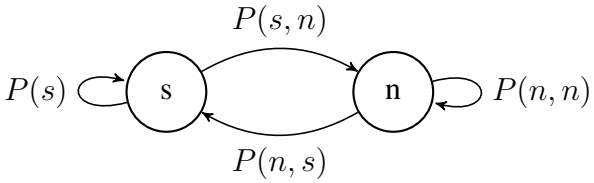


Figura 4.1: Estrutura da cadeia de Markov para capturar padrões sequenciais nos resultados das partidas disputadas por um time.

O objetivo dessa modelagem é utilizar o estado atual e as probabilidades das transições como atributos para as técnicas de aprendizagem de máquina. Assim como nos *atributos de desempenho*, os locais das partidas também foram levados em consideração. Logo, foram construídas três cadeias para cada um dos tipos (gols marcados e gols sofridos): uma para os jogos em casa, outra para os jogos como visitante e outra para todos os jogos; totalizando seis cadeias. Neste trabalho, esse novo conjunto de atributos é chamado de **Atributos Sequenciais**.

4.2.3 Atributos do Mercado

Sabendo que as *odds* (cotações) do mercado representam implicitamente probabilidades, pode-se transformar essas cotações em valores padronizados entre 0 e 1. Para esse fim, basta dividir 1 (um) pelo valor de cada *odd*. Em seguida, é necessário remover o *overround* visando extrair a real probabilidade definida por cada casa de apostas. Para isso foi utilizado o método de Shin [38] (ver Seção 2.3.3).

Uma vez estimadas as probabilidades reais para cada casa de apostas, pode-se extrair os atributos que representam a opinião geral do mercado. Para esse fim, foram extraídos quatro atributos para cada resultado: a probabilidade máxima (max), a probabilidade mínima (min), a probabilidade média (avg) e o desvio padrão (std). Formalmente, considere $A = \{a_1, a_2, ..., a_n\}$ o conjunto de probabilidades extraídas de n casas de apostas para um determinado resultado, logo a extração do novos atributos é dada por $max(A)$, $min(A)$, $avg(A)$ e $std(A)$. Neste trabalho, esse conjunto de atributos é denominado **Atributos do Mercado**.

4.2.4 Conjunto de Dados Final

A Tabela 4.2 apresenta o conjunto final de atributos usados para a tarefa de predição de *ambas marcam*, que incluem **atributos de desempenho**, **atributos sequenciais** e **atributos de mercado**.

Atributos de desempenho	Atributos sequenciais	Atributos de mercado
Nº Jogos disputados (em casa)	p_{ss} - para marcar gols (em casa)	BTTS - Prob. Sim (avg)
Média de vitórias (em casa)	p_{sn} -para marcar gols (em casa)	BTTS - Prob. Não (avg)
Média de empates (em casa)	p_{ns} - para marcar gols (em casa)	ML - Prob. Time da casa (avg)
Média de derrotas (em casa)	p_{nn} - para marcar gols (em casa)	ML - Prob. Time visitante (avg)
Média de gols marcados (em casa)	p_{ss} - para sofrer gols (em casa)	ML - Prob. empate (avg)
Média de gols sofridos (em casa)	p_{sn} - para sofrer gols (em casa)	BTTS - Prob. Sim (mac)
Média de pontos por partida (em casa)	p_{ns} - para sofrer gols (em casa)	BTTS - Prob. Não (max)
Média de jogos marcando gols (em casa)	p_{nn} - para sofrer gols (em casa)	ML - Prob. Time da casa (max)
Média de jogos sofrendo gols (em casa)	p_{ss} - para marcar gols (visitante)	ML - Prob. Time visitante (max)
Nº de jogos disputados (Visitante)	p_{sn} - para marcar gols (visitante)	ML - Prob. empate (max)
Média de vitórias (Visitante)	p_{ns} - para marcar gols (visitante)	BTTS - Prob. Sim (min)
Média de empates Visitante)	p_{nn} - para marcar gols (visitante)	BTTS - Prob. Não (min)
Média de derrotas (Visitante)	p_{ss} - para marcar gols (visitante)	ML - Prob. Time da casa (min)
Média de gols marcados (Visitante)	p_{sn} - para sofrer gols (visitante)	ML - Prob. Time da visitante (min)
Média de gols sofridos (Visitante)	p_{ns} - para sofrer gols (visitante)	ML - empate (min)
Média de pontos por partida (Visitante)	p_{nn} -para sofrer gols (visitante)	BTTS - Prob. Sim (std)
Média de jogos marcando gols (Visitante)	Estado atual (em casa)	BTTS - Prob. Não (std)
Média de jogos sofrendo gols (Visitante)	Estado atual (visitante)	ML - Prob. Time da casa (std)
		ML - Prob. Time Visitante (std)
		ML - Prob. empate (std)

Tabela 4.2: Conjunto final de atributos utilizados para a tarefa de predição de *ambas marcam*

4.3 Classificadores

Nesta seção, serão apresentados os classificadores utlizados neste trabalho e a motivação para essas escolhas.

4.3.1 Classificador baseado em classe majoritária

Uma abordagem simples para prever se ambas as equipes marcarão em uma partida é identificar a distribuição geral dos resultados para cada campeonato. Se a maioria das partidas, no conjunto de treinamento, terminar com ambas as equipes marcando, o classificador prevê

sim; caso contrário, *não*. Neste trabalho, esse modelo é identificado como **Classificador de Classe Majoritária (CCM)**. Para definir qual a classe majoritária para cada campeonato, basta analisar a distribuição dos resultados para *ambas marcam* (*sim* ou *não*). A Figura 4.2 mostra as distribuições para *sim* considerando todos os campeonatos (lado esquerdo) e por campeonato (lado direito).

Considerando todos os campeonatos, obseva-se que $50,61\%$ das partidas terminaram com ambas as equipes marcando. Embora haja uma pequena tendência para *sim*, a diferença em relação a classe *não* é estatisticamente significativa (a Tabela B.1 do Apêndice apresenta os valores do teste t).

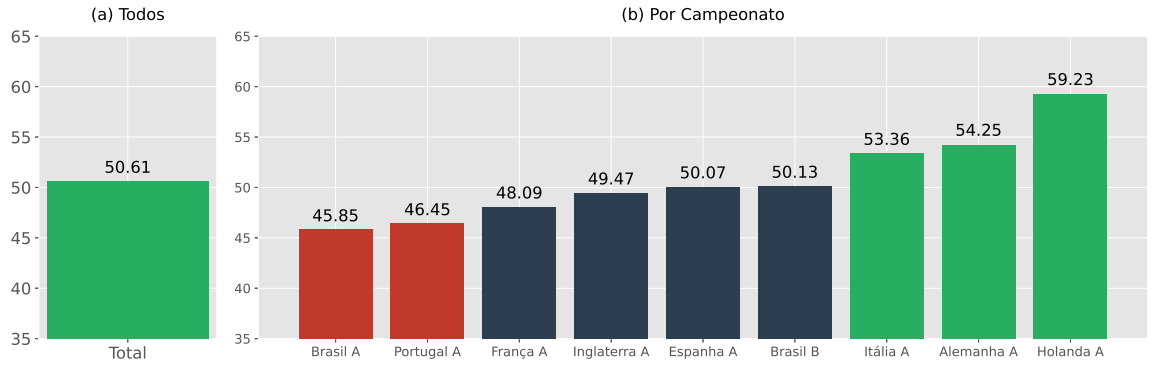


Figura 4.2: Distribuição de jogos que terminaram com ambas as equipes marcando, considerando todos os campeonatos combinados (a) e os campeonatos individualmente (b).

Em relação às distribuições por campeonato, nota-se que nos campeonatos *Itália A*, *Alemanha A*, *Holanda A*, a classe majoritária é *sim* (ambas equipes marcam), enquanto em *Brasil A* e *Portugal A*, a classe majoritária é *não*. Para os campeonatos *França A*, *Inglaterra A*, *Espanha A*, *Brasil B*, não há diferença estatística significativa, o que é confirmado pelo altos valores *de p-values*. Porém, para efeitos de predição, o **CCM** segue a classe majoritária, independente da significância do teste, ou seja, para *França A* e *Inglaterra A*, ele prevê *não* e para *Espanha A* e *Brasil B*, prevê *sim*.

Essas diferenças observadas nas distribuições, ainda que sutis, fornecem evidências de que um modelo que segue estritamente a classe majoritária (também conhecido como *regra zero*) pode fornecer melhores previsões do que um modelo aleatório. Dessa forma, esse classificador deve servir com uma linha de base ou limite inferior para o problema. É esperado que modelos mais sofisticados sejam capazes de superá-lo.

4.3.2 Classificadores baseados em Distribuição de Poisson

Classificadores baseados em modelos de Poisson têm sido amplamente usados para prever o número de gols em partidas de futebol e têm servido como uma boa linha de base para diversos trabalhos relacionados, como [104], [86], [72], [92]. Embora não tenham sido projetados para o problema de *predição de ambos marcam*, eles fazem a predição de gols, ou seja, podem facilmente ser ajustados para prever as probabilidades para *ambos marcam*.

O modelo padrão que serve de base para vários outros é o proposto por Maher [74]. A premissa central desse modelo é que o número de gols previstos para a equipe da casa e equipe visitante em uma determinada partida são variáveis independentes de Poisson, cujas médias são determinadas pelas respectivas qualidades de ataque e defesa de cada equipe. Formalmente, considerando h o time da casa, v o time visitante, X e Y o número de gols marcados por cada um respectivamente, então:

$$X \sim \text{Poisson}(\alpha_h \, \beta_v \, \gamma)$$

$$Y \sim \text{Poisson}(\alpha_v \, \beta_h)$$

no qual X e Y são independentes, $\alpha_i, \beta_i > 0$, $\forall i \in (h, v)$, α_i mede a força de ataque, β_i mede a força de defesa e $\gamma > 0$ é um parâmetro que se refere ao *fator casa* [1]. Neste trabalho, esse modelo é identificado como **Classificador de Maher (CMH)**.

Em [75] é apresentada uma melhoria para o modelo proposto por [74]. Os autores mostraram que o modelo não é totalmente adequado para prever partidas com poucos gols e introduziram um parâmetro de dependência para os seguintes resultados da partida: $0-0$, $1-0$, $0-1$ e $1-1$. Além disso, eles também propuseram duas funções de ponderação para dar mais importância aos jogos mais recentes. Neste trabalho, esse modelo é identificado como **Classificador de Dixon e Coles (CDC)**.

Em [76] é apresentada uma extensão para o framework de [75]. Os autores desenvolveram um modelo linear generalizado dinâmico Bayesiano para atualizar as estimativas dos parâmetros ao longo do tempo usando o método de Monte Carlo via Cadeias de Markov (MCMC). Neste trabalho, esse modelo é identificado como **Classificador de Rue e Salvesen (CRS)**.

[1] *fator casa* é um fator que representa uma força adicional para o time que joga em casa

Para todos os classificadores supracitados, é possível obter uma predição em formato de matriz $S \in \mathbb{R}^{a \times b}$, em que os índices a, b representam os placares da partida, sendo a os números de gols do time da casa e b, do time visitante. Os valores da matriz são as probabilidades de cada resultado possível da partida. Assim, a previsão dos classificadores para o problema de *ambas marcam* pode ser definida da seguinte forma:

$$P_{nao} = S[0,0] + S[0,1] + S[1,0]$$

$$P_{sim} = 1 - P_{nao}$$

4.3.3 Classificadores de aprendizagem de máquina

Neste trabalho são avaliados três classificadores de aprendizagem de máquina: *Gaussian Naive Bayes* (**GNB**), *Logistic Regression* (**RLO**) e *Gradient Boosting* (**XGB**). Esses classificadores são usados para predição em uma ampla gama de domínios complexos, incluindo futebol, conforme detalhado no Capítulo 3.

4.3.4 Classificadores baseados no mercado

Neste trabalho, são definidos dois classificadores baseados nas cotações do mercado de apostas: o **Classificador baseado na média de Mercado** (**CMM**) e o **Classificador baseado na casa de apostas mais justa** (**CCJ**).

O **CMM** prevê com base na opinião média de 19 casas de apostas, ou seja, define as probabilidades para *sim* e *não* de acordo com os atributos *BTTS - Prob. Sim (avg)* e *BTTS - Prob. Não (avg)* do conjunto de dados. Esse é um forte preditor sob a hipótese de eficiência do mercado. Assumindo a hipótese de mercado eficiente [105], nenhum apostador ou classificador pode vencer o mercado no longo prazo, consequentemente, esse classificador representa um limiar superior para os desempenhos de outros classificadores.

Para a construção do **CCJ**, foi analisado qual casa de apostas tem a menor média de *booksum* (a soma das probabilidades implícitas - ver Seção 2.3.2) no conjunto de treinamento. Espera-se que uma casa de apostas com baixo *booksum* seja um opção melhor para apostar, dado que as *odds* serão mais justas. Dessa forma, apostar contra ele pode permitir maiores oportunidades de lucro (mais detalhes são mostrados na Seção 4.5).

A Figura 4.3 mostra que a casa de apostas *1xBet* apresenta o *booksum* mais baixo, ou seja, o *1xBet* parece oferecer cotações mais justas quando comparada com outras casas de apostas. Desta forma, **CCJ** prevê baseado estritamente nas probabilidades de *1xBet*.

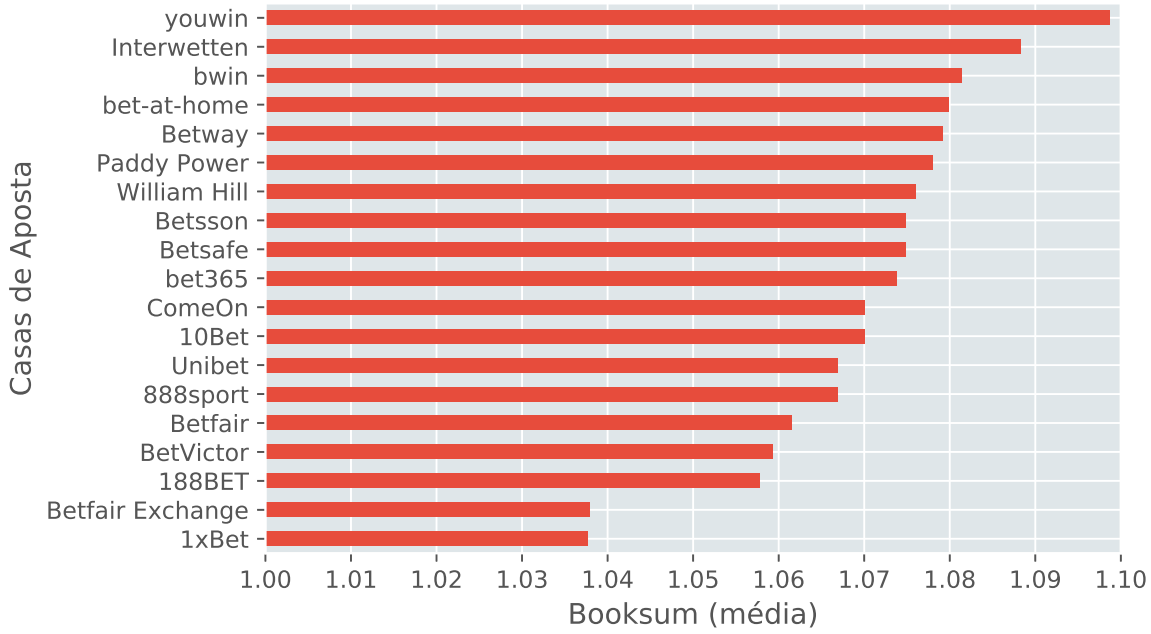


Figura 4.3: Média do *Booksum* das casas de apostas.

4.4 Experimentos

Nesta seção, são apresentados os detalhes dos experimentos a fim de reponder as seguintes questões de pesquisa:

1. PP1: Quão difícil é o problema de predição de ambas marcam?
2. PP2: Os classificadores avaliados são capazes de superar as casas de apostas?
3. PP3: Os classificadores são úteis para desenvolver estratégias de apostas lucrativas?

Para responder às questões PP1 e PP2, foi examinado o desempenho dos classificadores definidos na seção anterior. Para responder à questão PP3, foi analisado o desempenho de diferentes estratégias de apostas com base nas previsões desses mesmos classificadores.

4.4.1 Divisão do Conjunto de Dados

Inicialmente, o conjunto de dados foi dividido em $66,67\%$ das instâncias para treinamento e as $33,33\%$ restantes para teste, ou seja, o conjunto de treinamento contém partidas das primeiras quatro temporadas (2013-2016), enquanto o conjunto de teste contém partidas das duas últimas temporadas (2017-2018). Para todos os experimentos, os classificadores foram treinados usando a estratégia de janela crescente [106], ou seja, em cada iteração, as instâncias de teste da iteração anterior são adicionadas como instâncias de treinamento e usadas para prever a próxima janela de tempo. No caso deste trabalho, após cada rodada, as partidas avaliadas do conjunto de teste são adicionadas ao conjunto de treinamento para a previsão da próxima rodada, ou seja, todos os classificadores são retreinados a cada rodada.

4.4.2 Implementação dos Classificadores

Para implementação dos classificadores baseados em Poisson, **CMH**, **CDC** e **CRS**, foi utilizado o pacote *goalmodel* [2]. No caso dos classificadores de aprendizagem de máquina **GNB** e **LRE** foi usado o framework *scikit-learn*[3]. Por fim, para o **XGB**, foi usada a biblioteca *XGBoost*[4].

4.4.3 Otimização dos Hiperparâmetros

Para ajustar os hiperparâmetros de *XGB*, foi usada uma pesquisa de grade aleatória (*random search*) com validação cruzada com 5-folds, 100 iterações e os seguintes valores: *n_estimators* = [0, 1, 2 ... 998, 999, 1000], *max_depth* = [1,2,3,4,5], *learning_rate* = [0,1, 0,05, 0,01, 0,001], *colsample_bytree* = [.6, .7, .8, .9,1], *subsample* = [.7, .8, .9,1], e *min_child, min_weight* = [1, 2, 3, 4].

Os classificadores baseados em Poisson utilizam apenas os placares das partidas como atributos. Para fins de previsão, é útil pesar a influência de cada resultado da partida, de forma que as partidas mais antigas tenham menos influência do que as mais recentes. Essa ponderação é controlada por um parâmetro de tempo ξ. Neste trabalho, o parâmetro tempo ξ

[2] `https://github.com/opisthokonta/goalmodel`

[3] `https://scikit-learn.org/stable/`

[4] `https://xgboost.readthedocs.io/en/latest/`

foi otimizado considerando o *número de dias* entre as partidas como unidade de tempo. Para isso, foram avaliados valores entre 0 e $0,008$ (semelhante ao trabalho de [72]). Por exemplo, o valor ideal encontrado para o modelo que prevê a primeira rodada da temporada de 2017 foi $\xi = 0,001$.

Enquanto os classificadores baseados em Poisson precisam apenas dos placares anteriores, os classificadores de aprendizagem de máquina podem ser alimentados com diferentes conjuntos de atributos. Assim, para esses classificadores, foram realizados experimentos com diferentes conjuntos de atributos, a fim de avaliar quais desses conjuntos apresentam melhor desempenho.

