UNIVERSIDADE FEDERAL DE OURO PRETO
INSTITUTO DE CIÊNCIAS EXATAS E BIOLÓGICAS
DEPARTAMENTO DE COMPUTAÇÃO

ANANDA MENDES SOUZA
Orientador: Anderson Almeida Ferreira
Coorientador: Renato Lopes Moreira

PREDIÇÃO DE RESULTADOS DE JOGOS DO CAMPEONATO BRASILEIRO DE FUTEBOL FEMININO

Ouro Preto, MG
2024

UNIVERSIDADE FEDERAL DE OURO PRETO
INSTITUTO DE CIÊNCIAS EXATAS E BIOLÓGICAS
DEPARTAMENTO DE COMPUTAÇÃO

ANANDA MENDES SOUZA

PREDIÇÃO DE RESULTADOS DE JOGOS DO CAMPEONATO BRASILEIRO DE FUTEBOL FEMININO

Monografia II apresentada ao Curso de Ciência da Computação da Universidade Federal de Ouro Preto como parte dos requisitos necessários para a obtenção do grau de Bacharel em Ciência da Computação.

Orientador: Anderson Almeida Ferreira
Coorientador: Renato Lopes Moreira

Ouro Preto, MG
2024

SISBIN - SISTEMA DE BIBLIOTECAS E INFORMAÇÃO

S729p Souza, Ananda Mendes.
Predição de resultados de jogos do campeonato brasileiro feminino. [manuscrito] / Ananda Mendes Souza. - 2024.
66 f.: il.: color., gráf., tab..

Orientador: Prof. Dr. Anderson Almeida Ferreira.
Coorientador: Prof. Me. Renato Lopes Moreira.
Monografia (Bacharelado). Universidade Federal de Ouro Preto. Instituto de Ciências Exatas e Biológicas. Graduação em Ciência da Computação .

1. Campeonato Brasileiro (Futebol). 2. Futebol feminino. 3. Aprendizado do computador. 4. Futebol - Jogos. I. Ferreira, Anderson Almeida. II. Moreira, Renato Lopes. III. Universidade Federal de Ouro Preto. IV. Título.

CDU 796.332

Bibliotecário(a) Responsável: Paulo Vitor Oliveira - CRB6/2551

MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DE OURO PRETO
REITORIA
INSTITUTO DE CIENCIAS EXATAS E BIOLOGICAS
DEPARTAMENTO DE COMPUTACAO

FOLHA DE APROVAÇÃO

Ananda Mendes Souza

Predição de Resultados de Jogos do Campeonato Brasileiro de Futebol Feminino

Monografia apresentada ao Curso de Ciência da Computação da Universidade Federal
de Ouro Preto como requisito parcial para obtenção do título de Bacharel em Ciência da Computação

Aprovada em 7 de Fevereiro de 2024.

Membros da banca:

Anderson Almeida Ferreira (Orientador) - Doutor - Universidade Federal de Ouro Preto
Renato Lopes Moreira (Coorientador) - Mestre - Universidade Federal de Ouro Preto
Jadson Castro Gertrudes (Examinador) - Doutor - Universidade Federal de Ouro Preto
Pedro Henrique Lopes Silva (Examinador) - Doutor - Universidade Federal de Ouro Preto

Anderson Almeida Ferreira, Orientador do trabalho, aprovou a versão final e autorizou seu depósito na Biblioteca Digital de Trabalhos de Conclusão de Curso da UFOP em 7/02/2024.

Documento assinado eletronicamente por **Anderson Almeida Ferreira**, **PROFESSOR DE MAGISTERIO SUPERIOR**, em 16/02/2024, às 18:30, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

A autenticidade deste documento pode ser conferida no site http://sei.ufop.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0 , informando o código verificador **0665676** e o código CRC **73EA3A13**.

R. Diogo de Vasconcelos, 122, - Bairro Pilar Ouro Preto/MG, CEP 35402-163
Telefone: 3135591692 - www.ufop.br

Ao longo da minha vida, minha família sempre foi sinônimo de futebol. Seja aos domingos com os gritos (e xingamentos) durante os jogos do Flamengo ou nas intermináveis discussões pós jogo, não dá pra falar de Ananda sem falar de futebol. E eu dedico este trabalho a pessoa que me ensinou a amar esse esporte que me trouxe tantas alegrias: o meu "painho" Aloisio.
Com amor, Ananda.

Agradecimentos

À minha querida família, expresso meu sincero agradecimento pelo apoio constante. A jornada acadêmica foi marcada por desafios, e a presença de vocês foi o pilar que sustentou meu percurso. O amor e os ensinamentos são inestimáveis.

Aos amigos que se tornaram companheiros de jornada, desde os primeiros encontros do 19.1 até aqueles que surgiram ao longo do curso, agradeço por cada momento compartilhado. Cada um de vocês deixou uma marca especial em minha vida, tornando essa experiência universitária mais rica e significativa.

Especial reconhecimento à Gerardo, que me incentivou a iniciar essa jornada em outro estado, e as amigas de longa data, Carol e Fernanda, que compartilham comigo as vitórias desde a adolescência. A amizade de vocês foi um suporte valioso.

À Universidade Federal de Ouro Preto, manifesto minha profunda gratidão por ser o cenário que abrigou uma parte crucial da minha formação acadêmica. Agradeço a todos os professores e colaboradores que, com dedicação e comprometimento, foram fundamentais para meu desenvolvimento ao longo dos anos. Não posso deixar de expressar minha gratidão ao Instituto Federal da Bahia, foi nessa instituição que descobri meu gosto pela área da tecnologia e adquiri as bases que me impulsionaram para a graduação.

Aos programas de permanência estudantil, dedico meu agradecimento pela oportunidade de ingressar na graduação e pelo suporte financeiro que facilitou minha jornada acadêmica.

Por último, agradeço a Deus por todas as oportunidades, bênçãos e proteção concedidas ao longo dessa jornada.

Cada um dos citados neste agradecimento tiveram um papel fundamental e especial nesta conquista. A jornada pode ter sido desafiadora, mas a gratidão que sinto é eterna.

"O campo é o único lugar onde eu sou realmente eu mesmo."

Pelé

Resumo

No contexto esportivo, o uso de aprendizado de máquina tem se mostrado promissor na análise e previsão de resultados de competições esportivas. O Campeonato Brasileiro de Futebol Feminino é uma das principais competições do futebol feminino no país, e seu crescimento tem sido notável ao longo dos anos. No entanto, ainda existem desafios na análise e compreensão dos resultados das partidas. Dito isso, este trabalho tem como objetivo explorar a aplicação de aprendizado de máquina na predição de resultados de jogos do Campeonato Brasileiro de Futebol Feminino, buscando fornecer informações valiosas sobre o desempenho das equipes, o impacto de diferentes fatores em suas atuações e, consequentemente, auxiliar na tomada de decisões estratégicas para aprimorar o desempenho das equipes e melhorar a competitividade do campeonato. A metodologia adotada engloba a coleta de dados provenientes do site *Footystats*, constituindo um conjunto de exemplos de treinamento robusto composto por informações de cinco ligas, Estados Unidos, Itália, Inglaterra, Espanha, Copa Europeia. Adicionalmente, os jogos do Campeonato Brasileiro de 2022, totalizando 134 partidas, são utilizados como conjunto de testes. O processo inclui análise exploratória dos dados e a construção de modelos de predição por meio de técnicas de indução de classificadores. Os resultados obtidos confirmam a complexidade associada a esta tarefa, ressaltando a dificuldade frequentemente encontrada na literatura para alcançar taxas de acerto superiores a 60%.

Palavras-chave: Aprendizado de máquina; Campeonato Brasileiro de Futebol Feminino; Predição de resultados de partidas.

Abstract

In the sporting context, machine learning has shown promise in analyzing and predicting the results of sporting competitions. The Brazilian Women's Football Championship is one of the country's main women's football competitions, and its growth has been remarkable over the years. However, there are still challenges in analyzing and understanding match results. Thus, this work aims to explore the application of machine learning in predicting game results in the Brazilian Women's Football Championship, seeking to provide valuable information about the performance of teams, the impact of different factors on their performances, and, consequently, assist in making strategic decisions to improve team performance and improve the competitiveness of the championship. The methodology adopted encompasses gathering the data from the Footystats website, constituting a robust training set composed of information from five leagues: USA, Italy, England, Spain, and the European Cup. Additionally, the 2022 Brazilian Championship games, totaling 134 matches, are used as a set of tests. The process includes exploratory data analysis and inferring of prediction models through classifier induction techniques. The results confirmed the complexity associated with this task, highlighting the difficulty frequently found in the literature to achieve success rates above 60%.

Keywords: Machine Learning. Brazilian Women's Football Championship. Match outcome prediction.

Lista de Ilustrações

Lista de Tabelas

Lista de Abreviaturas e Siglas

ABNT	Associação Brasileira de Normas Técnicas
DECOM	Departamento de Computação
UFOP	Universidade Federal de Ouro Preto
IA	Inteligência Artificial
AM	Aprendizado de Máquina
RL	Regressão Logística
NB	Naive Bayes
KNN	K-Nearest Neighbors
ADQ	Análise Discriminante Quadrática
ADL	Análise Discriminante Linear
DT	Decision Tree
MLP	Multi-Layer Perceptron
SVM	Support Vector Machine
RF	Random Forest
SVML	Support Vector Machine com kernel linear
SVMR	Support Vector Machine com kernel RBF
ET	Extra Trees
CSV	Valores Separados por Vírgulas
xG	Expectativa de Gols
xA	Expectativa de Assistências
XGB	Extreme Gradient Boosting
MDA	Mean Decrease Accuracy
LNF	Liga Nacional de Futsal
LINAF	Liga Nacional de Futebol
CBF	Confederação Brasileira de Futebol

Lista de Símbolos

ξ	Letra grega Xi
α	Letra grega Alpha
β	Letra grega Beta
γ	Letra grega gamma
τ	Letra grega Tau
ϕ	Letra grega Phi

Sumário

1 Introdução

O futebol feminino tem ganhado cada vez mais destaque e reconhecimento nos últimos anos. Porém, ainda enfrenta desafios em termos de análise de dados e previsão de resultados. A falta de informação disponível sobre partidas e jogadoras dificulta a criação de modelos precisos para prever o desempenho de equipes ou individualmente de jogadoras por analistas e treinadores. Além disso, a maioria dos bancos de dados disponíveis se concentra no futebol masculino e recentemente algumas plataformas começaram a incluir dados completos, como, por exemplo, estatísticas dos campeonatos, das partidas, dos times e das jogadoras do futebol feminino. No entanto, como afirma Detoni (2022), o futebol feminino no Brasil é frequentemente subestimado e recebe pouca atenção da mídia e do público em geral. Isso resulta em recursos limitados disponíveis para coletar e analisar dados do futebol feminino.