No primeiro experimento, foram usados apenas atributos derivados dos resultados das partidas, ou seja, *Atributos de desempenho* e *Atributos sequenciais*. No segundo experimento, foram usados apenas recursos derivados das cotações das casas de apostas, ou seja, *Features de Mercado*. Por fim, no terceiro experimento, foram usados todos os atributos disponíveis. Neste trabalho, para distinguir cada combinação entre um classificador e um conjunto de atributos, foi adicionado ao rótulo do classificador um sublinhado seguido por uma letra maiúscula que identifica o grupo de atributos usado. Por exemplo, para o classificador **GNB** temos respectivamente: **GNB_P** (Atributos de desempenho e sequenciais), **GNB_M** (Atributos de mercado) e **GNB_A** (Todos os atributos disponíveis). A mesma lógica é usada para **RLO** e **XGB**. O conjunto de dados, código-fonte e documentação podem ser encontrados no Github [5].

4.4.4 Métricas de Avaliação

Para responder as questões de pesquisa PP1 e PP2, foi avaliado em termos de acurácia (ACC) e *Brier Score* (BRS) o desempenho de cada modelo. A acurácia é a porcentagem de previsões corretas, enquanto o *Brier Score* mede o desempenho do modelo na previsão da probabilidade de cada classe. Neste trabalho, todas as probabilidades são expressas como porcentagens. Formalmente, considerando N o número de jogos, m uma instância de um jogo, $\hat{y}_m$ a probabilidade prevista pelo classificador para *ambas marcam* e o_m o resultado real (0 se não acontecer; 1, caso contrário), pode-se definir o *Brier Score* formalmente como:

[5] `https://github.com/igormago/doutorado`

$$BRS = \frac{1}{N}\sum_{m=1}^{N}(\hat{y}_m - o_m)^2 \tag{4.1}$$

Para a questão de pesquisa PP3, os modelos foram avaliados em termos de lucratividade (*LUC*) e Retorno do Investimento (*RoI*). A lucratividade mede quanto dinheiro se ganha ou perde (saldo) depois de realizar uma aposta com base em uma determinada estratégia. Em outras palavras, se uma aposta acerta determinado resultado, um lucro é obtido com base nas probabilidades oferecidas pela casa de apostas. Por outro lado, se a aposta falha, perde-se todo o valor apostado. Assim, *lucratividade* pode ser definida como a soma desses ganhos e perdas, considerando um conjunto de jogos, enquanto o *RoI* mede o ganho ou a perda obtida em um conjunto de apostas em relação à quantidade de dinheiro apostado.

4.5 Discussão dos Resultados

Nesta seção, as questões de pesquisa deste trabalho são respondidas e discutidas a partir dos resultados dos experimentos.

4.5.1 PP1: Quão difícil é o problema de predição de ambas marcam?

Para responder essa questão, pode-se analisar a acurácia de dois classificadores empíricos: **CCM** e **CMM**.

Sob uma perspectiva puramente de eficiência de mercado, não deveria ser possível haver um preditor melhor do que a própria "opinião"do mercado. Assim, de certa forma, o **CMM** pode ser considerado um limite superior, ou seja, não deve ser possível conceber um modelo preditivo melhor do que ele. Então, comparando **CMM**, o melhor classificador sob uma hipótese de eficiência de mercado, com **CCM**, a linha de base mais simples que se pode pensar, é possível avaliar a dificuldade do problema, afinal quanto maior a diferença entre os dois, maior é a previsibilidade da tarefa.

Os resultados mostram que, considerando todas as partidas no conjunto de teste, o **CMM** tem uma acurácia de $55,34\%$ contra $51,51\%$ de **CCM** (ver Figura 4.4). Sabendo que *ambas marcam* é um problema de classificação binária, é possível observar que o desempenho de ambos os classificadores é apenas um pouco melhor do que um classificador aleatório que

teria uma acurácia média de 50%. Além disso, a diferença de desempenho entre **CMM** e **CCM** também é relativamente pequena (menos de 4%), o que indica que pode não haver muito espaço para melhorias. Portanto, esses resultados fornecem boas evidências de que a predição de *ambos marcam* pode ser considerada um problema difícil.

A comparação entre CCM e CMM também pode ser feita considerando os jogos agrupados por campeonato (ver Figura 4.4). Nesse cenário, mediante o Teste de Wilcoxon (ver Tabela B.2 do Apêndice), observa-se que **CMM** não é significativamente melhor que **CCM** nos campeonatos *Alemanha A*, *Holanda A* e *Espanha A*. No *Brasil B* especificamente, **CMM** apresentou desempenho ainda pior do que **CCM**. Para os outros campeonatos, **CMM** apresentou desempenho significativamente superior. Esses resultados reforçam a dificuldade do problema, no qual, em alguns casos, o mercado não consegue superar nem uma abordagem tão simples como o **CCM**. Além disso, esses resultados também dão indícios que alguns campeonatos podem ser mais fáceis de prever do que outros.

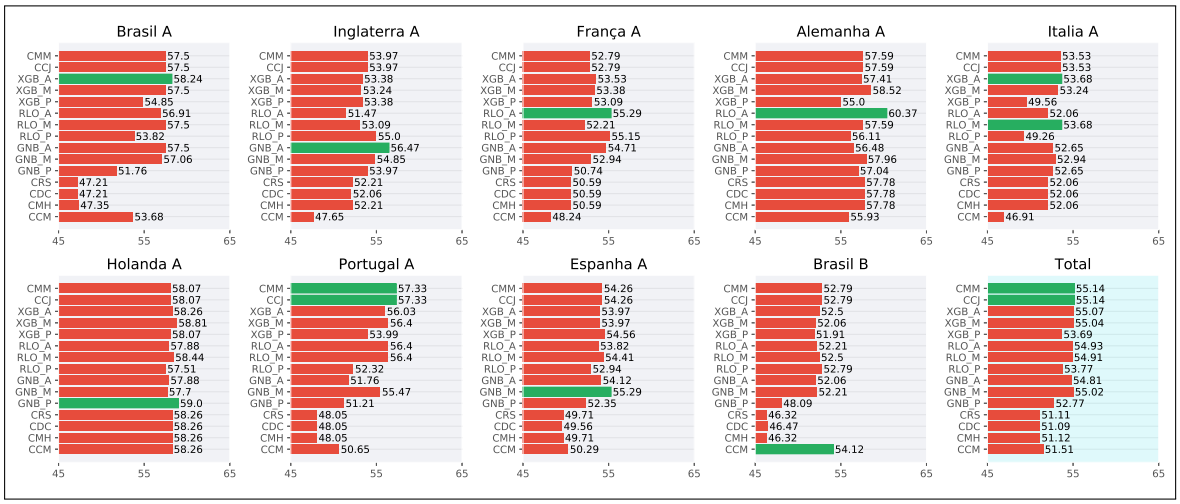


Figura 4.4: Desempenho dos classificadores em termos de acurácia.

4.5.2 PP2: Os classificadores avaliados são capazes de superar as casas de apostas?

Esta questão de pesquisa visa analisar se os classificadores treinados com informações sobre partidas de futebol são capazes de fazer previsões mais precisas do que o mercado, representado neste trabalho pelos classificadores **CMM** e **CCJ**. A Figura 4.4 apresenta o desempenho

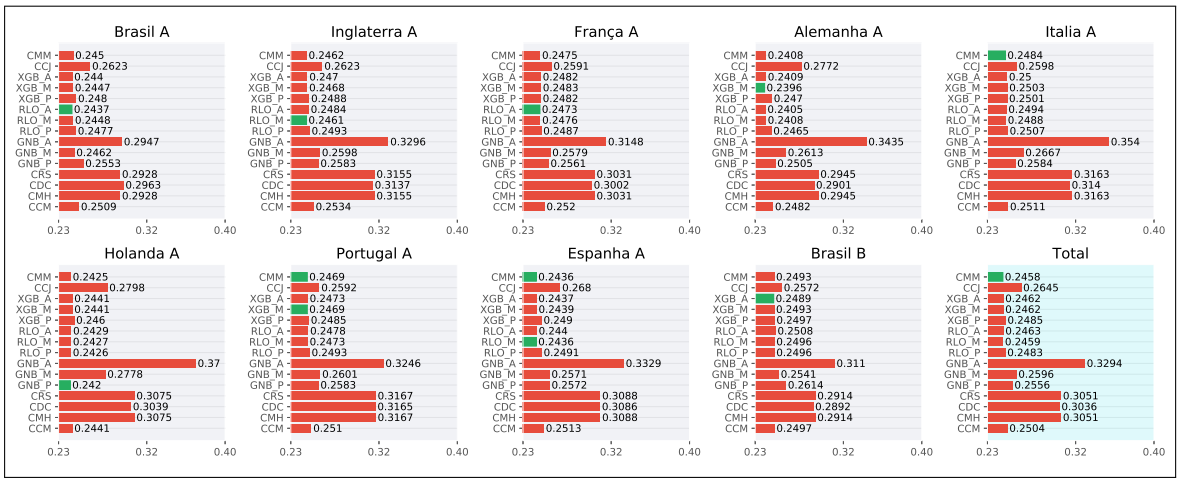


Figura 4.5: Desempenho dos classificadores em termos de *Brier Score*.

dos classificadores em termos de acurácia, enquanto que a Figura 4.5 o representa em termos de *Brier Score*.

Primeiramente, pode-se analisar os classificadores quando são alimentados apenas com dados extraídos do desempenho das equipes, ou seja, quando eles não usam dados de mercado. Nesse cenário, considerando todos os campeonatos **RLO_P** [ACC: $53,77$, BRS: $0,2483$] e **XGB_P** [ACC: $53,69$, BRS: $0,2485$] obtiveram um desempenho melhor do que os outros. No entanto, eles não foram capazes de superar as casas de apostas. Os classificadores baseados em Poisson, em particular, não foram capazes sequer de superar o **CCM**.

O baixo desempenho dos *classificadores baseados em Poisson* pode ser explicado pelo fato de esses modelos serem dependentes do desempenho individual de cada equipe. Dado que, a cada nova temporada, novas equipes são promovidas e rebaixadas, algumas equipes aparecem apenas no conjunto de teste, ou seja, não há informações prévias sobre essas equipes no conjunto de treinamento. A estratégia de avaliação usada (janela crescente) ameniza o problema, pois, rodada após rodada, esses classificadores aprendem um pouco mais sobre os "novos"times, mesmo aqueles que foram promovidos recentemente. No entanto, de forma geral, o desempenho dos classificadores baseados em Poisson, é inicialmente prejudicado pela limitação de dados sobre novas equipes. Nos classificadores de aprendizagem de máquina esse problema é mitigado, pois eles não aprendem padrões por equipe, individualmente, ou seja, mesmo para equipes recém-promovidas, esses classificadores podem

aproveitar os padrões aprendidos sobre o desempenho de qualquer equipe recentemente promovida, no conjunto de treinamento.

Uma segunda visão do problema é avaliar como os classificadores se comportam quando alimentados com informações de mercado. Analisando os *classificadores de aprendizagem de máquina* quando alimentados apenas com atributos do mercado, é possível observar que eles tem melhoria no desempenho. Nesse caso, a diferença entre a accurácia dos melhores classificadores é bem pequena: **XGB_M** [ACC: 55.04, BRS: $0,2462$], **RLO_M** [ACC: $54,91$, BRS : $0,2459$] e **GNB_M** [ACC: 55.02, BRS: $0,2596$], e não há diferença estatística entre esses classificadores e **CMM** [ACC: $54,91$, BRS: $0,2459$]. Uma hipótese para essa diferença muito pequena é que, se o mercado é eficiente quando os classificadores são alimentados apenas por informações de mercado, eles acabam reproduzindo o mesmo comportamento preditivo do próprio mercado.

Para verificar se isso é verdade, foram calculados os coeficientes de Kappa entre os classificadores de aprendizagem de máquina e os classificadores de mercado. Esse coeficiente pode ser usado para descrever o nível de concordância entre os classificadores. Quanto mais próximo de 1, mais forte é a concordância entre eles. A Figura 4.6 mostra os valores obtidos.

Os resultados confirmam que os *classificadores de aprendizagem de máquina* quando alimentados com atributos de mercado têm uma correlação quase perfeita com **CMM** ($k > 0,8$), enquanto que quando alimentados com características de desempenho têm apenas correlação moderada ($0,4 < k < 0,6$) ou correlação discreta ($0,2 < k < 0,4$). Em trabalhos futuros, pode-se investigar mais a fundo as características dos jogos nos quais os classificadores divergiram do mercado, a fim de encontrar eventuais padrões de ineficiência de mercado.

Por fim, quando são usados todos os atributos disponíveis, os *classificadores de aprendizagem de máquina* não apresentaram melhora significativa em relação aos experimentos anteriores, o que pode indicar que os *atributos de mercado* já contém as informações mais relevantes e, portanto, parecem incorporar todas as informações contidas nos atributos de desempenho e de padrões sequenciais.

Assim, PP2 pode ser respondido a partir de duas perspectivas. Dado que a diferença entre os melhores classificadores de aprendizagem de máquina e classificadores de mercado é mínima, pode-se assumir que seus desempenhos são equivalentes. Nesse sentido, a resposta

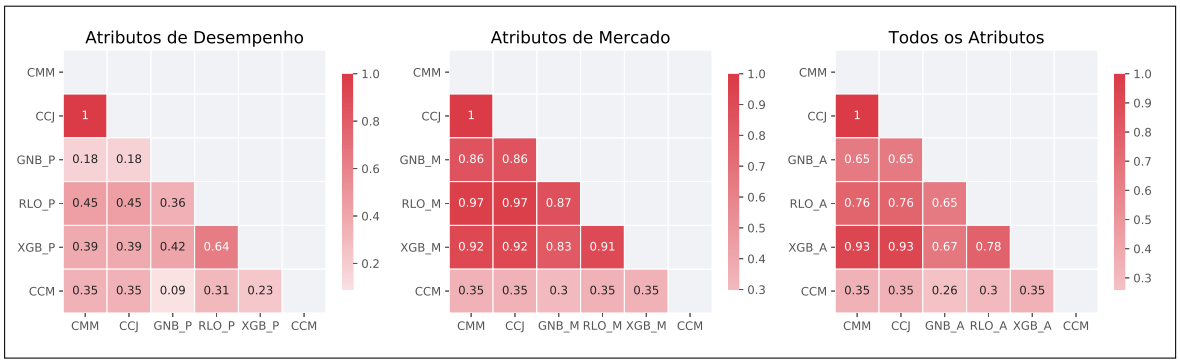


Figura 4.6: Coeficiente de Kappa para avaliar a correlação entre as predições dos classificadores.

para PP2 é negativa. Porém, em alguns campeonatos, os classificadores baseados em aprendizagem de máquina são tão bons quanto **CMM** e um pouco melhores que o **CCJ**, em termos de *Brier Score*. Esse resultado fornece indicações de que os classificadores de aprendizagem de máquina têm uma calibração melhor na previsão das probabilidades *BTTS* do que **CCJ**, o que permitiria criar estratégias lucrativas contra a casa de apostas $1xBet$. Assim, nesta perspectiva, considerando apenas casa de apostas mais justa, a resposta ao PP2 é positiva.

4.5.3 PP3: Os classificadores são úteis para desenvolver estratégias de apostas lucrativas?

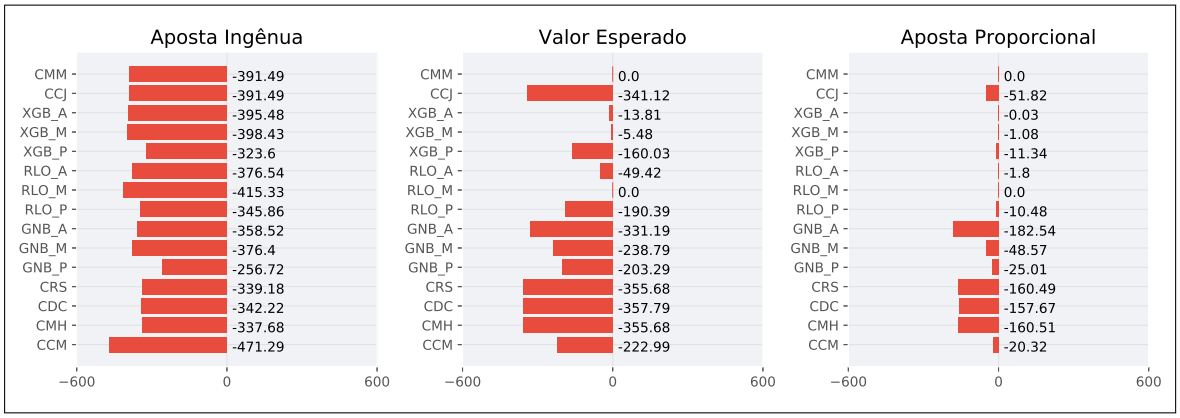


Figura 4.7: Lucratividade obtida através de diferentes estratégias de apostas: AI, VE and AP

Esta questão de pesquisa tem como objetivo avaliar se os classificadores são úteis para o desenvolvimento de estratégias lucrativas no mercado de apostas. Para tanto, foram con-

sideradas três estratégias de apostas denotadas neste trabalho como *Apostas ingênuas* (AI), *Valor esperado* (VE) e *Apostas proporcionais* (AP). Foi avaliada a aplicação dessas três estratégias para cada um dos classificadores em termos de *Lucratividade* (LUC) e *Retorno do Investimento* (RoI).

AI é uma estratégia direta adotada por apostadores menos experientes que acabam apostando no resultado que acham que vai acontecer, independentemente das *odds* oferecidas pelas casas de apostas. Assim, nessa abordagem ingênua, é apostado $\$1$ no resultado (*sim* ou *não*) previsto por um determinado classificador.

Na segunda estratégia, as apostas são feitas de acordo com o *Valor esperado* (VE). Essa estratégia leva em consideração a diferença entre a probabilidade definida pelo mercado e pelo classificador. É uma estratégia amplamente avaliada em trabalhos relacionados, como [72], [107] e [92]. Formalmente, o valor esperado de uma aposta em uma partida é dado pela Equação 4.2, onde A representa um determinado resultado (*sim* ou *não*), $P(A)$ é a probabilidade do resultado A de acordo com o classificador e $odds(A)$ é a probabilidade do mercado para o resultado A. Assim, a estratégia é apostar apenas em resultados cujo valor esperado seja positivo, $VE(A) > 0$.

$$VE(A) = P(A) \ * \ odds(A) \ - \ 1 \tag{4.2}$$

A terceira estratégia é a de *Aposta proporcional* (AP). A AP leva em consideração as chances e o valor esperado para encontrar a aposta ideal em uma aposta dada pela Equação 4.3. A estratégia AP é semelhante ao critério de Kelly [108], mas difere por não levar em consideração o saldo total do apostador. Neste trabalho, fixa-se a aposta máxima em $\$1$ por partida, como feito por [72].

$$AP(A) = \frac{(odds(A) + 1)P(A) - 1}{odds(A)} \tag{4.3}$$

Inicialmente, todas as estratégias foram testadas, levando-se em conta as probabilidades médias oferecidas pelo mercado. Como esperado, em um mercado eficiente, nenhum classificador foi capaz de obter uma estratégia lucrativa. A Figura 4.7 mostra que todos os classificadores apresentaram lucratividade negativa ao considerar todos os jogos do conjunto de teste.