A partir de 2019, observou-se uma melhora significativa no cenário do futebol feminino, impulsionada pelo sucesso da Copa do Mundo da França. Nesse ano, ocorreram marcos importantes, como a inclusão de um grande número de clubes tradicionais no futebol masculino no cenário feminino, devido à obrigatoriedade de equipes femininas para os times masculinos da Série A do Brasileirão e da Libertadores. Além disso, houve o retorno das transmissões do Campeonato Brasileiro feminino na televisão (Nunes, 2023).

Desde 2022, o Brasileirão feminino passou a contar com três divisões, e o calendário nacional foi ampliado para incluir a Supercopa feminina. Ao longo dos últimos anos, muitos clubes tradicionais no futebol masculino consolidaram sua presença na elite do futebol feminino. Atualmente, a Série A1 é composta por 13 times de renome, como Corinthians, Palmeiras, Flamengo, Cruzeiro, Santos, Internacional, Grêmio, São Paulo, Bahia, Athletico-PR, Atlético-MG, Avaí/Kindermann e Ceará. Mesmo que nem sempre o “peso da camisa” seja revertido em valorização e profissionalização para o feminino (Nunes, 2023).

Com o crescimento do futebol feminino no país, tem havido um aumento significativo na quantidade de dados disponíveis referentes às competições e times. Atualmente, plataformas reconhecidas, inclusive sites de apostas, passaram a oferecer cobertura do Campeonato Brasileiro Feminino e outras competições (Estadao, 2022). Embora ainda haja lacunas e áreas a serem aprimoradas, esse avanço representa um importante começo para a área de análise de dados no futebol feminino, visto que a inclusão de plataformas com cobertura do futebol feminino proporciona aos analistas, treinadores e fãs acesso a estatísticas. Isso possibilita uma análise mais detalhada do desempenho das equipes e o desenvolvimento de modelos preditivos mais precisos (Rosa, 2022).

Apesar da crescente disponibilidade de dados, ainda são escassos os estudos na área de análise de dados do futebol feminino no Brasil (Peconick, 2018). Mais especificamente, a

utilização desses dados para desenvolver modelos de previsão ainda é um campo pouco ou quase nada explorado na modalidade feminina. Essa lacuna representa uma oportunidade para a expansão do conhecimento e o avanço da análise de dados no futebol feminino brasileiro. A criação de trabalhos dedicados à predição de resultados pode contribuir para identificar áreas em que as equipes precisam melhorar para competir em um nível mais alto. Isso pode ajudar os treinadores e jogadores a identificar áreas de fraqueza e desenvolver estratégias para melhorar o desempenho.

Buscando resolver essa problemática, neste trabalho, será abordado a temática das previsões de resultados, por meio de aprendizado de máquina, para jogos do Campeonato Brasileiro de Futebol Feminino série A, utilizando métodos de aprendizado de máquina. Essa técnica visa melhorar a precisão e a estabilidade dos modelos de aprendizado de máquina, reduzindo o risco de *overfitting* (sobreajuste) e melhorando a precisão do modelo (Dietterich, 2000). A originalidade reside na exploração de técnicas de aprendizado de máquina treinadas com outras ligas do futebol feminino e aplicadas especificamente ao Campeonato Brasileiro de Futebol Feminino Série A, considerando as particularidades e desafios únicos dessa modalidade esportiva. Devido à escassez de dados disponíveis referentes ao Campeonato Brasileiro, houve a necessidade da construção da base de dados por meio da incorporação de informações provenientes de outras ligas.

A plataforma *Footystats* está sendo utilizada para obter os dados. Essa plataforma se mostrou a mais adequada, pois ofereceu arquivos no formato de valores separados por vírgulas (do inglês *Comma-separated values*, CSV) com estatísticas detalhadas sobre a liga, times, jogadoras e partidas.

Portanto, este trabalho visa contribuir para o avanço da análise de dados no esporte e ajudar a preencher a lacuna existente na falta de informações sobre o desempenho das equipes e jogadoras do futebol feminino. Além disso, pode ajudar a identificar tendências e padrões de desempenho e permitindo decisões mais informadas. O uso de modelos de aprendizado de máquina também pode aumentar a visibilidade da modalidade e atrair investimentos e recursos para a sua melhoria.

1.1 Justificativa

A realização deste trabalho sobre previsões de resultados utilizando aprendizado de máquina para o Campeonato Brasileiro de Futebol Feminino é fundamentada em diversos aspectos relevantes no contexto da formação acadêmica e profissional, mas principalmente pela motivação pessoal da autora. Pois a paixão pelo futebol e a trajetória como entusiasta do esporte ao longo da vida despertaram o desejo de contribuir para o avanço do futebol feminino. Ao perceber a escassez de análise de dados nessa modalidade, fica claro a importância de explorar essa lacuna e fornecer uma ferramenta útil para prever os resultados dos jogos com maior precisão, este trabalho pode ajudar a aumentar o interesse e a visibilidade do futebol feminino no Brasil, além

de poder ser usado em outras ligas ao redor do mundo.

Segundo Leitner, Zeileis e Hornik (2020), o impacto positivo dessa maior visibilidade pode se refletir em um aumento significativo no número de espectadores, patrocinadores e investidores no esporte. Isso, por sua vez, pode gerar melhorias nas condições e infraestrutura da modalidade, proporcionando mais recursos e oportunidades para as jogadoras.

Além disso, é reconhecida a relevância do uso de técnicas avançadas, como aprendizado de máquina, para aprimorar o desempenho esportivo, conforme mencionado por (Mackenzie; Esporte, 2012):

> "métodos e recursos de treinamento que englobam a ciência e a tecnologia para a melhoria do desempenho são requisitos básicos no empreendimento de um atleta que almeja vitórias em competições de alto rendimento."(Mackenzie; Esporte, 2012, p. 144)

Ao aplicar essas técnicas no contexto do Campeonato Brasileiro de Futebol Feminino, será possível, não apenas oferecer informações significativas para treinadores, jogadoras e torcedores, mas também contribuir para o avanço do conhecimento nesse campo específico.

1.2 Objetivos

O objetivo principal deste trabalho é desenvolver e aplicar uma abordagem de previsões de resultados de jogos utilizando aprendizado de máquina para o Campeonato Brasileiro de Futebol Feminino, com o intuito de melhorar a análise de dados e fornecer percepções estratégicas para treinadores, analistas e gestores.

Especificamente este trabalho tem como objetivos:

- Investigar uma forma para coletar e analisar os dados disponibilizados pelo Footystats[1];
- Inferir modelos preditivos para o Campeonato Brasileiro de Futebol Feminino;
- Avaliar experimentalmente esses modelos.

1.3 Organização do Trabalho

Este trabalho é dividido em cinco capítulos que abordam diferentes aspectos do tema em questão. O capítulo atual é a introdução, que mostra uma visão geral do trabalho e sua estrutura.

No Capítulo 2, é realizada uma revisão bibliográfica abrangente, abordando trabalhos relacionados e fornecendo uma fundamentação teórica sólida.

[1] https://footystats.org/pt/

O Capítulo 3 é dedicado ao desenvolvimento do trabalho, sendo detalhado o tipo de pesquisa realizada e a metodologia adotada, incluindo as técnicas de aprendizado de máquina e a base de dados utilizada, visando atingir os objetivos propostos de predição de resultados e otimização do modelo para uma maior acurácia.

O Capítulo 4 descreve os resultados obtidos a partir da implementação do modelo e dos testes realizados. Por fim, o Capítulo 5 contém as considerações finais, destacando os objetivos alcançados e sugerindo possíveis direções para trabalhos futuros.

Essa estrutura organizada proporciona uma compreensão clara do trabalho e permite uma abordagem abrangente do tema, explorando diferentes aspectos relevantes.

2 Revisão Bibliográfica

Este capítulo, contextualiza o tema de aprendizado de máquina (AM) e predições de resultados no futebol. Inicialmente, tem-se uma fundamentação teórica sobre conceitos utilizados no decorrer deste trabalho, visando possibilitar ao leitor um melhor entendimento da metodologia aplicada. Em seguida, são descritos alguns trabalhos relacionados que abordaram estes assuntos em outras modalidades, como no futebol masculino, até aqueles que abordam o futebol feminino, tais como a predição de resultados da Copa do Mundo Feminina de 2019.

2.1 Fundamentação Teórica

Nesta seção, busca-se fornecer ao leitor um entendimento claro dos conceitos e termos abordados neste trabalho, especialmente no capítulo de metodologia. Com esse propósito, serão apresentadas as principais noções de AM, e métodos de predição, com o intuito de aprimorar a compreensão do tema em questão.

2.1.1 Campeonato Brasileiro Feminino

O Campeonato Brasileiro Feminino Série A1, conhecido como Brasileirão Feminino, representa a principal divisão do futebol feminino no Brasil. Criado em 2013 pela Confederação Brasileira de Futebol (CBF), o torneio envolve dezesseis clubes e segue um sistema de promoção e despromoção com a Série A2.

A proibição da prática do futebol por mulheres no Brasil foi instituída pelo decreto-lei 3199 em 1941, durante a ditadura do Estado Novo, sendo revogada somente em 1979, ainda durante a ditadura militar. O ressurgimento do futebol feminino aconteceu em 1979 com a Liga Nacional de Futebol (LINAF) organizando os primeiros campeonatos nacionais. Entre 1983 e 1989, a Taça Brasil de Futebol Feminino destacou o Radar como o primeiro campeão nacional, vencendo equipes consagradas do futebol masculino. O torneio passou por mudanças de nome, tornando-se o Torneio Nacional na década de 1990.

A CBF assumiu a organização em 1993, renomeando-o como Taça Brasil de Clubes. Em 1994, tornou-se o Campeonato Nacional Brasileiro de Futebol Feminino, mas foi cancelado em 1995. O torneio foi interrompido até 2006, quando a LINAF, em parceria com a Federação Paulista de Futebol Amador, o reintroduziu como Liga Nacional. Em 2007, a CBF instituiu a Copa do Brasil de Futebol Feminino, e em 2013, em parceria com a Caixa Econômica Federal, lançou o Campeonato Brasileiro. Inicialmente com uma temporada de três meses, o campeonato ofereceu apoio financeiro às equipes em 2015.

Em 2017, a CBF reformulou a competição, reduzindo a Série A1 para 16 times e criando a Série A2. A Copa do Brasil de Futebol Feminino foi cancelada devido à expansão do Campeonato Brasileiro. A partir de 2019, o torneio ganhou mais visibilidade na TV aberta, resultando em um aumento significativo no interesse do público (Exame, 2024).

Atualmente, o campeonato é disputado em quatro fases. Na primeira, os 16 clubes jogam no modelo de pontos corridos todos contra todos, em turno único. Nas quartas de final, o torneio assume um formato de mata-mata, com confrontos eliminatórios em dois jogos, considerando partidas de ida e volta. Os oito primeiros se classificam para as quartas de final e os quatro últimos são rebaixados para a Série A2. Os confrontos são determinados de acordo com a posição na tabela, o primeiro colocado contra o oitavo, o segundo contra o sétimo, e assim por diante.

2.1.2 Aprendizado de máquina

O aprendizado de máquina (ou *"machine learning"* em inglês) é um subcampo da inteligência artificial (IA), que se concentra no desenvolvimento de algoritmos e técnicas que permitem que os computadores aprendam a partir de dados e melhorem seu desempenho em tarefas específicas sem serem explicitamente programados. A principal vantagem é a sua habilidade de lidar com tarefas complexas e difíceis de serem definidas explicitamente, permitindo que os sistemas se adaptem e evoluam ao longo do tempo, reduzindo erros de projeto e a necessidade de manutenções constantes (Nilsson, 1997). Dentro desse campo científico, é possível identificar duas categorias fundamentais, conforme destacado por Haykin (2009): Aprendizado Supervisionado e Aprendizado Não Supervisionado.

As técnicas de indução de classificadores são um tipo de aprendizado supervisionado que busca obter um modelo de classificação, também chamado de classificador, capaz de aprender a relação entre as variáveis independentes ou preditoras e a variável dependente (classes ou labels) das instâncias presentes em um conjunto de treinamento.

Formalmente, considere um exemplo de treinamento como um par $(x_i, f(x_i))$, sendo x_i representa a entrada do exemplo e $f(x_i)$ é a saída (classe) associada a essa entrada. O objetivo central de um indutor é, dada uma coleção de exemplos de treinamento, aprender uma função h que possa aproximar a função f - que geralmente é desconhecida. Nesse contexto, a função h é denominada hipótese, e a meta é fazer com que $h(x_i)$ se aproxime o máximo possível de $f(x_i)$ para cada exemplo x_i.

Esse classificador é treinado com dados cujas respostas são conhecidas, permitindo-lhe aprender a fazer previsões precisas para novos dados nunca antes vistos. O poder de generalização do modelo é fundamental para garantir sua eficácia na tarefa de predição (Kotsiantis *et al.*, 2007).

Nas subseções seguintes, serão apresentados os diferentes classificadores mencionados neste trabalho, destacando suas características e aplicações específicas, proporcionando uma visão das abordagens utilizadas e das possibilidades de otimização para a construção do ensemble

final.

2.1.2.1 K-Nearest Neighbors

Essa abordagem, descrita inicialmente por Cover e Hart (1967) e fundamentada na ideia de atribuir a uma amostra não classificada a classificação mais popular dos K exemplos (instâncias/vizinhos) mais próximos de uma instância que se deseja classificar.

A métrica mais comum para determinação dos K-vizinhos mais próximos é a distância euclidiana, definida pela Equação (2.1).

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{i=1}^{d}(x_i - y_i)^2} \tag{2.1}$$

Tal que x e y são dois pontos no espaço d-dimensional, x_i e y_i são as coordenadas do ponto x e do ponto y na dimensão i. E a soma se estende sobre todas as dimensões, $i = 1$ até d.

Em termos mais simples, x e y representam dois pontos que estão sendo comparados em um espaço d-dimensional, onde d é o número de características ou dimensões que cada ponto possui. A fórmula da distância euclidiana mede a distância entre esses pontos levando em consideração todas as dimensões, calculando a diferença entre suas coordenadas em cada dimensão, elevando-as ao quadrado, somando esses quadrados e, finalmente, tirando a raiz quadrada da soma. Isso nos dá uma medida de quão distantes ou próximos esses pontos estão uns dos outros.

Após calcular as distâncias entre uma instância de teste x^* e todas as que compõem o conjunto de dados de treinamento, o próximo passo é selecionar os K vizinhos mais próximos. A classificação da instância x^* é determinada pela classe majoritária entre os K vizinhos mais próximos. A fórmula para a classificação é dada por:

$$\hat{y_*} = \arg\max_{y_i} \sum_{i=1}^{K} I(y_i = y^*) \tag{2.2}$$

sendo y^* é a classe prevista para a instância de teste x^*, y_i é a classe de um dos K vizinhos mais próximos, e I é uma função indicadora que é 1 se y_i é igual a y^* e 0 caso contrário.

Os parâmetros frequentemente utilizados no KNN são '*n_neighbors*' e '*weights*'. O '*n_neighbors*' determina o número de vizinhos mais próximos que serão considerados ao realizar uma previsão para uma nova instância. Valores mais altos suavizam o impacto de pontos de dados atípicos, mas podem tornar o modelo menos sensível a padrões locais. Quanto ao '*weights*', ele determina como os vizinhos contribuem para a previsão.

Os valores específicos para cada um foram escolhidos como [3, 5, 7, 9] para '*n_neighbors*' e ['*uniform*', '*distance*'] para '*weights*'. '*uniform*' significa que todos os vizinhos têm o mesmo

peso, enquanto '*distance*' atribui um peso maior aos vizinhos mais próximos.

2.1.2.2 Random Forest

O *Random Forest* é um classificador que utiliza o conceito de árvores de decisão. Essas árvores são estruturas simples, onde as folhas representam as classes e os nós não foliares representam atributos, baseados em testes, com um ramo para cada possível saída. Ao classificar um objeto, o processo inicia-se na raiz da árvore, aplicando-se o teste em cada nó e seguindo o ramo apropriado para aquela saída. Esse procedimento continua até alcançar uma folha, momento em que o objeto é classificado de acordo com a classe indicada naquela folha (Breiman *et al.*, 1984).

Esse método combina múltiplas árvores de decisão, sendo cada árvore dependente de uma amostra aleatória de características independentes. Essa amostra possui a mesma distribuição para todas as árvores na floresta, constituindo um conjunto de árvores. A característica notável desse método é a convergência do erro de generalização da floresta à medida que se aumenta a quantidade de árvores. Esse erro depende tanto da força individual de cada árvore quanto da correlação entre elas, conforme proposto por Breiman (2001).

Uma peculiaridade é a utilização de amostras aleatórias de *features* para a divisão de cada nó nas árvores. Esse processo cria um erro geralmente menor em comparação com outros classificadores, tornando o *Random Forest* mais robusto perante a presença de ruído nos dados (Breiman, 2001).

A função de margem do classificador *Random Forest*, expressa pela Equação (2.3), é definida para medir o quanto a média de votos ultrapassa a média de votos para todas as demais classes. Essa margem, representada por $mg(Y, h)$, é diretamente proporcional ao intervalo de confiança da classificação, refletindo a capacidade de decisão do modelo.

$$mg(Y, h) = \max_{j \neq Y} \left(\frac{1}{K} \sum_{k=1}^{K} I(h_k(X) = j) - I(h_k(X) = Y) \right) \tag{2.3}$$

sendo, j o índice da classe considerada, onde $j \neq Y$. K o número total de classes no problema. $h_k(X)$ a função de decisão da k-ésima árvore na floresta para a amostra X, essa função atribui a amostra a uma classe específica. I a função indicadora que retorna 1 se a condição dentro dos parênteses for verdadeira e 0 caso contrário.

Os parâmetros frequentemente empregados pelo *Random Forest* são '*n_estimators*' e '*max_depth*'. O primeiro diz respeito ao número de árvores na floresta, enquanto o segundo representa a profundidade máxima de cada árvore. Os valores selecionados para esses parâmetros foram [100, 200, 300, 400] para '*n_estimators*' e [None, 10, 20, 30] para '*max_depth*'.

O *Random Forest* destaca-se por sua abordagem aleatória e pela combinação eficaz de múltiplas árvores de decisão, resultando em um modelo poderoso e robusto para problemas de

classificação.

2.1.2.3 Extreme Gradient Boosting

O algoritmo *Extreme Gradient Boosting (XGBoost)* é uma abordagem baseada em árvores de decisão aplicável tanto para regressão quanto para classificação. Expandindo o conceito do *Gradient Boosting*, o *XGBoost* incorpora melhorias significativas, principalmente na normalização da função de perda, buscando mitigar a variância do modelo. Este método de *ensemble* é construído com árvores de decisão de profundidade reduzida (Chen; Guestrin, 2016).

O *XGBoost* segue a ideia de treinamento aditivo, onde as árvores já construídas nas iterações anteriores permanecem inalteradas, e cada nova árvore é construída com base no aprendizado dos resíduos da árvore anterior.

Na predição final, a contribuição de cada árvore é ponderada e somada para formar o resultado. O modelo é otimizado iterativamente para reduzir o erro de predição. A função objetivo do *XGBoost* incorpora termos de regularização, representados por ϕ , que controlam os parâmetros de aprendizagem, a função de perda (representando o erro predito versus observado), e um termo de regularização que previne o *overfitting*.

O *XGBoost*, ao minimizar uma função de perda pré-definida, coloca mais peso nos casos preditos erroneamente pelas árvores desenvolvidas anteriormente. Isso resulta em um modelo final que é uma combinação ponderada das contribuições de todas as árvores desenvolvidas. O *XGBoost* reduz a complexidade do modelo, mitigando a probabilidade de *overfitting* em comparação com o *Gradient Boosting* tradicional. Esta abordagem também lida com a expansão de *Taylor* na função de perda, proporcionando melhorias na medida em que a complexidade do aprendizado aumenta (Chen; Guestrin, 2016).

No caso do XGBoost, foram ajustados os parâmetros '*max_depth*' e '*learning_rate*'. O '*max_depth*' representa a profundidade máxima de cada árvore, enquanto o '*learning_rate*' determina a taxa de aprendizado do modelo. Essa taxa controla a magnitude com que os pesos do modelo são ajustados durante o treinamento, influenciando a rapidez com que o algoritmo aprende com os dados. Uma taxa alta permite ajustes mais rápidos, mas pode levar a oscilações indesejadas, enquanto uma taxa baixa pode resultar em treinamento lento. A combinação do '*learning_rate*' com técnicas de regularização pode ser empregada para controlar a complexidade do modelo e evitar *overfitting*, especialmente ao reduzir a taxa de aprendizado enquanto aumenta o número de árvores na floresta. Os valores escolhidos para esses parâmetros foram [3, 4, 5, 6] para '*max_depth*' e [0.1, 0.01, 0.001] para '*learning_rate*'.

2.1.3 Métricas de Avaliação

Após a criação de um classificador ou modelo de aprendizado de máquina, é essencial avaliá-lo para compreender o seu desempenho e eficácia. No caso específico da classificação,

existem várias métricas utilizadas para analisar a qualidade das previsões feitas pelo modelo, dependendo do contexto e dos requisitos específicos do problema (Matos *et al.*, 2009). Dentre elas: Taxa de Erro, Acurácia, Precisão (ou *Positive Predictive Value*), *Recall* (ou Sensibilidade) e *F1-score*.

Para entender melhor essas métricas de avaliação em um classificador, é necessário considerar alguns conceitos. Considere como exemplo um conjunto X e uma classificação com duas classes: a classe "A"será tratada como positiva, e a classe "B"como negativa.

- **Verdadeiros Positivos (VP):** São as instâncias que o classificador rotulou corretamente como pertencentes à classe "A", e de fato, elas eram da classe "A".
- **Falsos Positivos (FP):** Representam as instâncias que o classificador rotulou erroneamente como pertencentes à classe "A", mas na verdade, elas eram da classe "B".
- **Verdadeiros Negativos (VN):** Correspondem às instâncias que o classificador rotulou corretamente como pertencentes à classe "B".
- **Falsos Negativos (FN):** São as instâncias que o classificador rotulou de forma equivocada como pertencentes à classe "B", porém, elas pertenciam à classe "A".