Uma alternativa para tentar obter lucro é sempre apostar na casa de apostas que oferece a melhor oferta (as probabilidades máximas) para cada jogo. Contudo, para esta situação, seria necessário que o apostador tivesse contas em 19 casas de apostas, o que não parece razoável. Assim, foi adotada a estratégia de apostar contra uma única casa de apostas. Para isso, foi escolhida a casa de apostas *1xBet*, que geralmente oferece *odds* melhores que as outras casas de apostas, conforme mostrado na Seção 4.3.4. As Figuras 4.8, 4.9 e 4.10 apresentam a lucratividade e o RoI obtidos através das estratégias AI, VE e AP, respectivamente (as Tabelas B.3, B.4 e B.5 do Apêndice apresentam os valores do Teste de Wilcoxon para as estratégias mencionadas).

Em geral, nenhum classificador teve sucesso usando a estratégia AI. Somente no campeonato *Brasil A* mais de um classificador foi capaz de proporcionar um retorno positivo, porém com um ROI insignificante (ver Figura 4.8).

Com relação à estratégia VE, os resultados foram notavelmente melhores, mas os ganhos permaneceram pequenos. Em geral, o melhor classificador foi o próprio mercado (**CMM**), que obteve lucratividade de $24,30$ mas com apenas $1,95\%$ de RoI. Analisando por campeonato (ver Figura 4.9), em *Inglaterra A* e *Itália A*, os classificadores de aprendizagem de máquina não foram eficientes, mas nos demais, os classificadores **RLO** e **XGB** foram capazes de obter lucro em várias partidas, principalmente nos campeonatos *França A*, *Brasil B* e *Portugal A*. Os resultados mostram que pode haver campeonatos com mais oportunidades de lucro do que outros.

Em relação à estratégia AP, em comparação com VE, o ROI foi geralmente melhor, enquanto a lucratividade foi pior. Devido à pequena diferença entre as probabilidades previstas pelos classificadores e as previstas pelas casas de apostas, a estratégia AP acabou por estabelecer uma aposta bastante pequena. Assim, seria necessário elevar o risco, aumentando significativamente o valor máximo fixado por jogo para que essa estratégia atingisse um lucro consistente.

Nos experimentos anteriores, as estratégias foram avaliadas considerando apostas em todas as partidas. No entanto, é possível restringir a estratégia para apostar em apenas um grupo específico de jogos. Para aprofundar a análise dessa ideia, a lucratividade dos classificadores foi reavaliada considerando apenas um grupo de partidas em que a probabilidade de

"ambas as equipes marcarem"(sim), de acordo com o mercado, pertence a um determinado intervalo.

Para isso, foi calculada a lucratividade em todas as faixas de probabilidades possíveis (considerando números inteiros), usando as três estratégias, contra o bookmaker *1xBet*. Os resultados demonstraram que a estratégia VE foi a mais eficaz para todos os classificadores. A Figura 4.11 exibe os resultados para todos os intervalos possíveis. A Tabela 4.3 mostra os melhores resultados, destacando o classificador, a faixa de probabilidades (para *sim*) que ofereceu o melhor retorno, o número de partidas identificadas pelo classificador com valor esperado positivo (dentro da faixa), além da lucratividade e RoI obtido com as apostas.

Os resultados são promissores, pois todos os classificadores obtiveram algum retorno positivo. Mesmo os *classificadores baseados em Poisson*, que não haviam obtido bons resultados em termos de acurácia, foram capazes de lucrar, obtendo um RoI superior a 13%, em todos os casos. **CCM** teve um resultado muito semelhante aos classificadores baseados em Poisson. Pode-se observar que esses classificadores foram capazes de obter lucro em jogos em que a probabilidade de *sim* é considerada pequena para a casa de apostas (39% - 42%).

RLO_M obteve o maior lucro geral $(61,57)$, mas para isso teve que apostar em muitas partidas, o que resultou em um pequeno RoI $(3,25\ \%)$ em relação aos demais. Por outro lado, **CMM** apresentou o maior RoI $(13,53\ \%)$, mas lucratividade modesta $(38,32)$. Os classificadores **GNB** e **XGB** também tiveram bons resultados, mas curiosamente em diferentes faixas de probabilidade. Enquanto os classificadores **GNB** foram melhores quando a probabilidade de *sim* para *ambas marcam* estava entre $39\ \%$ e $51\ \%$, os classificadores **XGB** tiveram melhor desempenho quando essa probabilidade ficou entre $55\ \%$ e $64\ \%$. Esses resultados sugerem que classificadores distintos podem ser mais lucrativos em combinações diferentes. Em trabalhos futuros, pode-se estudar esses padrões e criar estratégias mais sofisticadas e inteligentes para identificar os grupos de jogos mais lucrativos.

Finalmente, assim como PP2, a questão de pesquisa PP3 pode ser respondida sob duas perspectivas. Quando os classificadores apostam em todas as partidas contra as *odds* médias do mercado, não é possível obter lucro com as estratégias avaliadas neste trabalho. Assim, nessa perspectiva, a resposta para PP3 é negativa. No entanto, quando considerada apenas a casa de apostas mais justa e quando o número de apostas é limitado, seja por campeonato

ou por um intervalo de probabilidades, os classificadores podem ser úteis para conceber estratégias de apostas lucrativas. Nessa perspectiva, a resposta para PP3 é positiva.

4.6 Considerações Finais

Neste capítulo, foi investigado um problema de pesquisa ainda inexplorado no domínio das previsões de futebol, ou seja, o problema de previsão de "ambas marcam". Foi avaliado o desempenho das previsões de vários classificadores em termos de acurácia e eficiência no mercado de apostas. Para isso, foi construído um conjunto de dados robusto, mediante um cuidadoso processo de *feature engineering*, composto por dados de 9 (nove) ligas nacionais e 19 (dezenove) casas de apostas.

Primeiramente, foi observado que o problema de predição de "ambas marcam"é uma tarefa difícil, onde a opinião média do mercado é apenas ligeiramente melhor do que um modelo simples que sempre prevê a classe majoritária. Em seguida, foi realizado um conjunto abrangente de experimentos que mostra que os classificadores de aprendizagem de máquina são capazes de superar o mercado em termos de acurácia, embora apenas em alguns cenários. Na verdade, as *features* mais importantes foram aquelas extraídas do próprio mercado de apostas, onde os classificadores basicamente aprenderam a imitar o mercado na maioria das partidas.

Outro experimento usou as previsões do classificador como entrada para estratégias de apostas. O pressuposto é que, se for possível vencer o mercado de forma sistemática em termos de lucratividade, haverá indícios de que o mercado é ineficiente. Foi possível observar que, ao apostar em todas as partidas, os classificadores não alcançam uma margem de lucro significativa. Por outro lado, ao selecionar campeonatos específicos ou determinados grupos de jogos, há indícios de que os classificadores podem ser mais lucrativos, tornando as abordagens de aprendizado de máquina ainda mais interessantes para esse tipo de problema.

Tabela 4.3: Resultado da Estratégia VE em um grupo de jogos selecionados

Classificador	Probabilidades	#Jogos	LUC	ROI
CCM	39% - 42%	303	33.68	11.12%
CMH	39% - 42%	320	46.56	13.03%
CDC	39% - 42%	319	45.91	13.39%
CRS	39% - 42%	320	46.56	13.03%
GNB_P	39% - 42%	304	31.71	10.43%
GNB_M	43% - 48%	1409	39.48	2.80%
GNB_A	50% - 51%	277	28.21	10.18%
RLO_P	50% - 63%	1904	61.97	3.25%
RLO_M	33% - 48%	432	43.68	10.11%
RLO_A	30% - 52%	1732	36.14	2.09%
XGB_P	55% - 63%	1138	40.62	3.57%
XGB_M	55% - 64%	760	31.29	4.12%
XGB_A	55% - 64%	806	51.85	6.43%
CMM	39% - 49%	350	38.32	13.53%

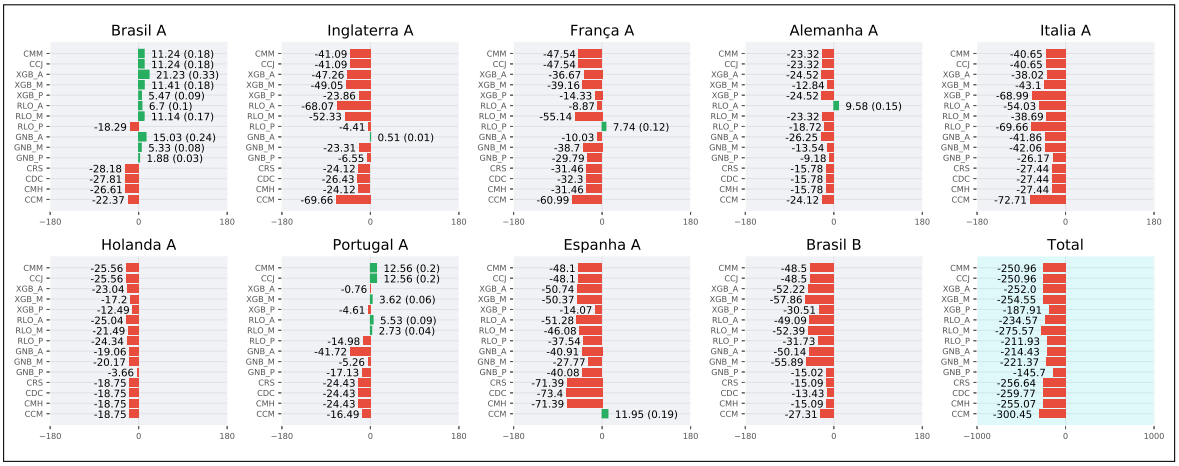


Figura 4.8: Lucratividade (e RoI) obtida através da estratégia Aposta Ingênua (AI) contra a casa de apostas *1xBet*.

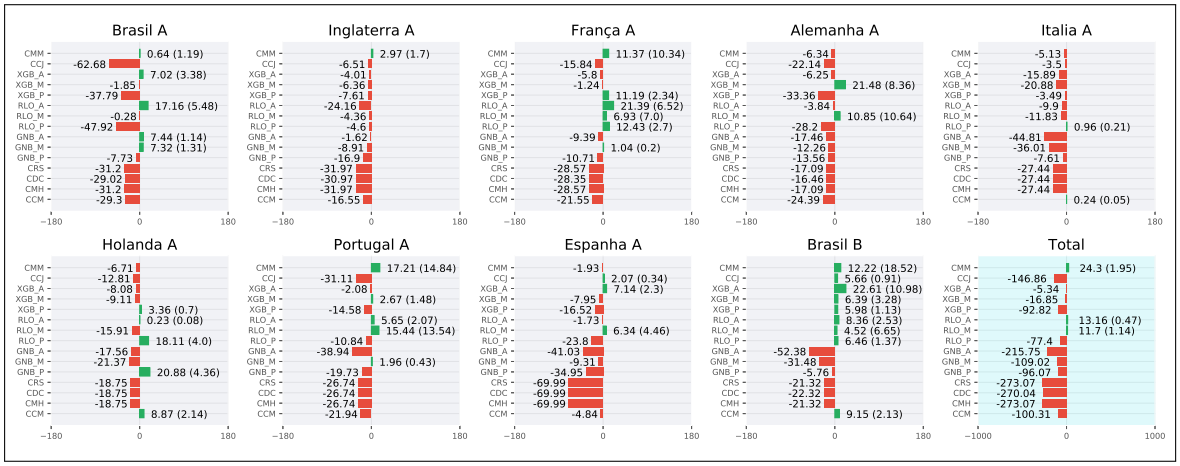


Figura 4.9: Lucratividade (e RoI) obtida através da estratégia Valor Esperado (VE) contra a casa de apostas *1xBet*.

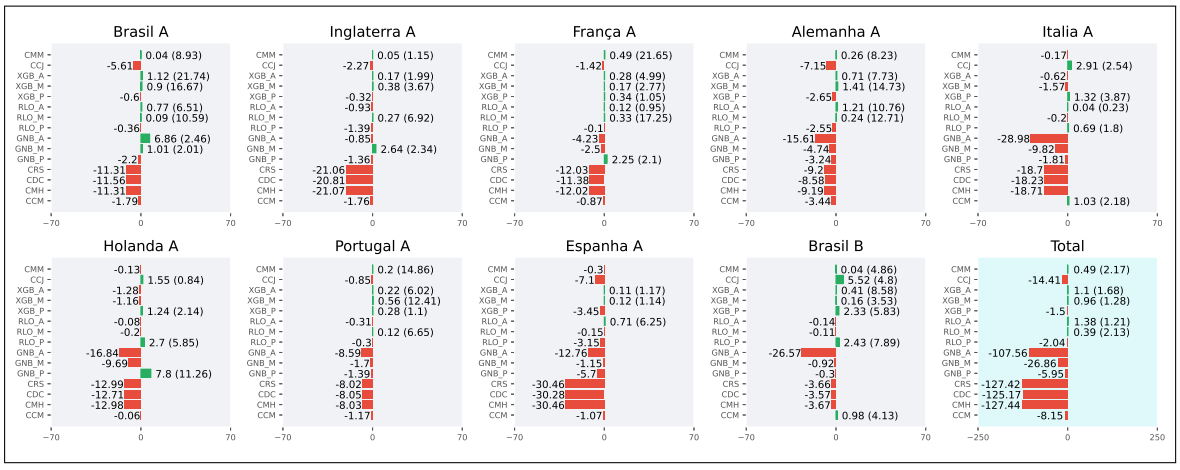


Figura 4.10: Lucratividade (e RoI) obtida através da estratégia Aposta Proporcional (AP) contra a casa de apostas *1xBet*.

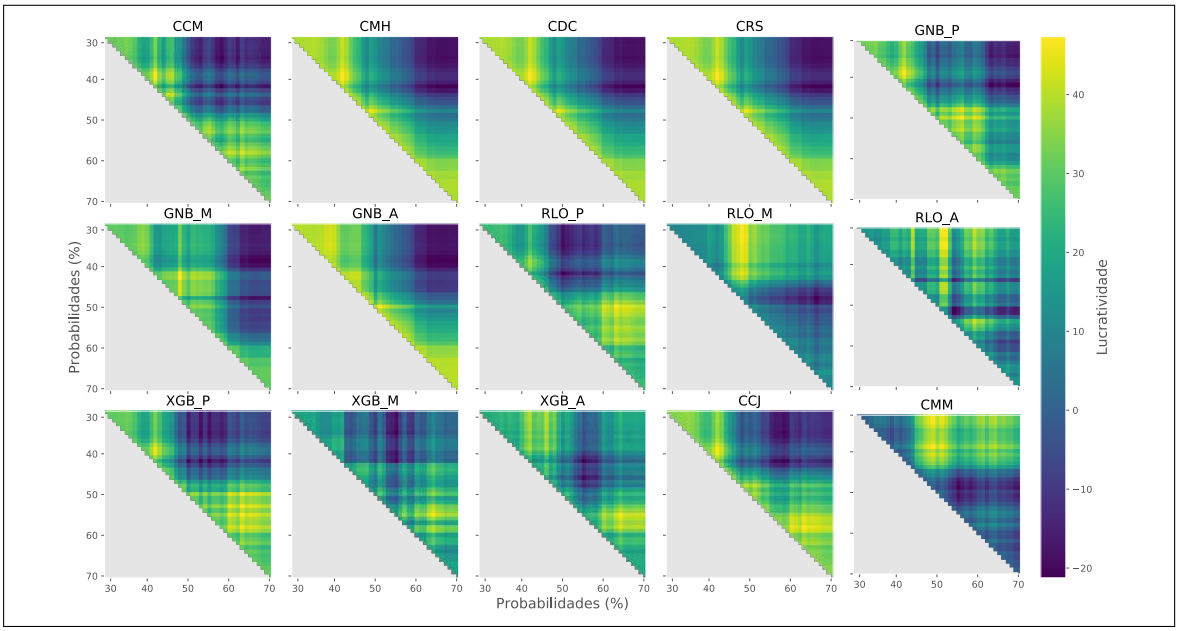


Figura 4.11: Lucratividade obtida através da estratégia Valor Esperado (VE) contra a casa de apostas *1xBet* considerando determinados intervalos de probabilidades. A coordenada Y é o início do intervalo de probabilidade, enquanto que a coordenada X é o final. A lucratividade obtida em cada intervalo é representada pela variação das cores.

Capítulo 5

Predição do Resultado durante a Partida

Neste capítulo, será detalhado o método proposto por este trabalho para prever o resultado final de uma partida, de forma contínua, do início até próximo do final do jogo.

O capítulo está dividido como segue. Primeiramente, são apresentados os conjuntos de dados coletados para o problema. São descritas as etapas de pré-processamento de dados e apresentado o conjunto de dados final usado para a modelagem de *predição durante a partida*. Em seguida, é aprensentada uma análise exploratória dos dados. Logo depois, são detalhados os classificadores e os experimentos executados, como também discutidos os resultados obtidos. Por fim, são apresentadas as conclusões e considerações finais sobre o capítulo.

5.1 Coleta de Dados

Dentre os conjuntos de dados disponíveis publicamente para previsões de futebol, até o limite do nosso conhecimento, nenhum deles inclui informações estruturadas sobre os eventos (chutes a gol, escanteios, cartões, etc.) que ocorrem no decorrer de uma partida. Portanto, para este trabalho, foi necessário estruturar um conjunto de dados próprio. Nesta seção, serão descritos os passos para construção desse conjunto de dados, de forma que os classificadores de aprendizagem de máquina possam ser treinados de forma eficiente.

5.1.1 Integração de Dados

Inicialmente, foram examinados dados de $1.436.529$ partidas de futebol ocorridas entre 1º de julho de 2017 e 31 de fevereiro de 2020. Os dados foram obtidos no portal *sportmonks.com*[1]. Apesar do alto número de partidas monitoradas, foi observado que apenas 143.104 delas, continham dados minuto a minuto sobre os eventos ocorridos em uma partida. Neste trabalho, esse conjuntos de dados é chamado de *streaming da partida* e inclui os seguintes eventos: gols, cartões amarelos, cartões vermelhos, chutes no gol, chutes para fora, ataques, ataques perigosos, escanteios e posse de bola.

Além das informações referentes aos eventos das partidas, há um outro tipo de informação muito relevante para predição de resultados: as chances calculadas pelo mercado de apostas. O *sportmonks.com* é uma boa fonte para coleta de *streaming da partida*, porém não fornece informações minuto a minuto sobre as probabilidades das casas de apostas. Dessa forma, para o mesmo período mencionado acima, foram coletados dados de 192.550 partidas disponíveis para apostas no portal *betfair.com*[2] [109]. Neste trabalho, esse conjunto de dados que contém as *odds* do mercado minuto a minuto é chamado de *streaming das odds*.

Após a obtenção dos dois conjuntos de dados, foi executada a tarefa de *entity matching* [110] para detectar instâncias que se referem ao mesmo jogo. Primeiramente, os jogos foram agrupados pelos atributos *data* e *campeonato*. Em seguida, foi aplicada uma função de similaridade para comparar os *nomes das equipes* e o *placar da partida*. Esse processo gerou um conjunto de dados integrado com 40.095 partidas.

5.1.2 Limpeza e Seleção de Dados

A primeira versão do conjunto de dados integrado apresentou dois problemas principais. O primeiro problema é que alguns jogos não dispunham de apostas contínuas durante o jogo. Dessa forma, esses jogos não tem as *odds* atualizadas constantemente e consequentemente, as chances calculadas pelo mercado não estão representadas corretamente minuto a minuto.