Tabela 2.1 – Matriz de confusão

		Detectada	
		Sim	**Não**
Real	**Sim**	Verdadeiro Positivo (VP)	Falso Negativo (FN)
	Não	Falso Positivo (FP)	Verdadeiro Negativo (VN)

Fonte: Autora

Neste trabalho, serão utilizadas a Acurácia e F1-Score como métricas de avaliação.

2.1.3.1 Acurácia

Mikhail e Ackermann (1976) afirmam que a acurácia (A) pode ser definida como o grau de proximidade que uma estimativa tem de seu parâmetro, ou seja, a proporção das predições verdadeiras em relação ao total de previsões realizadas, conforme apresentado na equação (2.4).

$$A = \frac{VP + VN}{VP + FP + VN + FN} \tag{2.4}$$

2.1.3.2 F1-Score

A *F1-Score* (F1) é uma métrica de avaliação usada para medir o equilíbrio entre a Precisão e a Revocação, conceito proposto por Turing (1950).

A Precisão (P) é a proporção de instâncias corretamente classificadas como positivas (classe "A") em relação ao total de instâncias classificadas como positivas pelo modelo. A fórmula é dada por:

$$P = \frac{VP}{VP + FP} \tag{2.5}$$

Já a Revocação (R) é a proporção de instâncias corretamente classificadas como positivas em relação ao total de instâncias verdadeiramente positivas no conjunto de dados. A fórmula é dada por:

$$R = \frac{VP}{VP + FN} \tag{2.6}$$

Dito isso, a F1-Score reflete o desempenho do modelo em prever corretamente todas as instâncias verdadeiramente positivas e também em evitar a classificação incorreta de falsos positivos (Saito; Rehmsmeier, 2015). Conforme definido na equação (2.7):

$$F1 = \frac{2 \times P \times R}{P + R} \tag{2.7}$$

2.1.3.3 Validação Cruzada *K-fold*

A técnica de validação cruzada *k-fold* (Burman, 1989) se baseia na subdivisão do conjunto de dados em k subconjuntos distintos. A cada iteração, o modelo é treinado com $k - 1$ desses subconjuntos, enquanto o restante é reservado para avaliação. Essa repetição do ciclo de treinamento e teste ocorre k vezes, abrangendo todos os subconjuntos.

Ao final do processo, a média dos desempenhos obtidos nas fases de treinamento e teste é calculada. Essa média oferece uma avaliação mais robusta, considerando diferentes combinações de dados de treino e teste ao longo das iterações.

A validação cruzada *k-fold* é especialmente valiosa em contextos onde os dados são escassos, permitindo uma avaliação mais confiável da capacidade de generalização do modelo.

2.2 Trabalhos Relacionados

Nesta seção, inicialmente, é realizada uma análise dos estudos de aprendizado de máquina voltados para o futebol masculino. Esses trabalhos se destacam pela abrangência, uma vez que contam com uma base de dados mais robusta, devido a quantidade de dados disponíveis,

comparada ao futebol feminino. As principais variações encontram-se nos algoritmos e campeonatos utilizados durante sua concepção, garantindo uma abordagem mais completa sobre o tema. Segundo Ganhor *et al.* (2020), apesar da quantidade de artigos disponíveis ser reduzida, é evidente um aumento relativo no interesse por predição de resultados nos últimos anos. Esse crescimento evidencia uma maior atenção e dedicação dos pesquisadores nesse campo específico, indicando o reconhecimento da importância das previsões de resultados no futebol e seu potencial para aprimorar a compreensão e o desempenho no esporte.

Os estudos realizados em Silva (2018) e Silva (2022) abordaram o Campeonato Brasileiro Série A masculino de 2016 e os campeonatos entre 2003 e 2020, respectivamente. O segundo estudo se destaca por ter um conjunto de dados substancialmente maior, permitindo uma análise mais abrangente e detalhada. Ambos os trabalhos coletaram uma variedade de informações estatísticas relacionadas às partidas, incluindo passes, chutes e gols, que foram utilizadas como variáveis na análise. No segundo estudo, foi necessário realizar um tratamento específico nos dados, aplicando a média de gols, uma vez que houve uma variação no formato do campeonato ao longo do tempo (de 24 para 20 clubes). Utilizando a regressão linear múltipla como ferramenta estatística, o primeiro trabalho obteve um R^2 de 93,09%, demonstrando uma forte correlação entre as variáveis analisadas do Campeonato Brasileiro de 2016 e os resultados finais do Campeonato Brasileiro de 2017. Por sua vez, o segundo trabalho, baseado em um conjunto de treinamento mais abrangente dos campeonatos de 2003 à 2020, obteve uma taxa de acerto de 100% na previsão das posições das equipes no Campeonato Brasileiro de 2021, mas apresentou uma precisão de apenas 35% na previsão das pontuações. Esses resultados evidenciam a importância crucial da quantidade de jogos analisados, destacando-a como um fator determinante na qualidade das previsões. Assim, ambos os estudos contribuem significativamente para o aprofundamento e compreensão desse tópico específico no campo do aprendizado de máquina aplicado ao futebol.

Outra técnica empregada para inferir resultados das partidas é a inferência de árvores de decisão. Um exemplo é o trabalho realizado por Silva, Barros e Albuquerque (2020), que utilizou um modelo estatístico baseado em árvore de decisão para prever resultados de jogos de futebol em diferentes campeonatos ao redor do mundo durante o ano de 2019, incluindo o Brasileirão Série A e a Superliga Argentina de Futebol. Além da aplicação de árvore de decisão, a pesquisa utilizou técnicas de avaliação, como a validação cruzada e análise descritiva, como ferramentas estatísticas para análise de dados e resultados. Após o estudo, constatou-se que o modelo proposto acertou 57,5% dos resultados dos jogos, o que sugere que a modelagem por meio de árvores de decisão é uma boa abordagem para a previsão de resultados de jogos de futebol, uma vez que alcança resultados comparáveis à literatura (Schneider, 2020; Silva, 2022).

A previsão de resultados de partidas também pode ser aplicada ao futsal. O estudo conduzido por Duarte e Coppini (2021) utilizou técnicas de aprendizado de máquina para prever os resultados de jogos da Liga Nacional de Futsal (LNF)[1], com base nos dados gerados durante

[1] https://lnfoficial.com.br

o primeiro tempo das partidas. Foram desenvolvidos dez modelos de previsão, baseados em: *Random Forest* (RF) (Breiman, 2001), *Gradient Boosting* (Friedman, 2002), Regressão Logística (RL) (Cox, 1972), *K-Nearest Neighbors* (KNN) (Cover; Hart, 1967), *Support Vector Machine* (SVM) (Vapnik, 1995), *Decision Tree* (DT) (Quinlan, 1986), *Naive Bayes* (Bayes, 1968), *Multi-Layer Perceptron* (MLP) (Rumelhart, 1986), Análise Discriminante Linear (ADL) (Fisher, 1936) e Análise Discriminante Quadrática (ADQ) (Fisher, 1936). Os resultados obtidos demonstraram que os modelos individuais apresentaram um bom desempenho na previsão de resultados específicos, como a vitória do time da casa, alcançando uma acurácia de até 95%. Além disso, foi criado um comitê de votação que combinava as previsões dos diferentes modelos. Esse comitê obteve um desempenho superior na previsão dos resultados gerais, alcançando uma acurácia de quase 79%. Esses resultados indicam que a combinação de vários modelos leva a uma resposta mais precisa no contexto geral.

O uso de modelos combinados também foi usado por Schneider (2020), com o objetivo de avaliar o desempenho de diversos algoritmos de classificação quando os mesmos são utilizados para prever resultados de partidas de futebol contidas entre as temporadas 2000-2001 e 2016-2017 da Premier League. Para alcançar esse objetivo, o estudo avaliou o desempenho de vários algoritmos de classificação, incluindo Regressão Logística, Análise Discriminante Linear, Análise Discriminante Quadrática, *K-Nearest Neighbors*, *Naive Bayes*, *Support Vector Machine* com *kernel* linear (SVML), *Support Vector Machine* com *kernel* RBF (SVMR), *Random Forest*, *Extra Trees* (ET) (Geurts; Ernst; Wehenkel, 2006) e um classificador *ensemble* com classificadores que possuem distribuições de previsões de instâncias significativamente diferentes. Como métrica de avaliação de desempenho dos algoritmos, utilizou-se a acurácia, F1-*score* e análises da matriz de confusão. Ao se observar a matriz de confusão gerada através da análise de um conjunto de dados de teste, constatou-se que a acurácia do classificador *ensemble* foi 1% maior (57%) do que a constatada nos classificadores individuais (56% - SVML e KNN). Assim, a utilização de um ensemble de classificadores pode apresentar-se como uma alternativa viável para aumentar a robustez e capacidade de generalização de um modelo de previsões.

No que se refere ao futebol feminino, Leitner, Zeileis e Hornik (2020) propõem uma abordagem híbrida para prever os resultados da Copa do Mundo Feminina da FIFA 2019, usando a técnica de floresta aleatória combinada com dois métodos diferentes de classificação: o método de classificação de *Poisson* (Karlis; Ntzoufras, 2003) e as habilidades baseadas nas probabilidades dos *bookmakers* (empresas que oferecem serviços de apostas em eventos esportivos e outros eventos). O modelo híbrido foi ajustado a todos os jogos da Copa do Mundo Feminina de 2011 e 2015[2] e, com base nas estimativas resultantes, a Copa de 2019 foi simulada 100.000 vezes para determinar as probabilidades de vitória para todas as 24 equipes participantes. Os resultados mostraram que o modelo híbrido superou outros modelos de previsão em termos de precisão na previsão dos resultados dos jogos da Copa do Mundo Feminina da FIFA 2019. A abordagem

[2] https://www.fifa.com/fifaplus/pt/home

proposta combina informações sobre as equipes, incluindo covariáveis e habilidades estimadas, para melhorar as previsões. A validação cruzada foi usada para estimar a acurácia do modelo híbrido proposto usando medidas de desempenho, como precisão e erro médio absoluto.

No âmbito da utilização de *baselines*, optou-se pelo modelo proposto por Dixon e Coles (1997), que apresenta um modelo paramétrico para prever resultados de partidas de futebol e identificar possíveis ineficiências no mercado de apostas. O estudo se concentra nas ligas e copas de futebol inglesas no período de 1992 a 1995, visando desenvolver uma estratégia de apostas lucrativa. O modelo considera diversas características, incluindo habilidades das equipes, efeito de jogar em casa, desempenho recente, capacidade de ataque e defesa, assim como a natureza do futebol em si. A abordagem adotada é fundamentada em uma distribuição de *Poisson* bivariada para o número de gols marcados por cada equipe, com parâmetros vinculados ao desempenho histórico. Os resultados indicam que a estratégia de apostas proposta, baseada nas probabilidades calculadas pelo modelo em comparação às oferecidas pelas casas de apostas, resulta em um retorno esperado positivo. Isso sugere que o modelo tem potencial para identificar oportunidades de apostas lucrativas, mesmo quando consideramos o viés incorporado nas probabilidades das casas de apostas.

Este estudo compartilha semelhanças com trabalhos anteriores ao utilizar o aprendizado de máquina para desenvolver um modelo de previsão de resultados no futebol. No entanto, se destaca por focar exclusivamente no contexto do futebol feminino brasileiro, proporcionando uma perspectiva única e relevante para essa modalidade. Além disso, este trabalho oferece um ambiente propício para futuras pesquisas na aplicação da abordagem híbrida no campo esportivo, especialmente para equipes e setores de análise de desempenho, permitindo estudos específicos e personalizados para a realidade do futebol feminino no país. Em suma, esta pesquisa contribui de forma complementar, expandindo e aprimorando os resultados já existentes na aplicação do aprendizado de máquina no futebol, oferecendo uma valiosa perspectiva no contexto feminino brasileiro.

3 Desenvolvimento

Neste capítulo, são apresentados os métodos e técnicas empregados na elaboração deste trabalho.

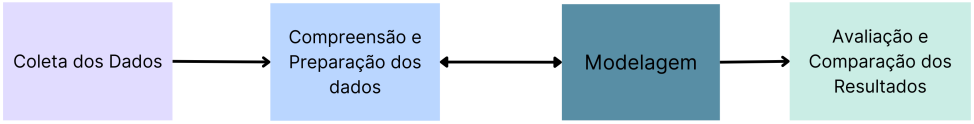
3.1 Metodologia

Nesta seção, são apresentados os métodos e procedimentos empregados na elaboração do trabalho, desde a coleta e pré-processamento dos dados até a definição e avaliação dos classificadores individuais.

Inicialmente, conforme ilustrado na Figura 3.1, foram conduzidas análises em várias plataformas de dados que continham informações relacionadas ao Campeonato Brasileiro feminino, com o propósito de selecionar aquela que melhor se adequasse aos objetivos deste estudo. Destaca-se que um dos aspectos fundamentais deste trabalho foi a integração de dados provenientes de times, jogadores e partidas de cinco ligas ao redor do mundo na temporada 2022-2023, abrangendo os países Estados Unidos, Itália, Inglaterra, Espanha e a Copa Europeia. Esse processo resultou na formação de um conjunto de treinamento composto por 780 partidas. O conjunto de teste, por sua vez, foi constituído por 134 jogos do Campeonato Brasileiro Feminino de 2022. Na sequência, visando garantir a qualidade dos resultados, foi feito a compreensão e pré-processamento dos dados. Essa etapa inclui a eliminação de atributos irrelevantes, o tratamento de valores ausentes, a transformação de dados, a normalização de atributos, a adição de um atributo para conter o resultado da partida.

No processo de implementação, realizou-se a seleção de atributos, estabelecendo as bases para a subsequente predição de resultados. Além disso, foram selecionados classificadores, representando distintos algoritmos de aprendizado de máquina, com base em seu desempenho em estudos anteriores. Posteriormente, avaliou-se os resultados preliminares obtidos, comparando-os com a *baseline* proposta por Dixon e Coles (1997).

Figura 3.1 – Metodologia utilizada neste trabalho

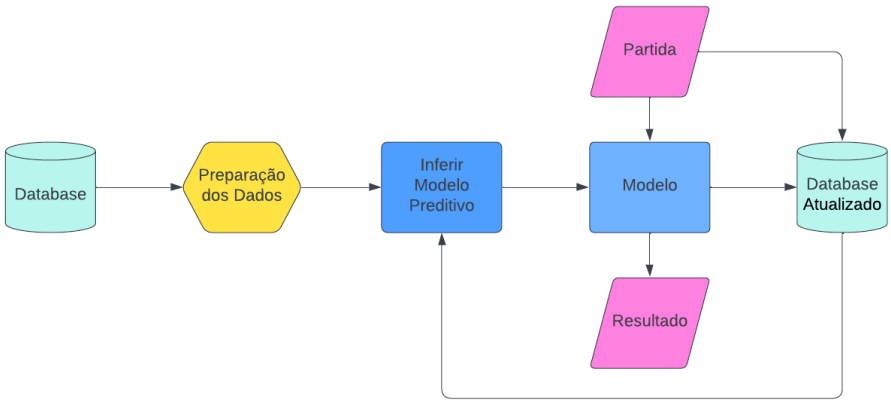


Fonte: Autora

No processo de aplicação do modelo preditivo, foram adotados dois cenários distintos: estático e incremental.

O cenário estático, referido como Cenário 1, utiliza os conjuntos de treino e teste de maneira isolada. Isso significa que ele realiza o treinamento com os 780 jogos disponíveis e, em seguida, faz a previsão dos resultados das 134 partidas de teste. Já o cenário incremental, ou Cenário 2, diferencia-se ao incrementar o conjunto de treinamento a cada iteração, ou seja, após prever o resultado de uma partida, os dados desta partida, juntamente com o resultado real, são adicionados ao conjunto de exemplos de treinamento. A Figura 3.2 mostra uma representação visual do fluxo adotado no cenário incremental neste estudo.

Figura 3.2 – Método Incremental



Fonte: Autora

3.2 Caracterização das bases de Dados

A primeira etapa da metodologia deste trabalho consiste na aquisição do conjunto de dados. A coleta de dados foi realizada a partir da plataforma *FootyStats*, um site de estatísticas e análises de futebol que oferece informações sobre diversas ligas ao redor do mundo. Para coletar esses dados foi necessário obter o pacote *premium*, e baixar os arquivos pelo aplicativo.

Foram escolhidos cinco ligas para formar a base de dados: *National Women's Soccer League* (EUA), *FA Women's Super League* (Inglaterra), *Serie A Woman* (Itália), Liga F (Espanha), *UEFA Women's Champions League* da temporada 2022/2023, totalizando 780 partidas. Essas ligas foram escolhidas devido à sua relevância internacional e à disponibilidade abrangente de dados.

O conjunto é composto por três arquivos distintos no formato CSV: partidas, jogadoras, times. Nas subseções a seguir, descreve-se cada um deles.

3.2.1 Partidas

Este conjunto de dados possui vários atributos que contêm informações relevantes sobre os jogos, como os nomes dos times da casa e visitantes, as médias pré-jogo de pontos por

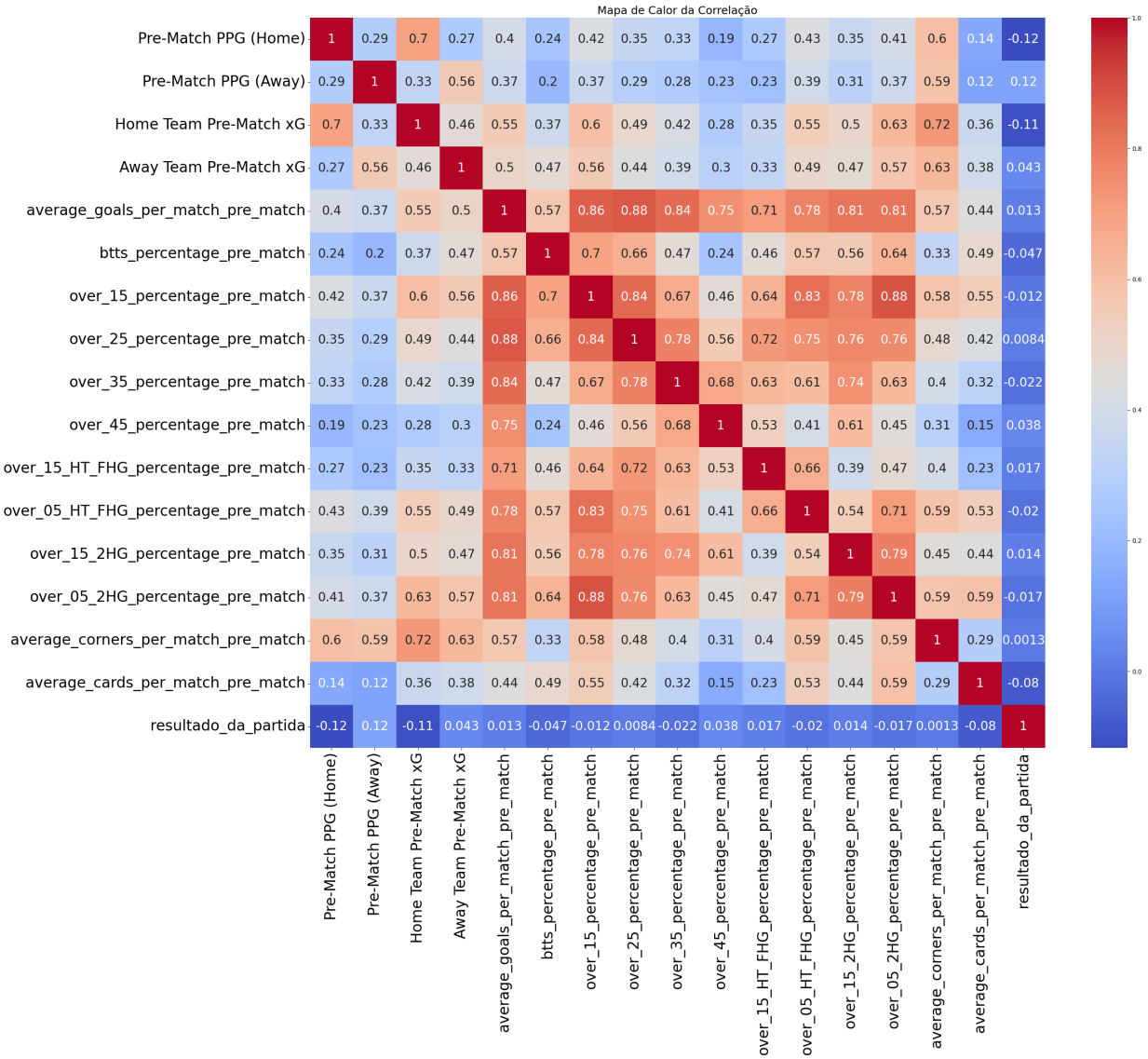
partida para os times da casa e visitantes, os gols marcados por cada time, o número total de gols, os gols no intervalo, os tempos em que os gols foram marcados, os escanteios, os cartões amarelos e vermelhos, os chutes a gol, as faltas cometidas, a posse de bola, os valores pré-jogo de expectativa de gols para cada time, as probabilidades de ambas as equipes marcarem, entre outras informações.

De forma que permite o estudo de diferentes aspectos, como o desempenho das equipes, o comportamento ao longo do jogo, as estatísticas de gols, escanteios, cartões, chutes a gol e posse de bola. Além disso, as informações pré-jogo podem ser utilizadas para avaliar o desempenho esperado das equipes.