[1] A documentação sobre como acessar os dados de spotmonks.com está disponível em `https://www.sportmonks.com/docs/`

[2] A documentação sobre como acessar os dados do *betfair.com* está disponível em `https://historicdata.betfair.com/`

Por esse motivo, todas as partidas em que não houve ajuste regular das *odds* foram removidas do conjunto de dados.

O segundo problema é que o tempo de jogo está indexado de forma diferente nos conjuntos de dados. No *streaming da partida*, o índice é o minuto da partida em que um evento ocorreu. Por exemplo, no minuto 6 (seis), há um cartão vermelho. Já no *streaming de odds*, o índice é a data/hora correspondente ao valor das *odds*. Assim, para sincronizar os *streamings*, o índice de *streaming de odds* foi ajustado para representar os minutos, a partir do cálculo da diferença do horário de início da partida e do horário do evento. Por exemplo, se no *streaming de odds* um jogo teve início às 4:02h, a informação referente às 4:08h corresponde ao minuto 6 (seis) do *streaming da partida*.

Após essa sincronização inicial, foi feita uma verificação na integridade do conjuntos de dados. A principal preocupação era verificar se havia mudanças significativas nas *odds* no momento em que há ocorrência de gols. O racional é que, quando ocorre um gol, a expectativa do mercado (representada pelas probabilidades) deve mudar imediatamente. Se isso não for observado, pode significar que as informações sobre o minuto do gol estão diferentes nos conjuntos de dados iniciais e, portanto, há uma ameaça à validade do conjunto de dados final. Sendo assim, foram eliminadas todas as partidas em que não há variação consistente nas *odds* após a ocorrência de gols, resultando em um conjunto de dados final com 9.416 partidas.

5.1.3 Representação Final dos Conjuntos de Dados

Até então, para cada partida, a coleção de dados contém um *streaming* com as frequências cumulativas de cada evento, para cada equipe. A Tabela 5.1 apresenta um exemplo de um *streaming da partida*, contendo alguns eventos durante os primeiros 10 (dez) minutos de uma partida.

Uma vez que as partidas de futebol podem ser vistas como uma sequência de eventos, é possível tratar o problema de predição de resultados como uma classificação de séries temporais. Para isso, os *streamings* das partidas (que tem dados cumulativos) devem ser transformados em séries temporais. Nesse sentido, foi extraída uma série temporal D que representa a diferença de desempenho entre duas esquipes, em relação a um determinado evento e. Formalmente, considere S_h o *streaming da partida* para o time da casa, e S_a para

o time visitante, pode-se calcular $D = S_h - S_a$. A Tabela 5.2 apresenta os resultados dessa transformação em relação ao exemplo da Tabela 5.1.

Apesar de D ser uma boa representação, as séries temporais geradas não capturam todas as informações sobre o andamento da partida. Por exemplo, suponha que em uma determinada partida, aos 30', o time da casa está ganhando por $1 - 0$. Ao mesmo tempo, em outra partida, a equipe da casa ganha por $3 - 2$. Em ambos os casos, o valor de $D^{[30]}$ é igual a 1, embora os placares das partidas tenham progredido de forma diferente. Para resolver esse problema, extraímos, para cada evento, a série temporal R que inclui a proporção do número de eventos associados à equipe da casa, em relação ao número total de eventos. Formalmente, se para um determinado minuto i, $S_h^{[i]} = S_a^{[i]}$, então $R^{[i]} = 0.5$, caso contrário, $R^{[i]} = \frac{S_h^{[i]}}{S_h^{[i]} + S_a^{[i]}}$. A Tabela 5.3 apresenta os resultados da transformação em relação ao exemplo na Tabela 5.1.

Por fim, sabe-se que os classificadores de aprendizagem de máquina podem funcionar melhor com dados padronizados (normalizados). Assim, os dados do *streaming de odds* foram transformados em probabilidades. Para isso, foram aplicados os conceitos da normalização básica, explicados na Seção 2.3.1, para realizar a transformação.

	Minutos									
Features	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Gols - Time da Casa	0	0	1	1	1	1	1	2	2	2
Gols - Time Visitante	0	0	0	0	0	1	1	1	1	1
Escanteios - Time da Casa	0	0	0	1	2	2	2	3	3	3
Escanteios - Time Visitante	0	1	1	1	1	1	1	1	1	1

Tabela 5.1: Exemplo de *streaming da partida* para alguns eventos

5.2 Análise Exploratória

Antes de iniciar o processo de modelagem foi relizada uma cuidadosa análise exploratória para conhecer melhor o conjunto de dados. Parte dessa análise foi focada principalmente em verificar se a ordem ou temporalidade dos eventos tem alguma influência no resultado final

	Minutos									
Features	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Gols	0	0	1	1	1	0	0	1	1	1
Escanteios	0	-1	-1	0	1	1	1	2	2	2

Tabela 5.2: Exemplo de transformação de *streaming da partida*: diferença no número de eventos (Notação: D)

	Minutos									
Features	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Gols	0.5	0.5	1.0	1.0	1.0	0.5	0.5	0.66	0.66	0.66
Escanteios	0.5	0.0	0.0	0.5	0.66	0.66	0.66	0.75	0.75	0.75

Tabela 5.3: Exemplo de transformação de *streaming da partida*: proporção no número de eventos associados ao time da casa em relação ao número total de eventos ocorridos (Notação: D)

da partida. Se esse aspecto temporal/sequencial for relevante, então é importante constuir modelos que possam tirar proveito desses padrões.

Para analisar esse comportamento, explorou-se inicialmente o seguinte problema: *Um time A está vencendo por 1-0 e sfore o gol de empate de um time B. O fato de ter empatado a partida, aumenta as chances do time B vencer?*

A Figura 5.1 apresenta a quantidade de jogos em que o placar estava em 1-1 em um determinado minuto. A linha azul representa a quantidade de jogos em que o *time que marcou primeiro* venceu, enquanto a linha laranja representa a quantidade de jogos que o *time que empatou* venceu. É possível observar que para o primeiro tempo não há diferença significativa, pois as linhas seguem um padrão semelhante. Isso significa que, se o empate acontece ainda no primeiro tempo, a ordem dos gols parece não ter relevância, pois ambas as equipes permanecem com chances iguais de vencer.

Já a partir do segundo tempo, um novo padrão começa a aparecer. Os times que empataram, acabam vencendo mais partidas. Essa diferença fica mais acentuada principalmente quando o empate ocorre entre os minutos 70 e 85. Uma hipótese para essa diferença pode

estar no aspecto motivacional das equipes. É possível que as equipes que estavam perdendo e reagiaram acabem ganhando uma motivação extra para os minutos finais. Assim, para esse problema, há sim uma relevância, ainda que discreta, na ordem e no tempo que os gols ocorreram.

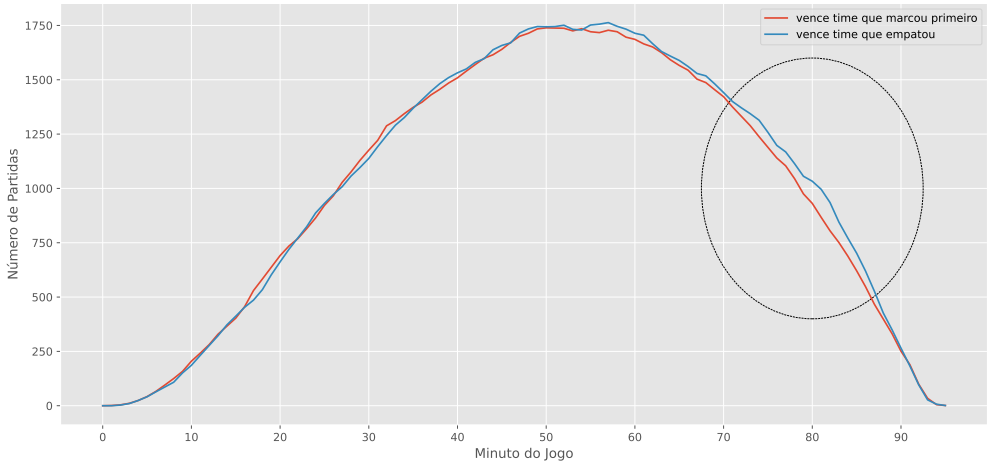


Figura 5.1: Quantidade de jogos em que o placar estava 1-1 em cada minuto

Para reforçar essa primeira impressão, foi analisada uma segunda situação, considerando os jogos em que o placar está 2-1. Esse placar pode ser construído de três formas:

- **Equilíbrio**: O *time A* faz 1-0, o *time B* empata 1-1 e o *time A* volta a marcar, 2-1.
- **Reação**: O *time A* abre 2-0 e depois o *time B* diminui para 2-1.
- **Virada**: O *time B* faz 1-0 e o *time A* vira o jogo para 2-1.

Dentro desse contexto, pode-se explorar o seguinte problema: *Dado que o placar está 2-1, as chances da equipe A vencer dependem de como o placar foi construído?*

A Figura 5.2 mostra as situações em que houve *vitória do time A* ou *empate*, para cada um dos cenários mencionados acima. Percebe-se que, quando há uma virada do *time A* até por volta dos 35 minutos, ele demonstra ter uma chance maior de vencer a partida. Também percebe-se que, quando há uma reação do *time B* após o minuto 70, as chances do *time A* vencer diminuem.

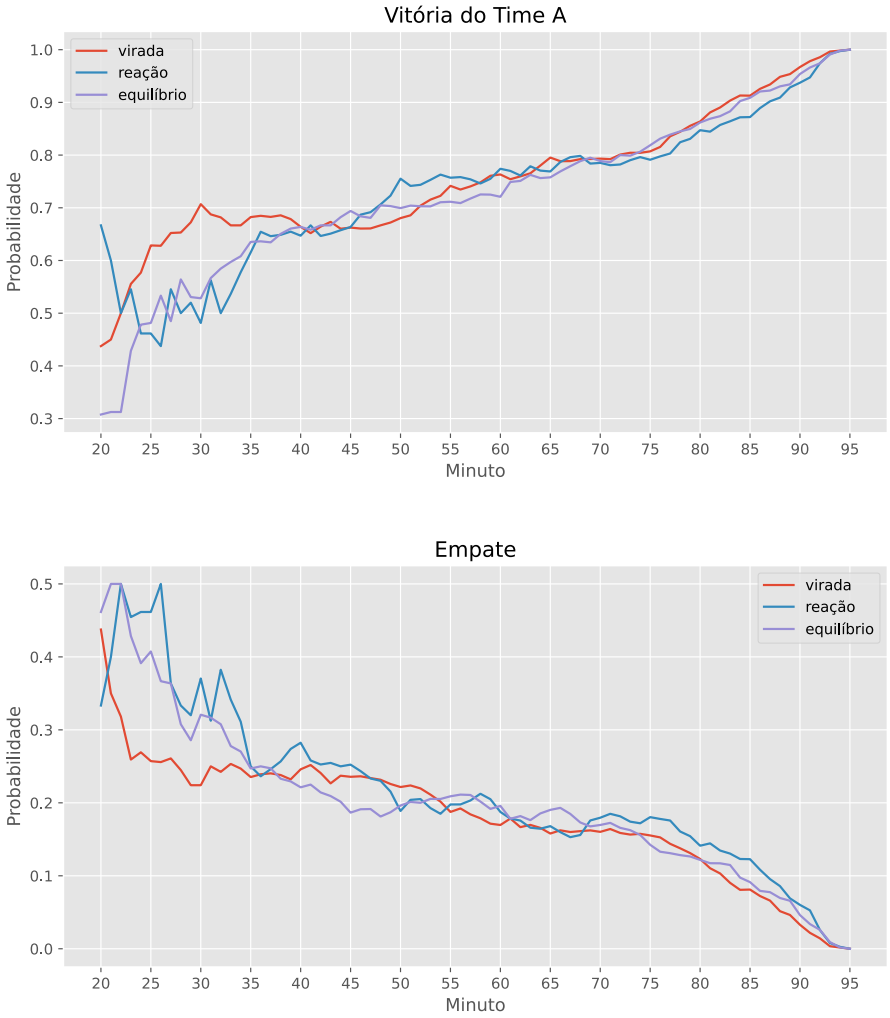


Figura 5.2: Chances de vitória do time que está vencendo e de empate, em cada minuto, para cada cenário: virada, reação e equilíbrio, quando o jogo está 2-1

As análises apresentadas confirmam os indícios de que a ordem e os minutos dos eventos podem ser relevantes para a construção de modelos. Outras análises secundárias como a importância do tempo em que ocorre um cartão vermelho e o tempo que ocorre o primeiro gol, também apontam para conclusões semelhantes. Em resumo, espera-se que técnicas que aprendem a partir da série temporal dos eventos de uma partida consigam fazer predições melhores do que técnicas que aprendem a partir da frequência acumulada. Por exemplo, considere um *agente A* que vai prever o resultado final a partir da informação de que o jogo está 1-1 aos 70 minutos de jogo. Considere um *agente B* que, além dessa informação, vai

saber quem marcou o gol primeiro e em qual minuto o gol foi marcado. Pelas análises realizadas, espera-se que o *agente B* consiga fazer uma predição mais acurada.

5.3 Modelos

Nesta seção, serão apresentados os classificadores experimentados neste trabalho e a motivação para essas escolhas.

5.3.1 Modelos baseados no estado do jogo

Neste trabalho, o *estado do jogo* é a representação de uma partida em determinado minuto, ou seja, um resumo de tudo o que aconteceu até aquele momento. Formalmente, para um determinado minuto i, define-se um estado de jogo $x^{[i]}$ como o vetor dos valores indexados ao índice (minuto) i nas séries temporais (D e R), além do tempo restante de jogo t, ou seja, $x^{[i]} = \{D^{[i]}, R^{[i]}, t\}$. Dessa forma, a partir de um conjunto dados que contém vários estados de jogo, é possível treinar um único modelo capaz de prever o resultado final de uma partida em qualquer minuto. Neste trabalho, essa estratégia é chamada de **modelo único (MU)**. A Figura 5.3 apresenta um exemplo, no qual cada vetor usado para treinamento representa um *estado do jogo* e esses estados são usados para treinamento de um modelo único.

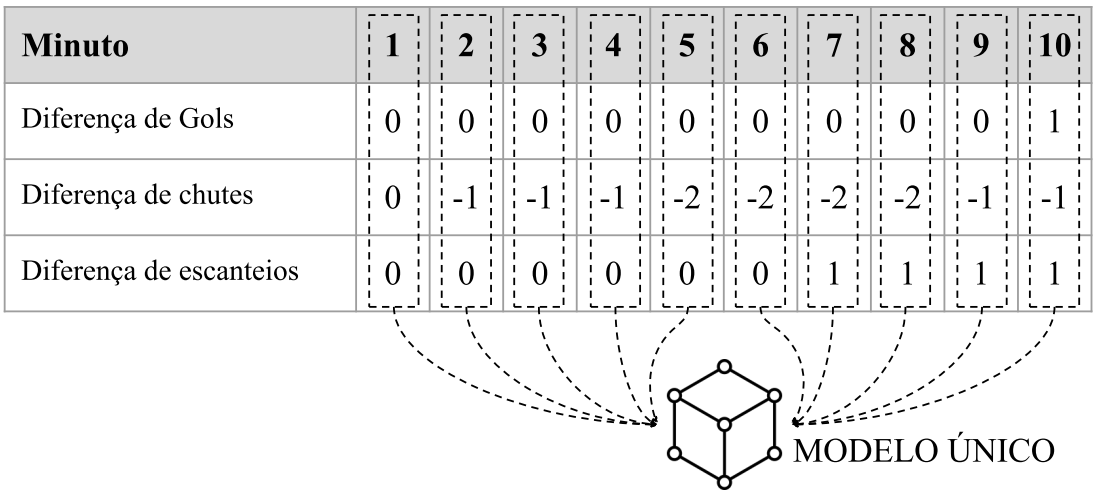


Figura 5.3: Exemplo de construção de um modelo único que faz previsões para qualquer minuto de jogo

Uma segunda estratégia é remover o recurso t (que representa o tempo) do vetor de estado do jogo e treinar um modelo separado para cada minuto. Observe que, neste cenário, haverá uma quantidade de modelos igual ao número de minutos de uma partida. Neste trabalho, essa estratégia é chamada de **modelos múltiplos** (MM). A Figura 5.4 apresenta um exemplo de múltiplos modelos, em que para cada minuto, um modelo é treinado individualmente.

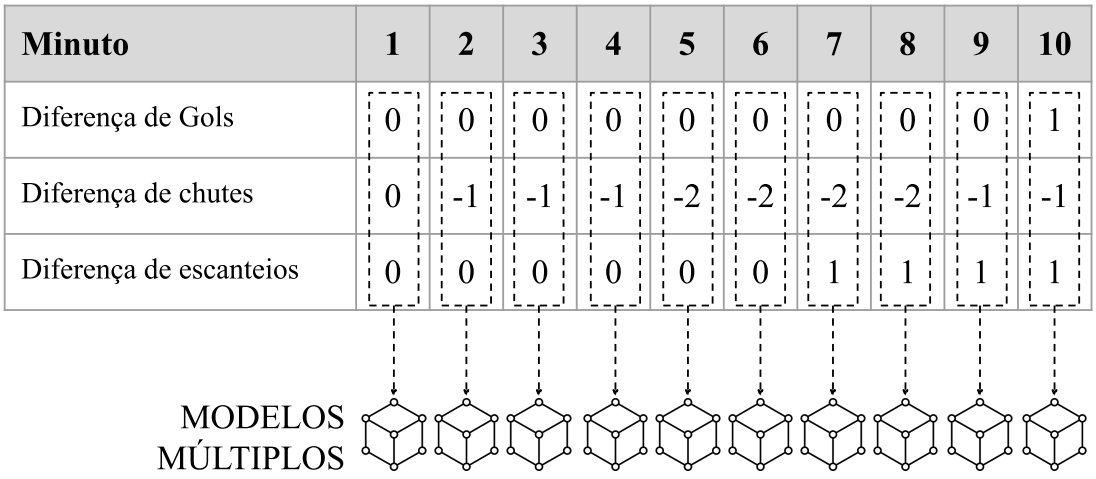


Figura 5.4: Exemplo de construção de múltiplos modelos que fazem previsões para cada minuto de jogo individualmente

Neste trabalho, para as duas estratégias, são avaliados três classificadores que foram usados com sucesso por trabalhos relacionados sobre previsão de futebol: *Gaussian Naive Bayes* (GNB), um classificador probabilístico simples, mas rápido, baseado no teorema de Bayes [111] [59]; Regressão Logística (RLG), um modelo linear popular e poderoso [71] [51]; e *Gradient Boosting* (XGB), um classificador *ensemble* que apresenta ótimo desempenho em cenários de predição pré-jogo [15] [93]. Essas estratégias também foram consideradas e avaliadas por Robberechts et al. [71].

5.3.2 Modelos baseados no progresso do jogo

Ao assistir a partidas de futebol, é natural que o público faça previsões de acordo com os acontecimentos observados no decorrer das partidas. O desempenho das equipes costuma oscilar e o controle do jogo pode "mudar de mãos"várias vezes. Para os desportistas, perceber essas variações parece trazer benefícios às previsões.

Embora o *estado do jogo* seja uma boa representação do progresso da partida, contendo um resumo do que aconteceu até um determinado minuto, ele não captura a ordem e o tempo dos eventos. Por exemplo, em uma partida de $1-1$, é interessante saber: qual time marcou o último gol? Este gol foi marcado recentemente ou no início da partida? Qual time tem atacado mais nos últimos minutos? Na análise exploratória realizada, foram demonstrados indícios de que essas informações são relevantes para refinar a predição.

Assim, é possível adotar uma estratégia para carregar os modelos com todos os dados de *streaming* minuto a minuto, transformando a tarefa em um problema de classificação de múltiplas séries temporais. Nesse caso, o progresso da partida pode ser representado formalmente por $Z^{[i]} = \{D^{[0,i]}, R^{[0,i]}\}$.

Nesse contexto, é possível construir modelos de duas maneiras. A primeira forma, leva em consideração que as séries temporais têm tamanhos diferentes para cada minuto de jogo. Assim, pode-se adotar a opção de treinar vários modelos, um para cada minuto da partida, sem fazer nenhuma alteração no conjunto de dados inicial. Neste trabalho, essa estratégia é denominada **modelos temporais múltiplos (MTM)**. A Figura 5.5 apresenta um exemplo de construção de múltiplos modelos a partir de múltiplas séries temporais.

Minuto	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Diferença de Gols	0	0	0	0	0	0	0	0	0	1
Diferença de chutes	0	-1	-1	-1	-2	-2	-2	-2	-1	-1
Diferença de escanteios	0	0	0	0	0	0	1	1	1	1

MODELOS TEMPORAIS MÚLTIPLOS

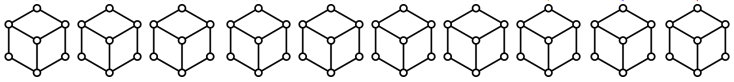


Figura 5.5: Exemplo de construção de múltiplos modelos a partir de múltiplas séries temporais.

Para a estratégia **MTM**, foram avaliadas arquiteturas de aprendizado profundo amplamente utilizadas para classificação de séries temporais multivariadas [28], a citar: *Fully*

Convolutional Network (FCN), *InceptionTime* e CNN-BiLstm. A FCN foi proposta pela primeira vez por Wang et al. [29] e são o estado da arte para vários problemas de outros domínios [28]. *InceptionTime* é um conjunto de modelos de Redes Neurais Convolucionais (CNN) profundas inspirado na arquitetura Inception-v4. Foi proposto por Fawaz et al. [1], que mostrou que esse modelo poderia ser mais preciso e mais rápido que o estado da arte.

Por fim, este trabalho usou uma combinação de ajuste manual e *gridsearch*, visando encontrar a melhor arquitetura e os melhores hiperparâmetros para o problema. Nos experimentos prévios, os melhores resultados foram obtidos usando uma CNN seguida por um LSTM bidirecional. Essa arquitetura também teve bons resultados em outros domínios, como no reconhecimento de ações em vídeos [112] [113].

A combinação de CNN com LSTM requer um *design* específico, uma vez que cada arquitetura possui características e pontos fortes próprios. A CNN é conhecida por sua capacidade de extrair o máximo de recursos possível das séries temporais. A LSTM mantém a ordem cronológica entre os eventos nas duas direções, tendo a capacidade de ignorar eventos irrelevantes.

Em resumo, o *CNN-BiLstm* otimizado para a predição durante as partidas contém uma camada convolucional, seguido por normalização em lote *batch normalization* e ativação ReLu. A normalização em lote é aplicada para acelerar a velocidade de convergência e ajudar a melhorar a generalização. Um LSTM bidirecional sucede o bloco convolucional e precede um novo *batch normalization*. A penúltima camada é um *Global Max Pooling* seguido por uma última camada densa com uma função *softmax* contendo 3 (três) filtros.

A última estratégia visa criar um modelo único capaz de prever qualquer minuto da partida. Para este cenário, é necessário um modelo que possa aprender e prever todos os minutos da partida ao mesmo tempo (um vetor de previsões). No entanto, é importante notar que, embora na etapa de treinamento seja possível carregar o classificador com múltiplas séries temporais completas de uma partida, as previsões para cada minuto não podem usar dados "do futuro". Assim, foi treinada uma LSTM, respeitando esta restrição, conforme mostrado na Figura 5.6. Neste trabalho, essa estratégia é denominada **modelo temporal único (MTU)**.

Finalmente, para **MTU**, também foi aplicada uma abordagem de aprendizagem multitarefa (**MTU_MT**). Nessa estratégia, além de prever o resultado final, o classificador prevê o resultado para o minuto seguinte. O racional é que, ao resolver ambas as tarefas simulta-

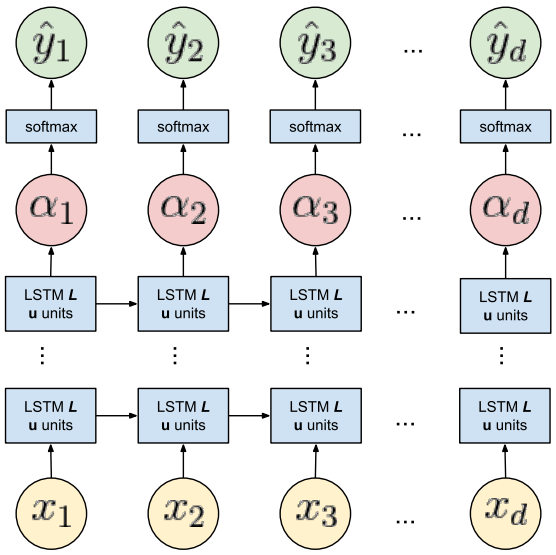


Figura 5.6: Arquitetura LSTM utilizada na abordagem de modelo único. x_t é o vetor de entrada para o minuto t, α_t é uma camada simples *feed forward*, $\hat{y}_t$ é a predição no minuto t, L representa o número de camadas, and u o número de unidades LSTM por camada

neamente, o classificador pode calibrar a previsão e melhorar a eficiência de aprendizagem para ambas as tarefas.

5.4 Experimentos

Nesta seção, são apresentados os detalhes dos experimentos a fim de responder as seguintes questões de pesquisa:

- PP4: Qual estratégia e modelo apresentaram melhor resultado para a tarefa de predição de resultados durante a partida?
- PP5: A partir de quanto tempo os modelos, usando informação coletada durante o jogo, superam a predição no mercado feita antes do jogo iniciar?
- PP6: Quão melhor ficam os modelos quando carregados com informações adicionais obtidas do mercado de apostas em comparação com modelos que são carregados com informações apenas do jogo?

- PP7: O classificador mais acurado é útil para desenvolvimento de uma estratégia lucrativa?

5.4.1 Divisão dos Dados e Configurações

O conjunto de dados foi dividido da seguinte forma: um conjunto de treinamento com 7.532 partidas e um conjunto de teste com as 1.884 restantes. Essa divisão respeita a ordem cronológica das partidas, ou seja, as partidas do conjunto de teste ocorreram após as partidas do conjunto de treinamento.

Para ajustar os hiperparâmetros de *XGB*, foi usada uma pesquisa de grade aleatória (*random search*) com validação cruzada com 5-folds, 100 iterações e os seguintes valores: *n_estimators* = [0, 1, 2 ... 998, 999, 1000], *max_depth* = [1,2,3,4,5], *learning_rate* = [0,1, 0,05, 0,01, 0,001], *colsample_bytree* = [.6, .7, .8, .9,1], *subsample* = [.7, .8, .9,1], e *min_child, min_weight* = [1, 2, 3, 4].

Para treinar as redes neurais, também foi usada uma pesquisa em grade para definir o número e tamanho dos filtros, *kernels* e camadas mais adequados ao problema. Durante o treinamento, foi usado o otimizador *Adam*, com a taxa de aprendizagem definida em $1e-4$. Foi fixado o número de épocas em 1.000 e o *batch size* como 256. No entanto, para cada época, foi usado 10% do conjunto de treinamento para validação, aplicando uma parada antecipada (*early stopping*) se o desempenho da rede não melhorasse após 10 épocas.

O conjunto de dados, código-fonte e documentação podem ser encontrados no Github [3].

5.4.2 Métricas de Avaliação

Para avaliar o desempenho dos classificadores, foi usado o *Ranked Probability Score* (RPS). O RPS é uma métrica formulada por Epstein [34] que passou a ser adotada para avaliação de modelos de predição de futebol a partir de [35].

A equação do RPS é definida por:

$$RPS(p_1, ...p_{r-1}, a_1, ...a_{r-1}) = \frac{1}{r-1} \sum_{i-1}^{r-1} (\sum_{j=1}^{1} (p_j - a_j))^2 \tag{5.1}$$

[3] `https://github.com/igormago/doutorado`

		RPS					
Classificador	**Estratégia**	**15'**	**30'**	**45'**	**60'**	**75'**	**90'**
GNB	MU	0.2365	0.2103	0.1849	0.1456	0.1125	0.0743
RLO	MU	0.2029	0.1822	0.1612	0.1288	0.0965	0.0554
XGB	MU	0.2009	0.1805	0.1613	0.1307	0.0934	0.0254
GNB	MM	0.2397	0.2156	0.1933	0.1503	0.1029	0.0272
RLO	MM	0.1996	0.1784	0.1594	0.1290	0.0913	0.0252
XGB	MM	0.2030	0.1813	0.1623	0.1316	0.0929	0.0254
FCN	MTM	0.2020	0.1790	0.1600	0.1291	0.0910	0.0254
InceptionTime	MTM	0.2009	0.1790	0.1599	0.1292	0.0910	0.0251
CNN-BiLSTM	**MTM**	**0.1985**	**0.1784**	**0.1592**	**0.1283**	**0.0899**	**0.0252**
LSTM	MTU	0.2019	0.1830	0.1651	0.1312	0.0919	0.0442
LSTM	MTU-MT	0.2001	0.1812	0.1628	0.1301	0.0919	0.0356

Tabela 5.4: Desempenho dos classificadores (em termos de RPS) a cada 15 minutos

onde r é o número de resultados possíveis em uma partida (r=3 no caso de previsões de resultados de futebol), p_j é a probabilidade prevista para um determinado resultado j, ou seja, $p \in [0, 1]$ para $j = 1, 2, ..r$, e a_j indica se o resultado j ocorreu, sendo $j = 1$ se ocorreu e $j = 0$ caso contrário. Dessa forma, quanto menor o valor da métrica, melhor o desempenho. A Tabela 5.4 apresenta os resultados considerando intervalos de 15 minutos.

5.5 Discussão dos Resultados

Nesta seção, as questões de pesquisa são respondidas e discutidas a partir dos resultados dos experimentos.

5.5.1 PP4: Qual estratégia e modelo apresentaram melhor resultado para a tarefa de predição de resultados durante a partida?

Para responder **PP4**, é necessário comparar os desempenho dos classificadores em termos de RPS (ver Figuras 5.7 e 5.8). Nessas figuras, o eixo-x denota os minutos decorridos, enquanto o eixo-y representa o RPS. Prezando pela legibilidade das figuras, foram plotados apenas os melhores classificadores. Para outras comparações, pode-se consultar a Tabela 5.4.

Comparando as estratégias MU e MM, pode-se observar que, para *GNB* e *XGB*, a estratégia MU é melhor durante quase toda a partida, passando a se equivaler nos 15 minutos

finais. Em contrapartida, para *RLO*, a estratégia MM apresentou melhor desempenho. Numa comparação geral, *RLO+MM* supera *GNB+MU* e *XGB+MU*. Assim, pode-se considerar *RLO+MM* como uma linha de base bastante forte para este problema. Robberechts et al. [71], mesmo usando um outro conjunto de dados, também chegou a uma conclusão semelhante.

Entre os modelos de aprendizado profundo, *CNN-BiLstm+MTM* apresentou resultados superiores as linhas de base *FCN* e *InceptionTime*. Comparando-o com *RLO+MM* (ver Figura 5.8), é possível observar que o desempenho é similar durante a maior parte da partida. Entretanto, *CNN-BiLstm* foi ligeiramente superior em um parte específica do segundo tempo: entre os minutos 70 e 85. A partir dos 85 minutos, todos os modelos voltam a ser equivalentes.

Uma possível explicação para esse resultado é que no primeiro tempo as séries temporais são muito curtas e, portanto, a ordem/minuto dos eventos não parece ser mais informativo do que o *estado do jogo*. Na segunda metade, porém, a série temporal é maior e, após um determinado momento (70º minuto), o modelo *CNN-BiLSTM* consegue tirar proveito da estrutura sequencial de eventos. Por fim, após os 85 minutos, o final do jogo está muito próximo e, como esperado, o placar da partida se torna muito mais relevante que os outros atributos. Assim, neste ponto, os modelos tornam-se novamente equivalentes.

Uma segunda observação importante é que o intervalo do jogo em que a estratégia MTM se sobressai se assemelha aos intervalos observados na análise exploratória, no qual as informações sequenciais têm mais relevância. Dessa forma, reforça a conclusão de que nesse momento das partidas, as técnicas de aprendizagem conseguem se beneficiar do uso de séries temporais.

Em relação a estratégia MTU, embora os modelos não tenha alcançado o desempenho da estratégia MTM, é possível notar que a abordagem multitarefa MTU-MT superou a abordagem de modelo temporal único MTU. No trabalho atual, a função de custo usada para a estratégia MTU-MT foi a soma das funções de custo para cada tarefa. Nesse sentido, trabalhos futuros podem tentar melhorar essa ideia inicial, incluindo novas tarefas e encontrando uma melhor função de custo que possa considerar pesos diferentes para cada tarefa. Assim, pode-se tentar obter um modelo único que tenham desempenho tão bom quanto estratégias com múltiplos modelos.

Por fim, a resposta para **PP4** é que *CNN-BiLstm+MTM* é a melhor abordagem para este problema, mas seguido de perto por uma estratégia mais simples como a *RLO+MM*.

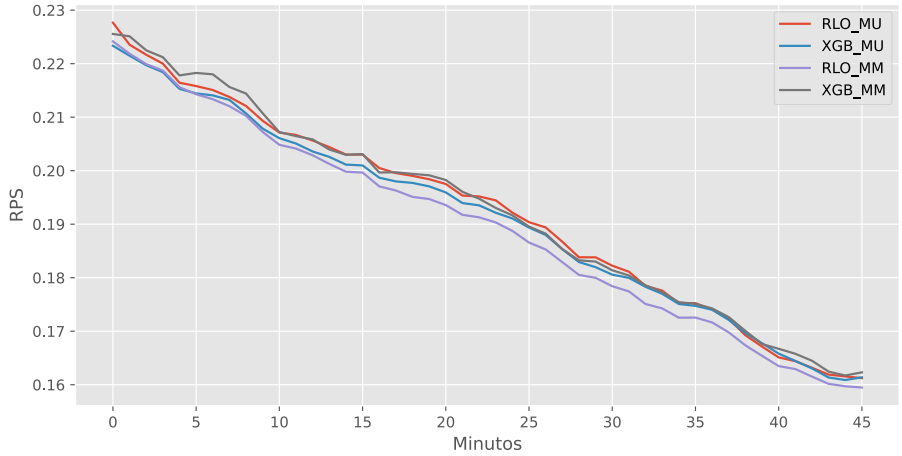
5.5.2 PP5: A partir de quanto tempo os modelos, usando informação coletada durante o jogo, superam a predição no mercado feita antes do jogo iniciar?

Para responder **PP5** (veja a Figura 5.9), pode-se comparar os melhores classificadores com as previsões do mercado de apostas antes do início das partidas. Os melhores classificadores precisaram de cerca de 14 minutos, em média, para fazer uma predição superior à predição inicial do mercado de apostas. Isso significa que, via de regra, os primeiros 14 minutos do tempo decorrido de um jogo possuem informações tão relevantes para a previsão quanto todo o conhecimento codificado pelo mercado de apostas antes do início da partida. Em outras palavras, mesmo sem qualquer informação prévia sobre a qualidade das equipes, 14 (quatorze) minutos é o tempo necessário para o classificador conhecer a força de cada equipe e poder fazer uma previsão tão precisa quanto a feita pelo mercado de apostas, que reúne todas as informações disponíveis sobre as equipes antes do início do jogo.

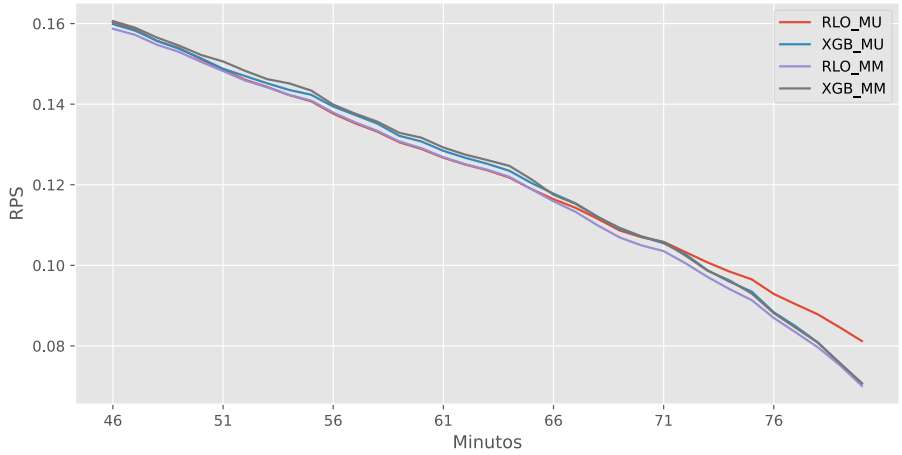
5.5.3 PP6: Quão melhor ficam os modelos quando carregados com informações do mercado em comparação com modelos que são carregados com informações apenas do jogo?

Para uma análise puramente estatística, construir modelos a partir de dados históricos é uma boa forma de entender certos padrões do jogo. Entretanto, sabe-se que há uma dificuldade natural para incluir uma diversidade de informações relevantes sobre as equipes, que muitas vezes, passam por aspectos subjetivos, como o ambiente político de um clube, motivação do elenco, jogadores fora da forma física ideal, etc. Todas essas informações podem ser úteis para uma boa análise pré-jogo e são levadas em conta pelo mercado de apostas.