O dataset contém 66 atributos, sendo 7 atributos categóricos nominais, 1 atributo categórico ordinal, 26 atributos discretos e 33 contínuos.

Uma análise incluiu a investigação das correlações entre os atributos das partidas. Como ilustrado na Figura 3.3, o foco foi especificamente nos atributos anteriores ao início das partidas, totalizando 23 atributos. Uma vez que os outros não contribuirão para a predição. Isso se deve ao fato de o modelo não ter acesso a informações sobre a partida que está prestes a prever, como o número de cartões, por exemplo.

Figura 3.3 – Mapa de calor da correlação no conjunto de dados de partidas.



Fonte: Autora

Os resultados, presentes na Figura 3.3, indicam que não há nenhum atributo no conjunto de dados que mantenha uma relação proporcional com o desfecho da partida, essa correlação negativa sugere uma influência inversamente proporcional entre o atributo pontos por jogo (PPG) e os resultados das partidas, indicando que esse atributo interfere pouco nos resultados finais das partidas.

A correlação de *Pearson* varia de -1 a 1, de modo que 1 indica que à medida que uma variável aumenta, a outra também aumenta proporcionalmente. 0 indica ausência de correlação linear, ou seja, não há uma relação linear entre as variáveis. E, -1 indica que quando uma variável aumenta, a outra diminui proporcionalmente.

3.2.2 Jogadoras

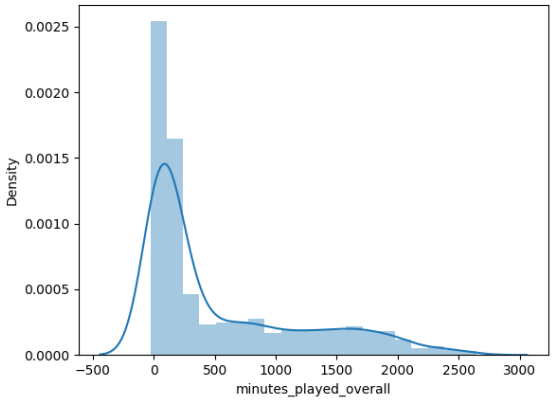
Este conjunto de dados é uma compilação detalhada de informações sobre as jogadoras inscritas nos campeonatos, abrangendo 277 atributos distintos. Entre esses atributos, encontramos 8 que são categóricos nominais, fornecendo informações sobre a liga, posição das jogadoras, seu atual clube, nacionalidade e outros aspectos relevantes.

O conjunto de dados também apresenta 142 atributos contínuos, incluindo estatísticas essenciais como o tempo total de jogo (minutos jogados), o número de partidas disputadas, a quantidade de gols marcados, assistências, cartões amarelos e vermelhos recebidos, bem como defesas de pênaltis, entre outros aspectos fundamentais do desempenho de uma jogadora.

Além disso, são 127 atributos discretos, fornecendo informações específicas sobre diversos aspectos técnicos e táticos das jogadoras. Isso inclui variáveis como expectativa de gols (xG) e expectativa de assistências (xA), que auxiliam na análise de desempenho, além de dados sobre o número de passes, chutes a gol, cruzamentos, dribles, duelos ganhos, interceptações, etc.

Para a análise da distribuição dos dados das jogadoras, foram selecionados dois atributos: "minutos jogados"e "idade", como evidenciado nas Figuras 3.4 e 3.5.

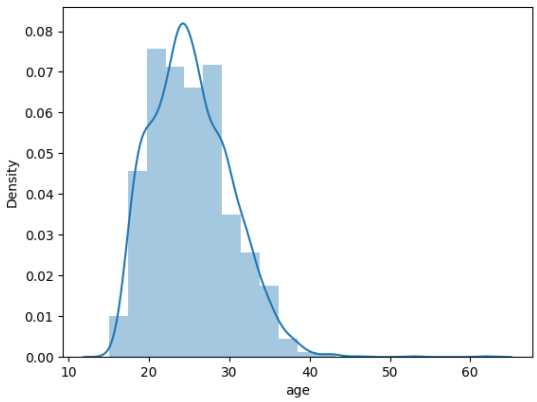
Figura 3.4 – Distribuição dos valores referentes a minutos jogados



Fonte: Autora

Lamentavelmente, o gráfico é afetado pela presença de muitos dados com valor 0, o que revela a lacuna de informações no contexto do campeonato feminino. Em média, cada time disputou aproximadamente 1400 minutos, o que implica que, para as jogadoras com valores registrados, esses geralmente variam entre 500 e 1000 minutos.

Figura 3.5 – Distribuição dos valores referentes a Idade das Jogadoras.



Fonte: Autora

A distribuição mostra que maioria das jogadoras encontra-se na faixa etária jovem, concentrando-se predominantemente entre os 20 e 30 anos. Este padrão sugere que jogadoras mais experientes podem estar inclinadas a se aposentar mais cedo ou buscar mercados secundários do futebol feminino.

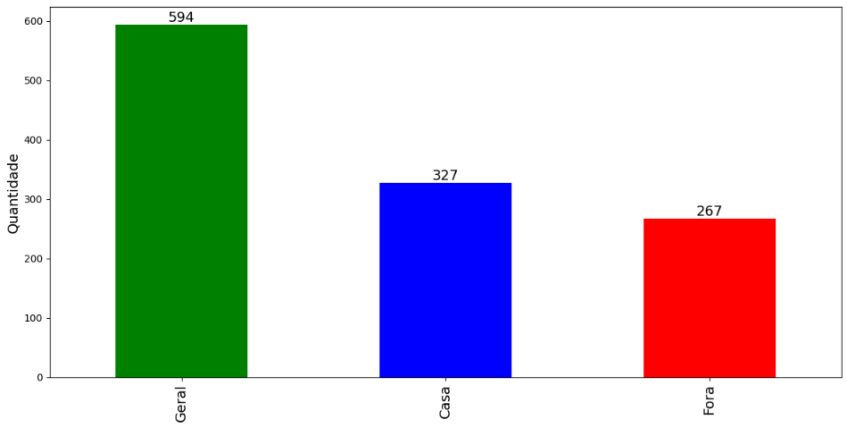
3.2.3 Times

Este conjunto de dados é uma fonte valiosa de informações estatísticas sobre as equipes. Ele contém uma ampla variedade de dados, incluindo o número de jogos disputados em casa e fora, vitórias, empates e derrotas, gols marcados e sofridos, posição na tabela, médias de gols por partida, porcentagens de jogos sem sofrer gols (*clean sheets*), entre outros. Essas informações permitem uma análise detalhada do desempenho das equipes em diferentes aspectos.

O conjunto de dados é composto por um total de 293 atributos, que estão distribuídos em: 4 atributos categóricos nominais, que incluem informações como o nome dos times e outras categorias não hierárquicas. Além disso, há 3 atributos categóricos ordinais, que representam dados com ordem e importância específica. O conjunto também inclui 67 atributos discretos, que capturam informações quantitativas distintas, e 219 atributos contínuos.

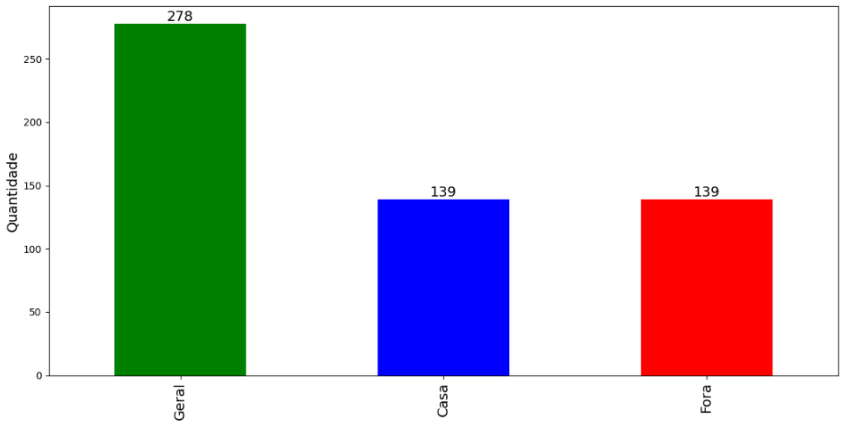
Para a análise da distribuição dos dados referentes aos times, foram selecionados atributos relacionados aos resultados, tais como a quantidade de vitórias, empates e derrotas. As Figuras 3.6, 3.7 e 3.8 ilustram esses aspectos.

Figura 3.6 – Quantidade de vitórias.



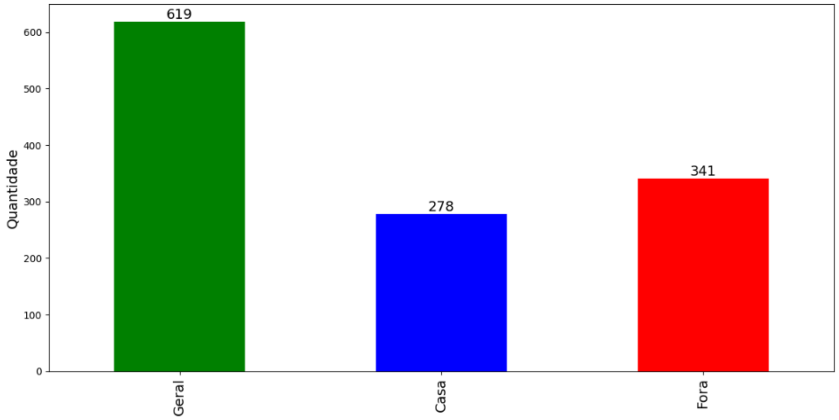
Fonte: Autora

Figura 3.7 – Quantidade de empates.



Fonte: Autora

Figura 3.8 – Quantidade de derrotas.



Fonte: Autora

Os gráficos evidenciam que jogar em casa é uma vantagem para alcançar vitórias, enquanto a maioria das derrotas ocorre fora de casa. Quanto aos empates, no conjunto de dados analisado, não parece haver uma distinção significativa entre partidas disputadas em casa ou fora.

3.3 Pré-processamento

Na segunda etapa do desenvolvimento, o pré-processamento desempenha um papel essencial, seguindo o conceito proposto por Batista *et al.* (2003). Seu objetivo é preparar os dados para a próxima fase, a extração de conhecimento, tornando o modelo de previsões mais efetivo.

Deste modo, considerando as características dos conjuntos de dados apresentados na seção anterior, nota-se que é necessário fazer algumas alterações nos dados. Inicialmente, atributos considerados irrelevantes por estarem relacionados à casas de apostas, visto que tais valores já passaram por um modelo anterior, foram removidos, tornando o conjunto de dados mais conciso e focado nas informações relevantes.

Posteriormente, foram identificados e eliminados os registros que continham valores relacionados a partida. Essa etapa foi especialmente importante no *dataset* de jogos, onde uma quantidade significativa de linhas estava afetada por esses valores. Essas ações visaram garantir a confiabilidade dos dados e evitar distorções na análise.