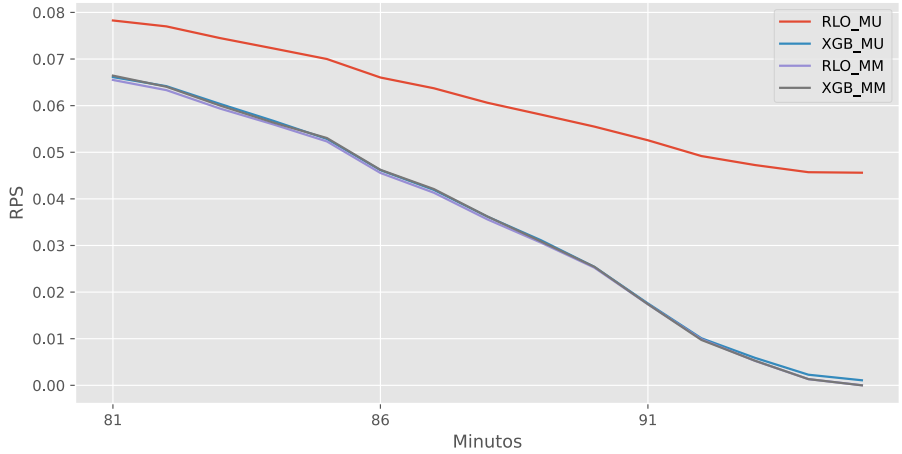
No âmbito da predição de resultados durante as partida, também há fatores importantes que não estão incluídos no conjunto de dados usado por este trabalho. Em outras palavras, apesar do *streaming da partida* ser uma excelente fonte de informação sobre o desenrolar



(a) Minutos: 0-45

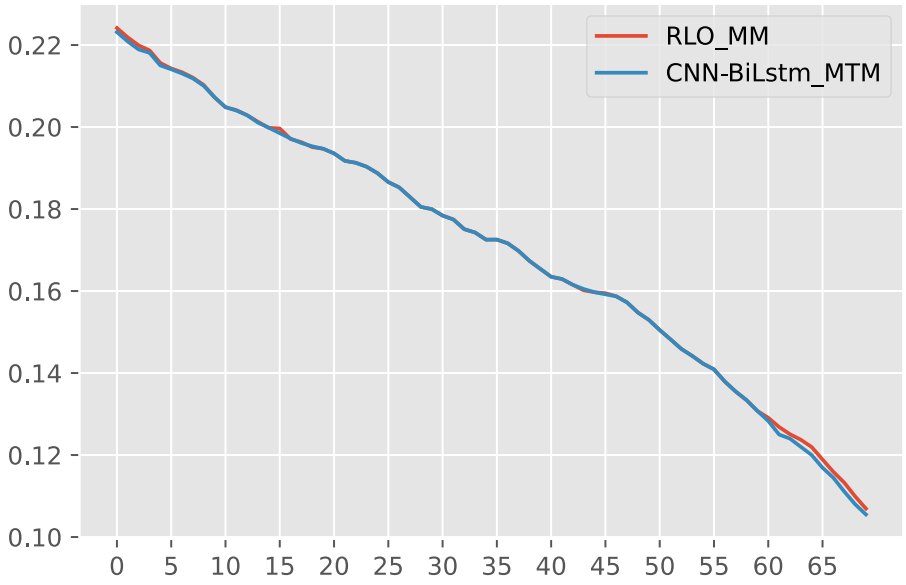


(b) Minutos: 46-80

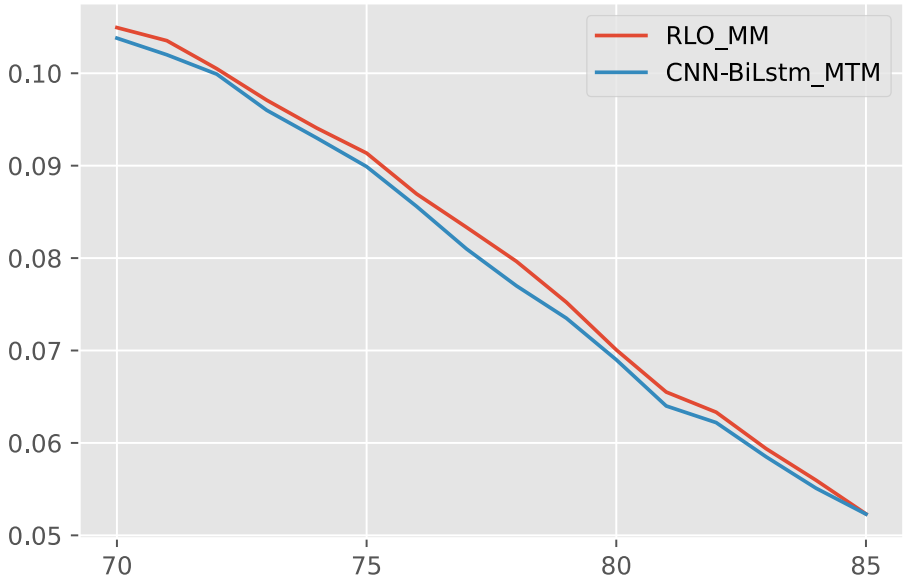


(c) Minutos: 80-96

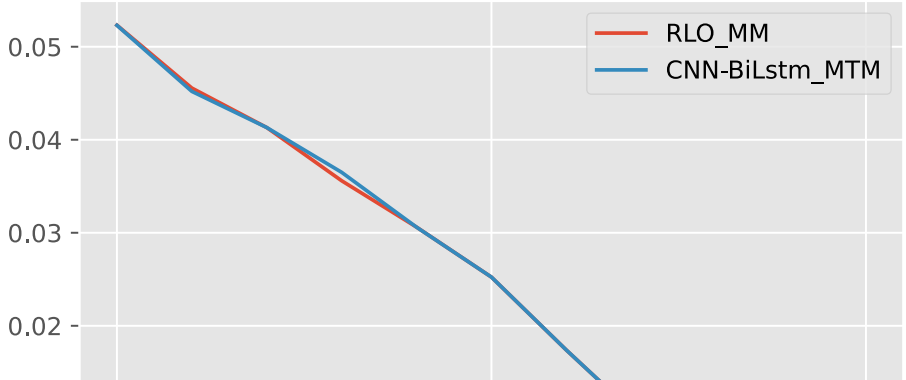
Figura 5.7: Desempenho dos classificadores usando as estratégias MU e MM (em termos de RPS)



(a) Minutos: 0-60



(b) Minutos: 60-85



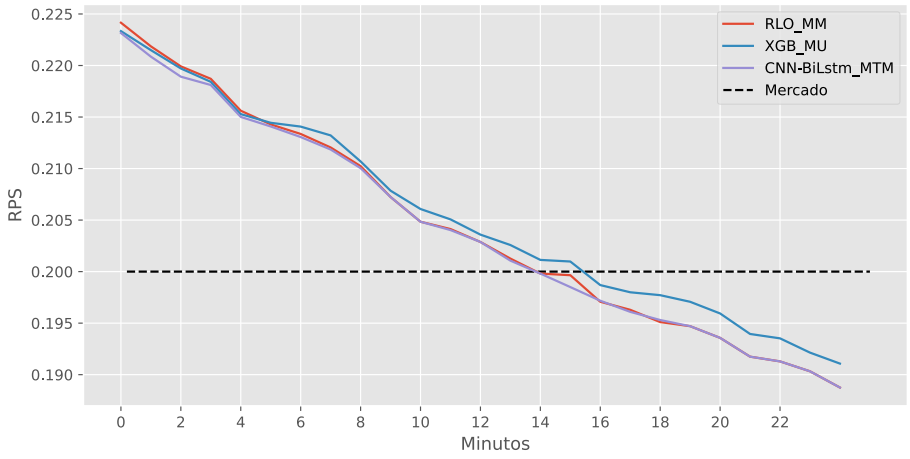


Figura 5.9: Comparando os melhores classificadores com a predição do mercado pré-jogo (A linha horizontal é o desempenho dos mercados de apostas antes do jogo)

de um jogo, ainda há uma série de informações que não está contida nele. Por exemplo, a substituição de um jogador importante por lesão, a má-atuação de determinados jogadores-chaves, jogadas bem trabalhadas que não resultam obrigatoriamente em uma estatística de finalização a gol, etc. Esses aspectos relacionados a parte tática e técnica do jogo, também pode ser levados em conta por apostadores para refinar a predição.

Nesse contexto, considerando que as *odds* do mercado de apostas podem encapsular todas essas informações extras, que não estão contidas na base de dados usada por este trabalho, pode-se usar as próprias probabilidades implícitas das *odds* como atributos adicionais para a predição. Assim, é possível verificar o quão melhor os modelos podem ficar, se usarem essas informações adicionais.

Para responder **PP6**, o modelo *CNN-BiLstm* foi retreinado, desta vez, adicionando os *streaming de odds*. A Figura 5.10 apresenta os resultados desse novo experimento. Como esperado, houve uma melhora notável na previsão, principalmente no primeiro tempo da partida. Isso acontece principalmente porque as *odds* encapsulam um conhecimento prévio sobre a força das equipes pré-jogo. Essa melhoria vai diminuindo com o passar do tempo, à medida que o classificador obtem mais informações sobre a própria partida. Assim, a partir minuto 70, as informações do mercado deixam de trazer benefício para o modelo. Isso provavelmente ocorre devido ao aprendizado sobre as equipes dentro do próprio jogo e a

proximidade do final da partida, que faz com que o placar se torne cada vez mais informativo quando comparado com os demais atributos.

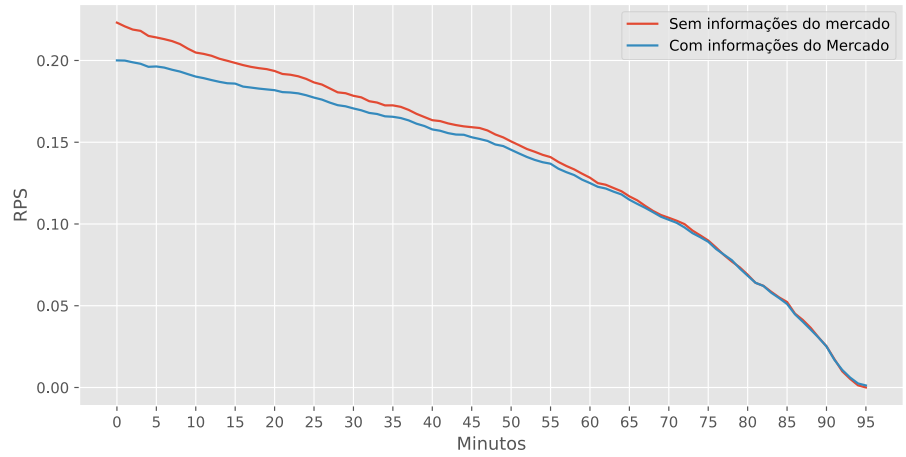


Figura 5.10: Comparando o melhor classificador com e sem informações do mercado

5.5.4 PP7: O classificador mais acurado é útil para desenvolvimento de uma estratégia lucrativa?

Assim como discutido na Seção 4.5.3, a precisão de um classificador não tem correlação direta com lucratividade no mercado de apostas. Dessa forma, ao responder PP7, busca-se compreender o desempenho do classificador no mercado de apostas, como também avaliar o grau de eficiência desse mercado.

Nesse contexto, foi realizado o seguinte experimento. Considere que para cada minuto do jogo, é realizada uma aposta de 1$ nos três resultados possíveis: o favorito, o azarão e o intermediário (nem favorito, nem azarão) - de acordo com a predição do melhor classificador *CNN-BiLstm*. A figura 5.11 apresenta os resultados.

Ao analisar os resultados, percebe-se que para apostas no resultado intermediário não houve nenhuma oportunidade de lucro significativa durante a partida. Por outro lado, para apostas no resultado favorito, há um padrão relativamente bem definido. Apostas no primeiro tempo não foram lucrativas, enquanto apostas no segundo tempo conseguiram gerar lucros, ainda que discretos. Aparentemente, as melhores oportunidades estão no começo do segundo tempo. O melhor resultado ocorreu no minuto 52, em que foi possível obter um lucro de

$79.05\$$. Considerando que foram feitas apostas em 1.984 jogos (conjunto teste), o RoI é de aproximadanete 4%.

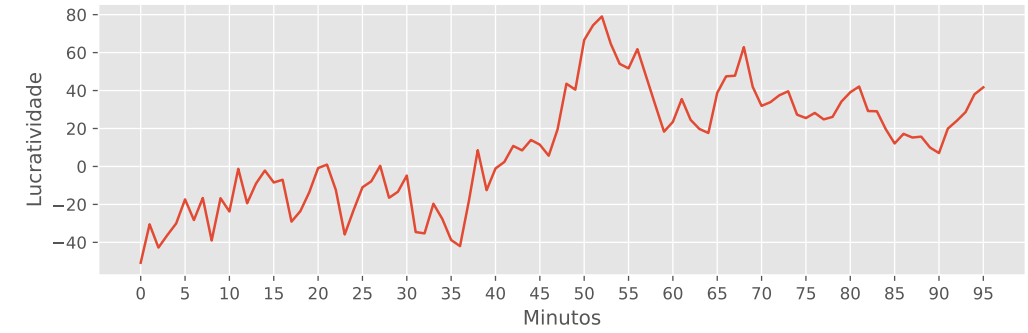
As apostas nos azarões também aparesentam uma padrão relativamente claro. Há algumas oportunidades de lucro no começo da partida. Essas oportunidades vão desaparencedo à medida que o jogo avança. No gráfico é possível ver alguns picos de lucro nos minutos 35 e 65. Entretanto, pelo padrão observado, é provável que esses picos sejam *outliers*. Um trabalho futuro pode aumentar a amostra de teste, para verificar se isso se confirma.

De forma geral, foram encontrados alguns bons indícios de que o mercado de apostas, durante as partidas, não são totalmente eficientes. Entretanto, com uma estratégia simples, que segue estritamente a predição de um bom classificador, o RoI obtido não chega ser um grande atrativo. Trabalhos futuros podem usar o conjunto de dados construídos por este trabalho para avaliar e encontrar estratégias mais inteligentes de apostas, que possam vir a gerar um retorno mais significativo sobre o investimento.

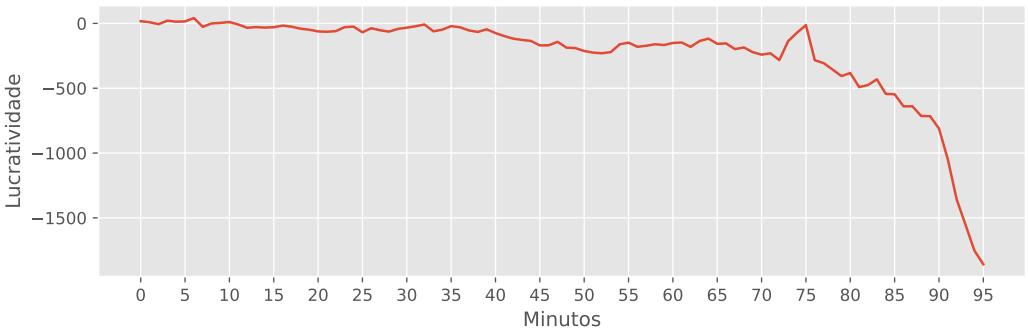
5.6 Considerações Finais

Neste Capítulo, foi apresentado um dos primeiros *benchmarks* para a previsão de resultados de futebol durante a partida. Foi avaliada uma ampla gama de modelos diferentes sob diferentes estratégias, incluindo redes neurais profundas para séries temporais multivariadas. Dentre as principais descobertas estão:

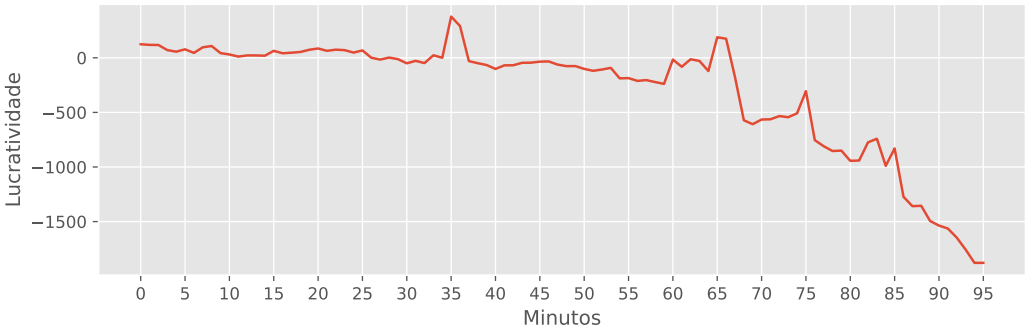
- A previsão de resultados de futebol durante a partida é um problema difícil, no qual linhas de base simples, como modelos treinados minuto a minuto, são muito difíceis de serem batidos (PP4);
- Classificadores alimentados com informações minuto a minuto precisam de cerca de 14 minutos para superar as melhores previsões feitas antes da partida (PP5);
- Informações do mercado são úteis para melhorar o desempenho dos classificadores, principalmente no início das partidas (PP6).
- Há bons indícios de que o mercado de apostas não é totalmente eficiente durante as partidas (PP7).



(a) Apostas no favorito



(b) Apostas no intermediário



(c) Apostas no azarão

Figura 5.11: Comparando a lucratividade em diferentes cenários de aposta

Capítulo 6

Conclusão

Neste trabalho, o problema central é a predição de resultados em futebol, tema bastante discutido pela ciência nas últimas décadas. A tarefa de predição foi explorada a partir de duas visões: antes do jogo iniciar (pré-jogo) e durante o jogo.

Para a predição pré-jogo foi escolhido um alvo de predição ainda pouco explorado na literatura: a predição de "ambos marcam- problema em evidência no mercado de apostas. Para esse fim, foi necessária a construção de um *dataset* contendo dados históricos de partidas e cotações de casas de apostas, envolvendo nove campeonatos nacionais. Esse *dataset* passou por um cuidadoso processo de *feature engineering*, em que o autor deste trabalho aplicou seu conhecimento sobre o domínio para gerar um conjunto de atributos relevante para o processo de aprendizagem. A partir desse conjunto, diferentes classificadores de aprendizagem de máquina foram treinados e comparados com outros modelos relevantes da litetura (baseados em Poisson), bem como com o mercado de apostas. A partir desses experimentos, algumas perguntas de pesquisa puderam ser respondidas, como apresentado na Tabela 6.1.

Em linhas gerais, devido à imprevisibilidade inerente ao futebol, a tarefa se mostrou bastante difícil. Fica a percepção de que, em um trabalho futuro, pode-se incluir novos tipos de dados para tentar trazer mais informações para os modelos. Entretanto, em termos de acurácia de predição, não deve haver muita margem para melhoria. Essa percepção pode ser corroborada principalmente quando os classificadores são usados como base para estratégias de apostas. De uma forma geral, o mercado se mostrou eficiente em muitos cenários, o que pode sugerir que na maioria das vezes o conhecimento gerado pelos classificadores se assemelha ao conhecimento usado pelo mercado de apostas. Ainda assim, foi possível encontrar

PP1. Quão difícil é o problema de predição de ambas marcam? Os classificadores de aprendizagem de máquina, apesar de superarem alguns modelos propostos pela literatura, foram, no geral, apenas um pouco melhores que classificadores simples baseados na classe majoritária. Assim, pode-se confirmar a dificuldade da tarefa.
PP2. Os classificadores avaliados são capazes de superar as casas de apostas? Considerando as probabilidades médias do mercado, os melhores classificadores não foram capazes de superar. Entretanto, considerando a casa de apostas que oferece as odds mais justas (melhores cotações), o *Gradient Boosting* conseguiu ser um pouco superior.
PP3. Os classificadores são úteis para desenvolver estratégias de apostas lucrativas? Considerando as *odds* médias do mercado, nenhum classificador foi capaz de obter lucro com as estratégias avaliadas neste trabalho. Entretanto, quando considerada apenas a casa de apostas mais justa e quando o número de apostas é limitado (por campeonato ou por um intervalo de probabilidades), foi possível conceber algumas estratégias de apostas lucrativas.