Os atributos remanescentes em cada conjunto foram[1]:

- **Partidas:** ‘date_GMT’, ‘home_team_name’, ‘away_team_name’, ‘Pre-Match PPG (Home)’, ‘Pre-Match PPG (Away)’, ‘home_team_goal_count’, ‘away_team_goal_count’, ‘Home Team Pre-Match xG’, ‘Away Team Pre-Match xG’, ‘average_goals_per_match_pre_match’, ‘btts_percentage_pre_match’, ‘over_15_percentage_pre_match’, ‘over_25_percentage_pre_match’, ‘over_35_percentage_pre_match’, ‘over_45_percentage_pre_match’, ‘over_15_HT_FHG_percentage_pre_match’, ‘over_05_HT_FHG_percentage_pre_match’, ‘over_15_2HG_percentage_pre_match’, ‘over_05_2HG_percentage_pre_match’, ‘average_corners_per_match_pre_match’, ‘average_cards_per_match_pre_match’, ‘stadium_name’.
- **Jogadoras:** Devido à quantidade de atributos remanescentes (268), serão listados os atributos excluídos: ‘birthday’, ‘birthday_GMT’, ‘shirt_number’, ‘additional_info’, ‘xg_faced_per_90_overall’, ‘xg_faced_per90_percentile_overall’, ‘xg_faced_per_game_overall’, ‘xg_faced_total_overall’, ‘sm_minutes_played_per90_percentile_overall’.
- **Times:** Os atributos ‘common_name’ e ‘country’ foram removidos devido à falta de relevância para o estudo.

[1] Descrições sobre esses atributos estão no Apêndice A

Além das exclusões, foram realizadas modificações adicionais nos conjuntos de dados. Essas modificações tinham como objetivo tratar valores que inicialmente eram representados como -1, indicando a ausência de dados, e transformá-los em 0 para refletir essa condição, nos casos de dados ausentes que poderiam ser modificada. Ademais, foi necessário ajustar a forma como as datas dos jogos estavam sendo apresentadas, separando a data e a hora, a fim de aprimorar a análise e compreensão dos dados.

Para a predição, um atributo foi adicionado para representar o resultado das partidas, com os valores 0 para Empate, 1 para Vitória do Time Mandante e 2 para Vitória do Time Visitante.

Em seguida os atributos foram normalizados. A normalização por padronização é uma técnica importante para lidar com atributos contínuos que possuem diferentes escalas e intervalos. Ela visa transformar os atributos de forma que eles compartilhem a mesma medida de posição e variação, geralmente representada pela média e pelo desvio-padrão (Faceli *et al.*, 2021). A fórmula que descreve a transformação de um atributo j em um objeto i utilizando a padronização é dada por:

$$x_{ij}^{'} = \frac{x_{ij} - \bar{x}_j}{\sigma_j} \tag{3.1}$$

sendo $x_{ij}^{'}$ é o valor normalizado do atributo j no objeto i, x_{ij} é o valor original do atributo j no objeto i, $\bar{x}_j$ representa a média do atributo j e σ_j é o desvio-padrão do atributo j.

A aplicação dessa técnica garante que todos os atributos tenham média zero e desvio-padrão unitário, tornando-os comparáveis e evitando que atributos com escalas diferentes tenham influências desproporcionais em algoritmos de aprendizado de máquina.

Por fim, o processo de pré-processamento envolveu a integração dos dados provenientes dos três arquivos para representar os dados das partidas. Dessa forma, as jogadoras foram divididas conforme a posição (goleira, defensora, meio-campo ou atacante) e suas características (atributos) específicas, em seguida, foi calculada a média dos valores dos atributos das jogadoras de acordo com o time e a posição das mesmas. A Tabela 3.1 e a Figura 3.9 ilustram como foi realizado esse processo:

Tabela 3.1 – Divisão de atributos

Posição	**Atributos**
Goleira	Nome Completo, Idade, Posição, Clube Atual, Minutos Jogados no Total, Partidas sem Sofrer Gols, Gols Sofridos, Cartões Amarelos Recebidos, Média de Cartões por 90 Minutos, Defesas por Jogo, Valor de Mercado, Bloqueios por Jogo, Avaliação Total, Socos Totais, Percentual de Defesa em Penalidades, Penalidades Defendidas, Penalidades Cometidas.
Defensora	Nome Completo, Idade, Posição, Clube Atual, Minutos Jogados no Total, Gols Marcados, Assistências, Partidas sem Sofrer Gols, Gols Sofridos, Cartões Amarelos, Cartões Vermelhos, Minutos por Gol Marcado, Minutos por Gol Sofrido, Avaliação Média, Assistências por Jogo, Passes por Jogo, Ações Defensivas por Jogo, Desarmes por Jogo, Chutes por Gol Marcado, Expectativa de Gols por Jogo, Disputas Aéreas Ganhas por Jogo, Chutes Totais, Jogos Substituído (Out), Jogos Substituído (In), Jogos Iniciados, Intercepções Totais, Desarmes por Jogo, Penalidades Cometidas, Bolas na Trave.
Meio-Campista	Nome Completo, Idade, Posição, Clube Atual, Minutos Jogados no Total, Gols Marcados, Assistências, Partidas sem Sofrer Gols, Gols Sofridos, Cartões Amarelos, Cartões Vermelhos, Minutos por Gol Marcado, Avaliação Média, Assistências por Jogo, Passes por Jogo, Ações Defensivas por Jogo, Desarmes por Jogo, Chutes por Gol Marcado, Expectativa de Gols por Jogo, Disputas Aéreas Ganhas por Jogo, Chutes Totais, Jogos Substituído (Out), Jogos Substituído (In), Jogos Iniciados, Intercepções Totais, Desarmes por Jogo, Penalidades Cometidas, Bolas na Trave.
Atacante	Nome Completo, Idade, Posição, Clube Atual, Minutos Jogados no Total, Gols Marcados, Assistências, Partidas sem Sofrer Gols, Gols Sofridos, Cartões Amarelos, Cartões Vermelhos, Minutos por Gol Marcado, Avaliação Média, Assistências por Jogo, Passes por Jogo, Chutes por Gol Marcado, Expectativa de Gols por Jogo, Disputas Aéreas Ganhas por Jogo, Chutes Totais, Jogos Substituído (Out), Jogos Substituído (In), Jogos Iniciados, Intercepções Totais, Desarmes por Jogo, Penalidades Cometidas, Bolas na Trave.

A seleção desses atributos foi orientada pelas características específicas de cada posição. Por exemplo, defesas por jogo é um atributo específico para goleiras.

Figura 3.9 – Método utilizado para calcular as médias dos atributos de acordo com a posição e o time de cada jogadora.

```
mean_attributes_goalkeepers_by_team = goalkeepers.groupby('Current Club').mean()
mean_attributes_defenders_by_team = defenders.groupby('Current Club').mean()
mean_attributes_midfielders_by_team = midfielders.groupby('Current Club').mean()
mean_attributes_forwards_by_team = forwards.groupby('Current Club').mean()
```

Fonte: Autora

Dado que não era viável incluir todas as jogadoras no conjunto de treinamento, optou-se por calcular a média dos atributos com base na posição e no time de cada jogadora. Por exemplo, todos os valores dos atributos das goleiras de um determinado time foram reunidos, e a média foi calculada a partir desses dados.

Posteriormente, houve a junção dos dados referentes a jogadoras, equipes e partidas. A integração desses seguiu os seguintes passos:

1. **Identificação das Partidas:** Primeiramente, foi necessário identificar quais times estarão envolvidos em cada partida, selecionando os nomes das equipes pelo atributo *home_team_name* (equipe da casa) e *away_team_name* (equipe visitante).

2. **Seleção de Atributos:** Para cada partida, foram selecionados atributos específicos das equipe. Isso implica na extração das informações relevantes dos conjuntos de dados de equipes e das jogadoras.

3. **Concatenação dos Dados:** Com as informações tratadas e selecionadas, os dados provenientes de todos os arquivos foram concatenados em um único registro (instância da partida).

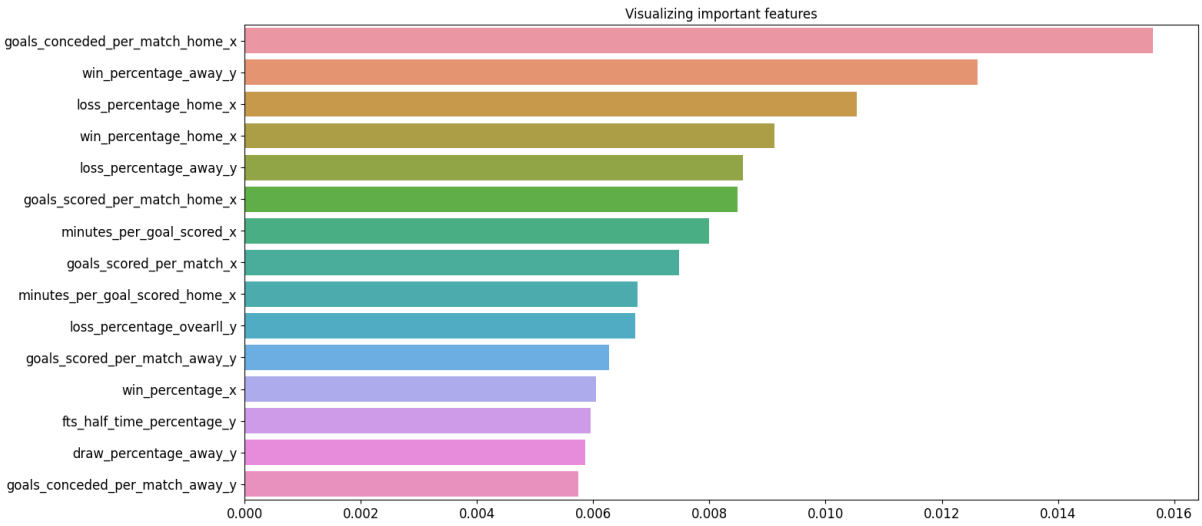
A descrição detalhada desses atributos pode ser encontrada no Apêndice A.

Para otimizar o conjunto de atributos, realizou-se a seleção de atributos com base em critérios bem definidos, estabelecendo as bases necessárias para o desenvolvimento da pesquisa. Para chegar na seleção final dos 200 melhores atributos (Tabela A), utilizou-se a técnica *Mean Decrease Accuracy* (MDA) para avaliar a importância das *features* em um modelo de *Random Forest* para regressão. O código configura uma busca em grade (*GridSearchCV*) para encontrar os melhores hiperparâmetros para o modelo. A validação cruzada é realizada usando *ShuffleSplit*. Com os parâmetros otimizados, o modelo é treinado e a técnica MDA é aplicada para avaliar a importância das *features*. A MDA consiste em embaralhar aleatoriamente os valores de uma *feature* por vez e observar o impacto no desempenho do modelo. Os resultados são então armazenados em um dicionário. Este dicionário registra as mudanças na métrica de desempenho (R^2) para cada *feature* quando seus valores são embaralhados. Finalmente, as *features* mais importantes

são identificadas ordenando os índices de acordo com a média das mudanças na métrica de desempenho.