Tabela 6.1: Perguntas de Pesquisas referentes à predição pré-jogo

algumas situações (intervalo de odds) em que houveram oportunidades de lucro. Nesse cenário, os classificadores, além de terem usado informações referentes ao desempenho das equipes, usaram informações advindas do próprio mercado. Essa combinação de atributos se mostrou crucial para a obtenção de uma estratégia lucrativa. Assim, este trabalho obteve bons indícios de que existem cenários de ineficiência de mercado, em alguns intervalos de *odds*, que podem serem explorados.

Para a predição de resultados durante a partida, este trabalho apresenta um dos primeiros *benchmarks* para o problema. Além disso, constrói o primeiro conjunto de dados estruturado que inclui eventos de partidas e *odds* do mercado de apostas, minuto-a-minuto. Além de ser um *dataset* rico para o estudo da predição de resultados e eficiência de mercado, pode ser relevante para diversos estudos que envolvem classificação de séries temporais. A partir desse *dataset*, foi possível avaliar diferentes estratégias de predição e diferentes classificadores. A partir desses experimentos, algumas perguntas de pesquisa puderam ser respondidas, como apresentado na Tabela 6.2.

Em uma análise final, este trabalho se mostra um ótimo ponto de partida para trabalhos futuros, deixando a percepção de que há margem para encontrar melhores modelos para o problema, a partir das estratégias apresentadas. Algumas dessas ideias de trabalhos futuros são discutidas na seção a seguir.

6.1 Trabalhos Futuros

A predição de resultados pré-jogo já é bastante discutida na literatura, entretanto ainda há uma diversidade de caminhos a serem explorados. Este trabalho indica as seguintes possibilidades de continuação:

Explorar novos alvos. A predição pré-jogo está quase sempre voltada para o resultado final da partida. Entretanto, atualmente, há muito interesse em outros alvos, principalmente por conta dos mercados de apostas. Este trabalho abordou um mercado ainda pouco explorado na literatura: a predição de "ambas marcam", no qual o objetivo é prever se ambas as equipes irão marcar gols em uma partida, ou não. Assim como o mercado de "ambas marcam", existem outros alvos que podem ser explorados como, por exemplo: o mercado *over/under*, no qual o objetivo é predizer se uma partida terá mais ou menos gols que uma determinada margem; o mercado de "*handicap*", no qual o objetivo é predizer se uma determinada equipe irá vencer por mais gols que uma determinada margem; entre outros.

Explorar novos tipos de dados. A grande maioria dos trabalhos avaliados levam em conta o histórico de resultados de partidas. Entretanto, no futebol não é raro que uma equipe de qualidade inferior acabe vencendo uma equipe de qualidade superior. Nesse contexto, pode-se considerar que os resultados das partidas, por si só, podem não ser tão representativos para medir a força de uma equipe. Uma alternativa para refinar essa medição é a utilização das estatísticas do jogo (*scouts*). Além disso, na Internet há uma diversidade de informações detalhadas sobre as equipes e jogadores. Assim, a exploração de novos atributos pode ser relevante para acrescentar detalhes não capturados pelos resultados das partidas. Algumas sugestões para aquisição de novos dados:

1. Pappalardo et al. [114] - Dados espaço-temporais de eventos em partidas de futebol;

2. *football-data.co.uk*[1] - *Scouts* de partidas;

3. *fifaindex.com*[2] - Informações sobre jogadores usadas em jogos de videogame;

4. *totalcorner.com*[3] - Dados detalhados sobre cotações de casas;

Propor novas ideias para modelagem de atributos e/ou usar aprendizagem profunda. As ideias aplicadas para a modelagem de atributos são fundamentais para o bom funcionamento de alguns classificadores. Nesse cenário, é preciso ter um bom conhecimento de domínio para a criação de atributos. Assim, propor novas estratégias de modelagem de atributos é sempre um caminho válido a ser explorado. Outra possibilidade é aplicar arquiteturas profundas de redes neurais destinadas a trabalhar com "dados brutos".

Otimizar modelos para o mercado de apostas. Neste trabalho, os modelos foram treinados para fazer predição de resultados e essas predições foram utilizadas para abastecer estratégias de apostas. Em um cenário focado especificamente para apostas esportivas, o modelo pode ser treinado diretamente visando a obtenção de lucros. Nesse caso, os modelos seriam treinados para "realizar"apostas e não para prever resultados.

A predição de resultados durante as partidas ainda é um assunto pouco explorado. Todas as ideias sugeridas para predição pré-jogo podem ser adaptadas para predição durante o jogo. Além dessas, este trabalho desperta para as seguintes possibilidades de continuação:

Explorar novas arquiteturas de redes neurais. A classificação de múltiplas séries temporais é uma área que está em constante evolução. Assim, é possível continuar avaliando novas estruturas de redes neurais que possam vir a superar os resultados alcançados neste trabalho. Entre as arquiteturas recentes com bons resultados, em outros domínios, que não foram exploradas neste trabalho, destacam-se: Transformes [115] e TapNet [116].

Construir um modelo único. A estratégia deste trabalho que apresentou melhor resultado utiliza múltiplos modelos treinados individualmente para cada minuto do jogo. Dessa forma, uma possível alternativa é desenvolver uma arquitetura única que possa prever para qualquer minuto do jogo com a mesma eficiência dos múltiplos modelos. A utilização de modelos multi-tarefa também se mostrou promissora e pode ser uma alternativa a ser aperfeiçoada.

[1] `https://www.football-data.co.uk/`
[2] `https://www.fifaindex.com/`
[3] `https://www.totalcorner.com/`

Avaliar a eficiência do mercado de apostas dentro do jogo. Focando exclusivamente no mercado de apostas é importante avaliar a eficiência do mercado de apostas durante a partida. Isso pode ser feito avaliando diferentes momentos do jogo ou então tentando criar uma estratégia lucrativa a partir das predições dos classificadores. O resultado pode ajudar a responder perguntas como: Qual o melhor momento para apostar durante uma partida? Seria melhor apostar logo após a ocorrência de gols? Seria melhor apostar a partir de um determinado minuto? Ou quando a *odd* atingir um determinado valor? Essas e muitas outras perguntas podem ser investigadas.

Por fim, em termos de aplicação, pode-se disponibilizar os melhores modelos em um serviço real que possa, a partir de eventos coletados das partidas, fazer predições em tempo real. Essa aplicação pode ser interessante para apostadores, imprensa esportiva e fãs em geral.

PP4: Qual estratégia e modelo apresentaram melhor resultado para a tarefa de predição de resultados durante a partida? Uma rede neural profunda, de arquitetura *CNN-BiLstm*, treinada individualmente para cada minuto, foi o melhor modelo para este problema. Entretanto, foi seguido de perto por um classificador simples como o de regressão logística.
PP5: A partir de quanto tempo os modelos, usando informações coletadas durante o jogo, superam a predição no mercado feita antes do jogo iniciar? Os melhores classificadores precisaram de cerca de 14 minutos, em média, para superar o mercado de apostas pré-jogo.
PP6. Quão melhor ficam os modelos quando carregados com informações adicionais obtidas do mercado em comparação com modelos que são carregados com informações apenas do jogo? Há uma melhora substancial na previsão, principalmente no primeiro tempo da partida. Essa melhoria vai desaparecendo com o passar do tempo, à medida que o classificador obtém mais informações sobre a própria partida. Assim, a partir do minuto 70, as informações prévias sobre as equipes deixam de ser relevantes. Por fim, a partir do minuto 82, o uso de informações prévias começa a piorar o classificador. Aparentemente, para essa parte final da partida, as informações adquiridas apenas durante o jogo parecem mais relevantes do que informações pré-jogo.
PP7. O classificador mais acurado é útil para desenvolvimento de uma estratégia lucrativa? As predições do classificador permitiram a criação de uma estratégia lucrativa, a partir de apostas realizadas principalmente no início do segundo tempo. Apesar do lucro discreto, os resultos apresentam bons indícios de que o mercado de apostas em futebol não é totalmente eficiente durante toda a partida.

Tabela 6.2: Perguntas de Pesquisas referentes à predição durante as partidas

Bibliografia

[1] Hassan Ismail Fawaz, Benjamin Lucas, Germain Forestier, Charlotte Pelletier, Daniel F Schmidt, Jonathan Weber, Geoffrey I Webb, Lhassane Idoumghar, Pierre-Alain Muller, and François Petitjean. Inceptiontime: Finding alexnet for time series classification. *Data Mining and Knowledge Discovery*, 34(6):1936–1962, 2020.

[2] Arvind T Mohan and Datta V Gaitonde. A deep learning based approach to reduced order modeling for turbulent flow control using lstm neural networks. *arXiv preprint arXiv:1804.09269*, 2018.

[3] Cadeias de markov – wikipédia, a enciclopédia livre. `https://pt.wikipedia.org/wiki/Cadeias_de_Markov`. (Acessado em 01/08/2018).

[4] Betfair. Betfair exchange | best odds online, back and lay betting. (Acessado em 02/08/2018).

[5] Github - durtal/betfair: R package for the betfair api. `https://github.com/durtal/betfaiR`. (Acessado em 01/08/2018).

[6] FIFA. Fifa - fifa.com. `www.fifa.com/`, 07 2017. (Acessado em 29/07/2018).

[7] M. J. Moroney. Facts from figures 1956. *I am also greatly indebted, for the more general parts of this discussion, to LHC Tippett, Statistics*, page 3, 1943.

[8] Christopher Anderson and David Sally. *The numbers game: Why everything you know about soccer is wrong*. Penguin, 2013.

[9] Raquel Aoki, Renato M. Assuncao, and Pedro O. S. Vaz de Melo. Luck is hard to beat: The difficulty of sports prediction. In *Proceedings of the 23rd ACM SIGKDD*

International Conference on Knowledge Discovery and Data Mining, pages 1367–1376. ACM, 2017.

[10] Leonard Mlodinow. *The drunkard's walk: How randomness rules our lives*. Vintage, 2009.

[11] Marcelo Leme de Arruda. Previsão de resultados em partidas de futebol. `http://www.chancedegol.com.br/previsao2.pdf`, 2013. (Acessado em 31/07/2018).

[12] EURO 2016 Prediction Competition. Euro 2016 prediction competition. `https://eu16prediction.cs.kuleuven.be/`. (Acessado em 01/08/2018).

[13] Machine Learning. Machine learning, special issue on machine learning for soccer - springer. `https://link.springer.com/journal/10994/topicalCollection/AC_5423abcd021c04f2678eea3a27580457/page/1`. (Acessado em 01/08/2018).

[14] Anthony Costa Constantinou. Dolores: A model that predicts football match outcomes from all over the world. *Machine Learning*, pages 1–27, 2018.

[15] O. Hubáček, G. Šourek, and F. Železnỳ. Learning to predict soccer results from relational data with gradient boosted trees. *Machine Learning*, pages 1–19, 2018.

[16] Heung-Pyo Lee, Paul Kyuman Chae, Hong-Seock Lee, and Yong-Ku Kim. The five-factor gambling motivation model. *Psychiatry research*, 150(1):21–32, 2007.

[17] Terra. Mercado brasileiro de apostas on-line pode movimentar r$ 6,7 bilhões ao ano. `https://goo.gl/zPxuXs`. (Acessado em 01/08/2018).

[18] BBC. Football betting - the global gambling industry worth billions - bbc sport. `https://www.bbc.com/sport/football/24354124`. (Acessado em 01/08/2018).

[19] Cbn - a rádio que toca notícia - mercado de apostas esportivas atrai milhares de brasileiros. `https://cbn.globoradio.globo.com/media/audio/331648/`

mercado-de-apostas-esportivas-atrai-milhares-de-br.htm. (Acessado em 24/02/2021).

[20] Lars Magnus Hvattum and Halvard Arntzen. Using elo ratings for match result prediction in association football. *International Journal of forecasting*, 26(3):460–470, 2010.

[21] Anthony Costa Constantinou and Norman Elliott Fenton. 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*, 9(1):37–50, 2013.

[22] Burton G Malkiel and Eugene F Fama. Efficient capital markets: A review of theory and empirical work. *The journal of Finance*, 25(2):383–417, 1970.

[23] Betsonly. Most popular football betting markets | betsonly.com. http://www.betsonly.com/sport-betting/popular-football-betting-markets, 05 2017. (Acessado em 02/02/2021).

[24] V. A. Padilha. mineracaodadosbiologicos-parte5.pdf. https://edisciplinas.usp.br/pluginfile.php/4125431/mod_resource/content/2/mineracaodadosbiologicos-parte5.pdf. (Acessado em 01/08/2018).

[25] Bayes' theorem (stanford encyclopedia of philosophy). https://plato.stanford.edu/entries/bayes-theorem/. (Acessado em 19/02/2021).

[26] Rahul Baboota and Harleen Kaur. Predictive analysis and modelling football results using machine learning approach for english premier league. *International Journal of Forecasting*, 2018.

[27] Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep learning*, volume 1. MIT press Cambridge, 2016.

[28] Hassan Ismail Fawaz, Germain Forestier, Jonathan Weber, Lhassane Idoumghar, and Pierre-Alain Muller. Deep learning for time series classification: a review. *Data Mining and Knowledge Discovery*, 33(4):917–963, 2019.

[29] Zhiguang Wang, Weizhong Yan, and Tim Oates. Time series classification from scratch with deep neural networks: A strong baseline. In *2017 International joint conference on neural networks (IJCNN)*, pages 1578–1585. IEEE, 2017.

[30] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

[31] Sepp Hochreiter. The vanishing gradient problem during learning recurrent neural nets and problem solutions. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 6(02):107–116, 1998.

[32] Sepp Hochreiter and Jürgen Schmidhuber. Lstm can solve hard long time lag problems. *Advances in neural information processing systems*, pages 473–479, 1997.

[33] Maryem Rhanoui, Mounia Mikram, Siham Yousfi, and Soukaina Barzali. A cnn-bilstm model for document-level sentiment analysis. *Machine Learning and Knowledge Extraction*, 1(3):832–847, 2019.

[34] E. S. Epstein. A scoring system for probability forecasts of ranked categories. *Journal of Applied Meteorology*, 8(6):985–987, 1969.

[35] A. C. Constantinou and N. E. Fenton. Solving the problem of inadequate scoring rules for assessing probabilistic football forecast models. *Journal of Quantitative Analysis in Sports*, 8(1), 2012.

[36] Bet 365. bet365 - apostas desportivas online. `https://www.bet365.com/#/HO/`. (Acessado em 02/08/2018).

[37] Michael A Smith, David Paton, and Leighton Vaughan Williams. Do bookmakers possess superior skills to bettors in predicting outcomes? *Journal of Economic Behavior & Organization*, 71(2):539–549, 2009.

[38] Hyun Song Shin. Measuring the incidence of insider trading in a market for state-contingent claims. *The Economic Journal*, 103(420):1141–1153, 1993.

[39] Erik Štrumbelj. On determining probability forecasts from betting odds. *International journal of forecasting*, 30(4):934–943, 2014.

[40] Premier League. Premier league football news, fixtures, scores & results. `https://www.premierleague.com/`. (Acessado em 01/02/2021).

[41] Kou-Yuan Huang and Wen-Lung Chang. A neural network method for prediction of 2006 world cup football game. In *Neural Networks (IJCNN), The 2010 International Joint Conference on*, pages 1–8. IEEE, 2010.

[42] João Gomes, Filipe Portela, and Manuel F Santos. Pervasive decision support to predict football corners and goals by means of data mining. In *New Advances in Information Systems and Technologies*, pages 547–556. Springer, 2016.

[43] Rodrigo G Martins, Alessandro S Martins, Leandro A Neves, Luciano V Lima, Edna L Flores, and Marcelo Z do Nascimento. Exploring polynomial classifier to predict match results in football championships. *Expert Systems with Applications*, 83:79–93, 2017.

[44] World football elo ratings. `https://www.eloratings.net/`. (Acessado em 01/08/2018).

[45] Christoph Leitner, Achim Zeileis, and Kurt Hornik. Forecasting sports tournaments by ratings of (prob) abilities: A comparison for the euro 2008. *International Journal of Forecasting*, 26(3):471–481, 2010.

[46] Leonardo Soares Bastos and Joel Mauricio Correa da Rosa. Predicting probabilities for the 2010 fifa world cup games using a poisson-gamma model. *Journal of Applied Statistics*, 40(7):1533–1544, 2013.

[47] Fabian Wunderlich and Daniel Memmert. The betting odds rating system: Using soccer forecasts to forecast soccer. *PloS one*, 13(6):e0198668, 2018.

[48] Stefan Dobravec. Predicting sports results using latent features: A case study. In *Information and Communication Technology, Electronics and Microelectronics (MIPRO), 2015 38th International Convention on*, pages 1267–1272. IEEE, 2015.

[49] Lei Le, Emilio Ferrara, and Alessandro Flammini. On predictability of rare events leveraging social media: a machine learning perspective. In *Proceedings of the 2015 ACM on Conference on Online Social Networks*, pages 3–13. ACM, 2015.

[50] Alasdair Brown, Dooruj Rambaccussing, J James Reade, and Giambattista Rossi. Forecasting with social media: evidence from tweets on soccer matches. *Economic Inquiry*, 56(3):1748–1763, 2018.

[51] Darwin Prasetio et al. Predicting football match results with logistic regression. In *Advanced Informatics: Concepts, Theory And Application (ICAICTA), 2016 International Conference On*, pages 1–5. IEEE, 2016.

[52] Josip Hucaljuk and Alen Rakipović. Predicting football scores using machine learning techniques. In *MIPRO, 2011 Proceedings of the 34th International Convention*, pages 1623–1627. IEEE, 2011.

[53] Pınar Tüfekci. Prediction of football match results in turkish super league games. In *Proceedings of the Second International Afro-European Conference for Industrial Advancement AECIA 2015*, pages 515–526. Springer, 2016.

[54] Lars Magnus Hvattum. Ordinal versus nominal regression models and the problem of correctly predicting draws in soccer. *International Journal of Computer Science in Sport*, 16(1):50–64, 2017.

[55] Gianluca Baio and Marta Blangiardo. Bayesian hierarchical model for the prediction of football results. *Journal of Applied Statistics*, 37(2):253–264, 2010.

[56] A Tsakonas, G Dounias, S Shtovba, and V Vivdyuk. Soft computing-based result prediction of football games. In *The First International Conference on Inductive Modelling (ICIM'2002). Lviv, Ukraine*. Citeseer, 2002.

[57] Alexander P Rotshtein, Morton Posner, and AB Rakityanskaya. Football predictions based on a fuzzy model with genetic and neural tuning. *Cybernetics and Systems Analysis*, 41(4):619–630, 2005.

[58] Alessandro Martins Alves, João Carlos Correia Baptista Soares Mello, Thiago Graça Ramos, Annibal Parracho Sant'Anna, et al. Logit models for the probability of winning football games. *Pesquisa Operacional*, 31(3):459–465, 2011.

[59] Niek Tax and Yme Joustra. Predicting the dutch football competition using public data: A machine learning approach. *Transactions on Knowledge and Data Engineering*, 10(10):1–13, 2015.

[60] S Mohammad Arabzad, ME Tayebi Araghi, S Sadi-Nezhad, and Nooshin Ghofrani. Football match results prediction using artificial neural networks; the case of iran pro league. *Journal of Applied Research on Industrial Engineering*, 1(3):159–179, 2014.

[61] Andreas Groll and Jasmin Abedieh. Spain retains its title and sets a new record–generalized linear mixed models on european football championships. *Journal of Quantitative Analysis in Sports*, 9(1):51–66, 2013.

[62] Verica Lazova and Lasko Basnarkov. Pagerank approach to ranking national football teams. *arXiv preprint arXiv:1503.01331*, 2015.

[63] Karol Odachowski and Jacek Grekow. Using bookmaker odds to predict the final result of football matches. In *International Conference on Knowledge-Based and Intelligent Information and Engineering Systems*, pages 196–205. Springer, 2012.

[64] Martin Crowder, Mark Dixon, Anthony Ledford, and Mike Robinson. Dynamic modelling and prediction of english football league matches for betting. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 51(2):157–168, 2002.

[65] Radha-Krishna Balla. Soccer match result prediction using neural networks, 2007.

[66] Engin Esme and Mustafa Servet Kiran. Prediction of football match outcomes based on bookmaker odds by using k-nearest neighbor algorithm.

[67] Darwin Choi and HKUST Sam K Hui. The role of surprise: understanding over-and underreactions using in-play soccer betting. *HKUST/NYU Stern working paper*, 2012.

[68] Chinwe Peace Igiri. Support vector machine—based prediction system for a football match result. *IOSR Journal of Computer Engineering (IOSR-JCE)*, 17(3):21–26, 2015.

[69] Tianxiang Cui, Jingpeng Li, John R Woodward, and Andrew J Parkes. An ensemble based genetic programming system to predict english football premier league games. In *Evolving and Adaptive Intelligent Systems (EAIS), 2013 IEEE Conference on*, pages 138–143. IEEE, 2013.

[70] Andreas Groll, Christophe Ley, Gunther Schauberger, and Hans Van Eetvelde. Prediction of the fifa world cup 2018-a random forest approach with an emphasis on estimated team ability parameters. *arXiv preprint arXiv:1806.03208*, 2018.

[71] Pieter Robberechts, Jan Van Haaren, and Jesse Davis. Who will win it? an in-game win probability model for football. *arXiv preprint arXiv:1906.05029*, 2019.

[72] Georgi Boshnakov, Tarak Kharrat, and Ian G McHale. A bivariate weibull count model for forecasting association football scores. *International Journal of Forecasting*, 33(2):458–466, 2017.

[73] A Owen. The application of hurdle models to accurately model 0-0 draws in predictive models of football match outcomes. In *Proceedings of MathSport International 2017 Conference*, page 295, 2017.

[74] Michael J Maher. Modelling association football scores. *Statistica Neerlandica*, 36(3):109–118, 1982.

[75] Mark J Dixon and Stuart G Coles. Modelling association football scores and inefficiencies in the football betting market. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 46(2):265–280, 1997.

[76] Havard Rue and Oyvind Salvesen. Prediction and retrospective analysis of soccer matches in a league. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 49(3):399–418, 2000.

[77] Ben Ulmer, Matthew Fernandez, and Michael Peterson. *Predicting Soccer Match Results in the English Premier League*. PhD thesis, Doctoral dissertation, Ph. D. dissertation, Stanford, 2013.

[78] Anito Joseph, Norman E Fenton, and Martin Neil. Predicting football results using bayesian nets and other machine learning techniques. *Knowledge-Based Systems*, 19(7):544–553, 2006.

[79] Taoya Cheng, Deguang Cui, Zhimin Fan, Jie Zhou, and Siwei Lu. A new model to forecast the results of matches based on hybrid neural networks in the soccer rating system. In *Computational Intelligence and Multimedia Applications, 2003. ICCIMA 2003. Proceedings. Fifth International Conference on*, pages 308–313. IEEE, 2003.

[80] John Goddard and Ioannis Asimakopoulos. Forecasting football results and the efficiency of fixed-odds betting. *Journal of Forecasting*, 23(1):51–66, 2004.

[81] David Forrest, John Goddard, and Robert Simmons. Odds-setters as forecasters: The case of english football. *International journal of forecasting*, 21(3):551–564, 2005.

[82] John Goddard. Regression models for forecasting goals and match results in association football. *International Journal of forecasting*, 21(2):331–340, 2005.

[83] Burak Galip Aslan and Mustafa Murat Inceoglu. A comparative study on neural network based soccer result prediction. In *Seventh International Conference on Intelligent Systems Design and Applications (ISDA 2007)*, pages 545–550. IEEE, 2007.

[84] Byungho Min, Jinhyuck Kim, Chongyoun Choe, Hyeonsang Eom, and RI Bob McKay. A compound framework for sports results prediction: A football case study. *Knowledge-Based Systems*, 21(7):551–562, 2008.

[85] Anthony Costa Constantinou, Norman Elliott Fenton, and Martin Neil. Profiting from an inefficient association football gambling market: Prediction, risk and uncertainty using bayesian networks. *Knowledge-Based Systems*, 50:60–86, 2013.

[86] Siem Jan Koopman and Rutger Lit. A dynamic bivariate poisson model for analysing and forecasting match results in the english premier league. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 178(1):167–186, 2015.

[87] Yoonjae Cho, Jaewoong Yoon, and Sukjun Lee. Using social network analysis and gradient boosting to develop a soccer win–lose prediction model. *Engineering Applications of Artificial Intelligence*, 72:228–240, 2018.

[88] Laura Hervert-Escobar, Neil Hernandez-Gress, and Timothy I Matis. Bayesian based approach learning for outcome prediction of soccer matches. In *International Conference on Computational Science*, pages 269–279. Springer, 2018.

[89] Pieter Robberechts and Jesse Davis. Forecasting the fifa world cup–combining result- and goal-based team ability parameters. In *International Workshop on Machine Learning and Data Mining for Sports Analytics*, pages 16–30. Springer, 2018.

[90] Gunther Schauberger and Andreas Groll. Predicting matches in international football tournaments with random forests. *Statistical Modelling*, 18(5-6):460–482, 2018.

[91] Leonardo Egidi, Francesco Pauli, and Nicola Torelli. Combining historical data and bookmakers' odds in modelling football scores. *Statistical Modelling*, 18(5-6):436–459, 2018.

[92] Siem Jan Koopman and Rutger Lit. Forecasting football match results in national league competitions using score-driven time series models. *International Journal of Forecasting*, 35(2):797–809, 2019.

[93] Rahul Baboota and Harleen Kaur. Predictive analysis and modelling football results using machine learning approach for english premier league. *International Journal of Forecasting*, 35(2):741–755, 2019.

[94] Johannes Stübinger, Benedikt Mangold, and Julian Knoll. Machine learning in football betting: Prediction of match results based on player characteristics. *Applied Sciences*, 10(1):46, 2020.

[95] Halvard Arntzen and Lars Magnus Hvattum. Predicting match outcomes in association football using team ratings and player ratings. *Statistical Modelling*, page 1471082X20929881, 2020.

[96] Júlio Lobão and Nuno Marques Rolla. Um outro olhar sobre a eficiência dos mercados: o caso das bolsas de apostas de tênis. *RAE-Revista de Administração de Empresas*, 55(4):418–431, 2015.

[97] Martin Spann and Bernd Skiera. Sports forecasting: a comparison of the forecast accuracy of prediction markets, betting odds and tipsters. *Journal of Forecasting*, 28(1):55–72, 2009.

[98] Lisandro Kaunitz, Shenjun Zhong, and Javier Kreiner. Beating the bookies with their own numbers-and how the online sports betting market is rigged. *arXiv preprint arXiv:1710.02824*, 2017.

[99] Christian Deutscher, Bernd Frick, and Marius Ötting. Betting market inefficiencies are short-lived in german professional football. *Applied Economics*, 50(30):3240–3246, 2018.

[100] Edward Wheatcroft. A profitable model for predicting the over/under market in football. *International Journal of Forecasting*, 36(3):916–932, 2020.

[101] Karen Croxson and J James Reade. Information and efficiency: Goal arrival in soccer betting. *The Economic Journal*, 124(575):62–91, 2014.

[102] Mark Richard and Jan Vecer. Efficiency testing of prediction markets: Martingale approach, likelihood ratio and bayes factor analysis. *Risks*, 9(2):31, 2021.

[103] BetExplorer. Betexplorer soccer stats - results, tables, soccer stats & odds. http://www.betexplorer.com/. (Acessado em 01/08/2018).

[104] Alun Owen. Dynamic bayesian forecasting models of football match outcomes with estimation of the evolution variance parameter. *IMA Journal of Management Mathematics*, 22(2):99–113, 2011.

[105] Burton G Malkiel. Efficient market hypothesis. In *Finance*, pages 127–134. Springer, 1989.

[106] Donna Ryan et al. *High performance discovery in time series: techniques and case studies*. Springer Science & Business Media, 2013.

[107] Giovanni Angelini and Luca De Angelis. Parx model for football match predictions. *Journal of Forecasting*, 36(7):795–807, 2017.

[108] John L Kelly Jr. A new interpretation of information rate. In *The Kelly Capital Growth Investment Criterion: Theory and Practice*, pages 25–34. World Scientific, 2011.

[109] Exchange historical data. `https://historicdata.betfair.com/`. (Accessed on 02/16/2021).

[110] Peter Christen. The data matching process. In *Data matching*, pages 23–35. Springer, 2012.

[111] K Talattinis, G Kyriakides, E Kapantai, and G Stephanides. Forecasting soccer outcome using cost-sensitive models oriented to investment opportunities. *International Journal of Computer Science in Sport*, 18(1):93–114, 2019.

[112] Amin Ullah, Jamil Ahmad, Khan Muhammad, Muhammad Sajjad, and Sung Wook Baik. Action recognition in video sequences using deep bi-directional lstm with cnn features. *IEEE access*, 6:1155–1166, 2017.

[113] Tanveer Hussain, Khan Muhammad, Amin Ullah, Zehong Cao, Sung Wook Baik, and Victor Hugo C de Albuquerque. Cloud-assisted multiview video summarization using cnn and bidirectional lstm. *IEEE Transactions on Industrial Informatics*, 16(1):77–86, 2019.

[114] Luca Pappalardo, Paolo Cintia, Alessio Rossi, Emanuele Massucco, Paolo Ferragina, Dino Pedreschi, and Fosca Giannotti. A public data set of spatio-temporal match events in soccer competitions. *Scientific data*, 6(1):1–15, 2019.

[115] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017.

[116] Xuchao Zhang, Yifeng Gao, Jessica Lin, and Chang-Tien Lu. Tapnet: Multivariate time series classification with attentional prototypical network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 6845–6852, 2020.

Apêndice A

Revisão Sistemática

A.1 Consulta na plataforma SCOPUS

Pesquisa Avançada:

(((TITLE-ABS-KEY (predict*) OR TITLE-ABS-KEY (forecast*) AND NOT TITLE-ABS-KEY (robot)) AND (TITLE-ABS-KEY (soccer) OR TITLE-ABS-KEY (football)) AND (TITLE-ABS-KEY (result*) OR TITLE-ABS-KEY (outcome*) OR TITLE-ABS-KEY (score*) OR TITLE-ABS-KEY (winner) OR TITLE-ABS-KEY (bet*)) AND (SUBJAREA (busi) OR SUBJAREA (comp) OR SUBJAREA (deci) OR SUBJAREA (econ) OR SUBJAREA (math))))

Última Execução: 07/03/2021

Artigos Retornados: 805

Apêndice B

Testes Estatísticos para predição pré-jogo de "ambos marcam"

Campeonato	t-statistic	p-value
Brasil A	-3.24409	0.00120
Brasil B	0.10256	0.91832
Inglaterra A	-0.41028	0.68166
França A	-1.48826	0.13689
Alemanha A	2.98221	0.00292
Itália A	2.62129	0.00885
Holanda A	6.57011	0.00000
Portugal A	-2.41681	0.01581
Espanha A	0.05128	0.95911
All	1.383128	0.16664

Tabela B.1: T-Test para comparar a distribuição das classes para BTTS - $H_0 : yes = no$ / $H_1 : yes \neq no$ (relativo a Figura 4.2)

Campeonato	t-statistic	p-value
Brasil A	3175.0	0.020544
Brasil B	546.0	0.207578
Inglaterra A	20043.0	0.013810
França A	20100.0	0.073009
Alemanha A	2064.0	0.355809
Itália A	22518.0	0.012284
Holanda A	650.0	0.888638
Portugal A	11203.5	0.017608
Espanha A	35295.0	0.171014
All	776985.0	0.000002

Tabela B.2: Wilcoxon-Test para avaliar se CMM é melhor que CCM em termos de acurácia

	CMH	CDC	CRS	GNB_P	GNB_M	GNB_A	RLO_P	RLO_M	RLO_A	XGB_P	XGB_M	XGB_A	CCJ	CMM
IPO	0.0000	0.3329	0.3173	0.5498	0.0117	0.0256	0.1850	0.0007	0.0081	0.4211	0.0017	0.0020	0.0016	0.0016
DAC	0.3329	0.0000	0.3139	0.5121	0.0144	0.0309	0.2054	0.0009	0.0101	0.4591	0.0022	0.0025	0.0021	0.0021
RAS	0.3173	0.3139	0.0000	0.5421	0.0121	0.0265	0.1892	0.0007	0.0084	0.4282	0.0018	0.0021	0.0017	0.0017
GNB_P	0.5498	0.5121	0.5421	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
GNB_M	0.0117	0.0144	0.0121	0.0000	0.0000	0.0000	0.0000	0.0000	0.7205	0.0000	0.0000	0.0000	0.0000	0.0000
GNB_A	0.0256	0.0309	0.0265	0.0000	0.0000	0.0000	0.0000	0.0000	0.0002	0.0000	0.0000	0.0000	0.0000	0.0000
LRE_P	0.1850	0.2054	0.1892	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0159	0.0000	0.0000	0.0000	0.0000
LRE_M	0.0007	0.0009	0.0007	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.3014	0.3014
LRE_A	0.0081	0.0101	0.0084	0.0000	0.7205	0.0002	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
XGB_P	0.4211	0.4591	0.4282	0.0000	0.0000	0.0000	0.0159	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
XGB_M	0.0017	0.0022	0.0018	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.5290	0.0001	0.0001
XGB_A	0.0020	0.0025	0.0021	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.5290	0.0000	0.0007	0.0007
FBO	0.0016	0.0021	0.0017	0.0000	0.0000	0.0000	0.0000	0.3014	0.0000	0.0000	0.0001	0.0007	0.0000	0.0000
AMK	0.0016	0.0021	0.0017	0.0000	0.0000	0.0000	0.0000	0.3014	0.0000	0.0000	0.0001	0.0007	0.0000	0.0000

Tabela B.3: Wilcoxon-Test para comparar a lucratividade dos classificadores seguindo a estratégia AI contra a casa de apostas 1xBet

	CMH	CDC	CRS	GNB_P	GNB_M	GNB_A	RLO_P	RLO_M	RLO_A	XGB_P	XGB_M	XGB_A	CCJ	CMM
IPO	0.0000	0.1510	0.0000	0.0218	0.6427	0.0606	0.0068	0.0000	0.0000	0.0042	0.0000	0.0000	0.0000	0.0000
DAC	0.1510	0.0000	0.1510	0.0241	0.6580	0.0572	0.0075	0.0000	0.0000	0.0047	0.0000	0.0000	0.0000	0.0000
RAS	0.0000	0.1510	0.0000	0.0218	0.6427	0.0606	0.0068	0.0000	0.0000	0.0042	0.0000	0.0000	0.0000	0.0000
GNB_P	0.0218	0.0241	0.0218	0.0000	0.0000	0.0000	0.0011	0.0000	0.0007	0.0002	0.0041	0.0002	0.0000	0.0000
GNB_M	0.6427	0.6580	0.6427	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
GNB_A	0.0606	0.0572	0.0606	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
LRE_P	0.0068	0.0075	0.0068	0.0011	0.0000	0.0000	0.0000	0.0000	0.0002	0.0031	0.0099	0.0042	0.0078	0.0000
LRE_M	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0684	0.0000	0.0000	0.0000
LRE_A	0.0000	0.0000	0.0000	0.0007	0.0000	0.0000	0.0002	0.0000	0.0000	0.0000	0.0054	0.0000	0.0000	0.0000
XGB_P	0.0042	0.0047	0.0042	0.0002	0.0000	0.0000	0.0031	0.0000	0.0000	0.0000	0.0000	0.0000	0.2443	0.0000
XGB_M	0.0000	0.0000	0.0000	0.0041	0.0000	0.0000	0.0099	0.0684	0.0054	0.0000	0.0000	0.0025	0.0000	0.5753
XGB_A	0.0000	0.0000	0.0000	0.0002	0.0000	0.0000	0.0042	0.0000	0.0000	0.0000	0.0025	0.0000	0.0000	0.0012
FBO	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0078	0.0000	0.0000	0.2443	0.0000	0.0000	0.0000	0.0000
AMK	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.5753	0.0012	0.0000	0.0000

Tabela B.4: Wilcoxon-Test para comparar a lucratividade dos classificadores seguindo a estratégia VE contra a casa de apostas 1xBet

	CMH	CDC	CRS	GNB_P	GNB_M	GNB_A	RLO_P	RLO_M	RLO_A	XGB_P	XGB_M	XGB_A	CCJ	CMM
IPO	0.0000	0.8960	0.1598	0.0007	0.0000	0.1534	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0341	0.0000
DAC	0.8960	0.0000	0.9148	0.0007	0.0000	0.1471	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0357	0.0000
RAS	0.1598	0.9148	0.0000	0.0007	0.0000	0.1536	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0342	0.0000
GNB_P	0.0007	0.0007	0.0007	0.0000	0.0469	0.1016	0.5941	0.1977	0.2177	0.9279	0.2027	0.2396	0.0000	0.2170
GNB_M	0.0000	0.0000	0.0000	0.0469	0.0000	0.0089	0.0000	0.0001	0.0001	0.0000	0.0001	0.0001	0.0000	0.0001
GNB_A	0.1534	0.1471	0.1536	0.1016	0.0089	0.0000	0.2005	0.0223	0.0355	0.2014	0.0201	0.0248	0.4742	0.0248
LRE_P	0.0000	0.0000	0.0000	0.5941	0.0000	0.2005	0.0000	0.0000	0.0000	0.1138	0.0000	0.0000	0.0000	0.0000
LRE_M	0.0000	0.0000	0.0000	0.1977	0.0001	0.0223	0.0000	0.0000	0.6385	0.0000	0.5908	0.0383	0.0000	0.0357
LRE_A	0.0000	0.0000	0.0000	0.2177	0.0001	0.0355	0.0000	0.6385	0.0000	0.0000	0.3574	0.1256	0.0000	0.2748
XGB_P	0.0000	0.0000	0.0000	0.9279	0.0000	0.2014	0.1138	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
XGB_M	0.0000	0.0000	0.0000	0.2027	0.0001	0.0201	0.0000	0.5908	0.3574	0.0000	0.0000	0.0174	0.0000	0.9018
XGB_A	0.0000	0.0000	0.0000	0.2396	0.0001	0.0248	0.0000	0.0383	0.1256	0.0000	0.0174	0.0000	0.0000	0.2246
FBO	0.0341	0.0357	0.0342	0.0000	0.0000	0.4742	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
AMK	0.0000	0.0000	0.0000	0.2170	0.0001	0.0248	0.0000	0.0357	0.2748	0.0000	0.9018	0.2246	0.0000	0.0000

Tabela B.5: Wilcoxon-Test para comparar a lucratividade dos classificadores seguindo a estratégia AP contra a casa de apostas 1xBet