A Figura 3.10 exibe alguns dos atributos. Dada a limitação de espaço, impossibilitando a apresentação de um *ranking* completo com 200 atributos, optou-se por destacar os 15 melhores.

Figura 3.10 – Os 15 melhores atributos selecionados



Fonte: Autora

A análise revela que a localização da partida, seja em casa ou fora, é um dos fatores mais significativos na determinação dos resultados, corroborado pelas Figuras 3.6 e 3.8.

Com essas transformações, buscou-se mitigar quaisquer lacunas ou ambiguidades nos dados, proporcionando uma base sólida para as investigações posteriores.

3.4 Indução dos Classificadores

Tendo os dados corretamente ajustados, a terceira etapa do desenvolvimento é a definição do modelo de predição de resultados. Para tanto, foram treinados classificadores para prever vitória, empate ou derrota de um time. O treinamento foi realizado buscando os melhores valores de parâmetros para cada técnica de indução de classificadores, via *Grid Search* e validação cruzada. Os classificadores utilizados foram: *Random Forest*, *K-Nearest Neighbors*, *XGBoost*. Cada classificador foi selecionado por sua diversidade e pela utilização em estudos anteriores, como referenciado por Schneider (2020) e Silva, Barros e Albuquerque (2020). A avaliação dos classificadores foi feita por meio da acurácia e F1-Score.

4 Avaliação Experimental

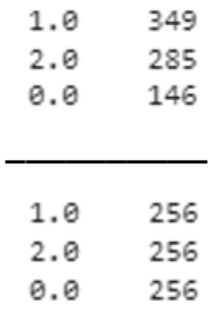
4.1 Configuração dos Experimentos

A implementação do código foi realizada em *Python*, fazendo uso de bibliotecas fundamentais como pandas para manipulação de arquivos CSV, *numpy* para cálculos diversos, *matplotlib* para representação gráfica, e *scikit-learn* para aplicação de técnicas de aprendizado de máquina (Pedregosa *et al.*, 2011), junto a outras ferramentas relevantes.

Quanto à estruturação dos dados de entrada, o conjunto de exemplos de treinamento abrangeu as temporadas de 2022/2023, totalizando 780 partidas. O conjunto de testes incorporou a temporada de 2022 do Campeonato Brasileiro Feminino, compreendendo 134 partidas.

Para a aplicação dos algoritmos de classificação, foi adotada uma abordagem de aprendizagem supervisionada, na qual os classificadores indicam se ocorrerá vitória do mandante (classe 1), empate (classe 0) ou vitória do visitante (classe 2). Após a atribuição das classes, os dados foram balanceados como mostra a figura 4.1 do antes e depois do balanceamento.

Figura 4.1 – Distribuição dos resultados das partidas antes e depois do balanceamento.



Fonte: Autora

Os dados foram divididos em X e Y, no qual o conjunto X são os melhores 200 atributos, obtidos através do método MDA descrito na seção 3.3, e o Y representa o resultado da partida. Em *Python*, a representação foi a seguinte:

```
X_train = merged_df[selected_features1]
Y_train = merged_df['resultado_da_partida']

X_test = brasileirao[selected_features1]
Y_test = brasileirao['resultado_da_partida']
```

sendo que '`merged_df`' e '`brasileirao`' são os conjuntos de treinamento e teste, respectivamente.

No processo de aplicação do modelo preditivo, foram adotados dois métodos distintos. O primeiro consiste na aplicação do modelo sobre os jogos de teste, mantendo o conjunto de treinamento homogêneo ao longo de todo o processo. Esta abordagem permite avaliar o desempenho do modelo em prever resultados sem a influência da atualização contínua do conjunto de treinamento. Já o segundo método adotado é o incremental, onde a cada partida predita, o resultado real dessa partida é incorporado ao conjunto de treinamento. Dessa forma, o modelo é constantemente atualizado com as informações mais recentes, proporcionando uma adaptação contínua às nuances do contexto dos jogos.

A escolha dos hiperparâmetros para cada algoritmo de classificação baseou-se em testes com o conjunto de treinamento, sendo a abordagem mais eficaz a exploração de diversas combinações e a avaliação do desempenho de cada modelo.

A seleção dos melhores parâmetros foi realizada utilizando a técnica de *GridSearch*. E os valores de cada parâmetro foram:

- ***Random Forest***: '*max_depth*': 30, '*n_estimators*': 200
- ***XGBoost***: '*learning_rate*': 0.1, '*max_depth*': 5
- **KNN**: '*n_neighbors*': 3, '*weights*': '*distance*'

Além da acurácia e *F1-Score*, para uma compreensão mais minuciosa da performance dos modelos, recorreu-se a matrizes de confusão por meio do método *confusion_matrix*, evidenciando a frequência de classificação para cada classe do modelo.

4.2 Baseline

O estudo adotou o modelo proposto por Maher (1982) como base, incorporando modificações essenciais para possibilitar a inclusão simultânea de conjuntos de dados incompletos e informações de diferentes divisões. Além disso, foram implementadas ajustes que permitem considerar flutuações no desempenho das equipes ao longo do tempo. A análise do modelo principal abrangeu dados provenientes de 92 times das três divisões do futebol inglês, abrangendo um período de três anos e totalizando 6.629 partidas. A aplicação dessa *baseline* envolveu os conjuntos de treinamento e teste descritos na Seção 4.1.

Ao examinar a distribuição dos resultados, observou-se uma proporção de frequências com 46% das partidas resultando em vitórias para os times da casa, 27% em empates e outros 27% em vitórias para os times visitantes. Dixon e Coles (1997) sugere uma abordagem mais sensível às tendências recentes para uma análise mais precisa e contextualizada.

Essas adaptações no modelo visam proporcionar uma estrutura mais flexível e abrangente, capaz de lidar com nuances específicas dos dados e refletir de maneira mais fiel a dinâmica do desempenho das equipes ao longo do tempo e em diferentes contextos de competição.

4.2.1 Modelo Matemático

O modelo do Maher (1982) considera apenas partidas dentro de um período pré-definido, por exemplo, desde o início da temporada, enquanto a exponencial negativa negocia mais fortemente os resultados das partidas à medida que o tempo passe. O modelo refinado do Dixon e Coles (1997) pode ser expresso matematicamente da seguinte forma:

$$L(\alpha_i, \beta_i, \rho, \gamma, i = 1, \ldots, n) = \prod_{k \in A_t} \{\tau_{\lambda_k, \mu_k}(x_k, y_k) \frac{e^{-\lambda} \lambda^{x_k}}{x_k!} \frac{e^{-\mu} \mu^{y_k}}{y_k!}\}^{\phi(t-t_k)} \tag{4.1}$$

sendo t_k representa o tempo em que a partida k foi jogada, $A_t = \{k : t_k < t\}$ (conjunto de partidas jogadas antes do tempo t), α, β, γ e τ são definidos como taxa de ataque, taxa de defesa, fator casa e o fator de correção, respectivamente. ϕ representa a função de ponderação não crescente. O artigo do Dixon e Coles (1997), definiu $\phi(t)$ como uma exponencial negativa com taxa ξ.

O fator de correção $\tau_{\lambda,\mu}(x, y)$ utilizado para lidar com o viés em torno de empates com poucos gols, especialmente em resultados como 0-0, tem uma definição específica para diferentes cenários:

- Se $x = y = 0$, ou seja, ambos os times não marcam gols, então $\tau_{\lambda,\mu}(x, y) = 1 - \lambda\mu\rho$. Isso implica uma redução na probabilidade de 0-0, levando em consideração os efeitos de empates de 0-0 em partidas de futebol;
- Se $x = 0$ e $y = 1$, ou seja, o time da casa não marca gols e o time visitante marca um gol, então $\tau_{\lambda,\mu}(x, y) = 1 - \lambda\rho$. Isso reflete uma correção para o resultado de 0-1;
- Se $x = 1$ e $y = 0$, ou seja, o time da casa marca um gol e o time visitante não marca gols, então $\tau_{\lambda,\mu}(x, y) = 1 + \mu\rho$. Isso é uma correção semelhante, mas para o resultado inverso de 1-0;
- Se $x = y = 1$, ou seja, ambos os times marcam exatamente um gol, então $\tau_{\lambda,\mu}(x, y) = 1 - \rho$. Isso representa uma correção para empates de 1-1;
- Para todos os outros casos, $\tau_{\lambda,\mu}(x, y) = 1$, indicando que não há correção necessária.

4.2.2 Avaliação do Modelo

Para determinar o valor de ξ, em (Dixon; Coles, 1997), o modelo é avaliado com base nas previsões de resultados de partidas. A função objetivo é redefinida como:

$$S(\xi) = \sum_{k=1}^{N} (\delta_k^H \log p_k^H + \delta_k^A \log p_k^A + \delta_k^D \log p_k^D) \tag{4.2}$$

sendo δ captura o resultado da partida (por exemplo, δ_k^H = 1 se a partida k terminou com vitória em casa), e os termos p são as probabilidades dos resultados da partida.

O ξ obtido no artigo foi 0.0065 e foi esse valor utilizado na implementação da *baseline*.

Neste trabalho, usou-se a implementação disponível em *Predicting Football Results With Statistical Modelling: Dixon-Coles and Time-Weighting*.

4.3 Resultados

A avaliação do desempenho preditivo dos classificadores nas partidas foi conduzida utilizando os módulos *accuracy_score* e *f1_score* da biblioteca *sklearn*. As técnicas de indução de classificadores empregadas foram: *Random Forest*, *XGBoost*, e KNN (*K-Nearest Neighbors*) (Zaki; Meira, 2020). Essas técnicas foram escolhidas com base nos resultados obtidos nos trabalhos relacionados, especialmente o do Schneider (2020).

Inicialmente, foram conduzidos testes exclusivamente com dados do Campeonato Brasileiro de 2022 para avaliar a viabilidade de utilizá-lo como conjunto de treino e teste. Contudo, os resultados iniciais não foram satisfatórios, e devido ao limitado volume de dados disponíveis, optou-se por construir um conjunto de exemplos de treinamento mais robusto. Os resultados podem ser observados na Tabela 4.1.

Tabela 4.1 – Desempenho dos classificadores com dados do Campeonato Brasileiro no cenário estático

Classificador	Acurácia	*F1-Score*
RF	0.5807	0.5401
KNN	0.5763	0.5553
XGB	0.5593	0.5245

Fonte: Produzida pela própria autora.

No contexto analisado, destaca-se o desempenho positivo do KNN. A utilização de exemplos apenas do Campeonato Brasileiro contribuiu para a formação de uma vizinhança mais próxima, resultando em um desempenho superior. Em contrapartida, os classificadores RF e XGB apresentaram resultados satisfatórios, mas ainda inferiores aos obtidos nos outros experimentos.

A Tabela 4.2 apresenta a acurácia média e o *F1-Score* obtidos por cada classificador no cenário estático.

Tabela 4.2 – Resultados dos classificadores no cenário estático

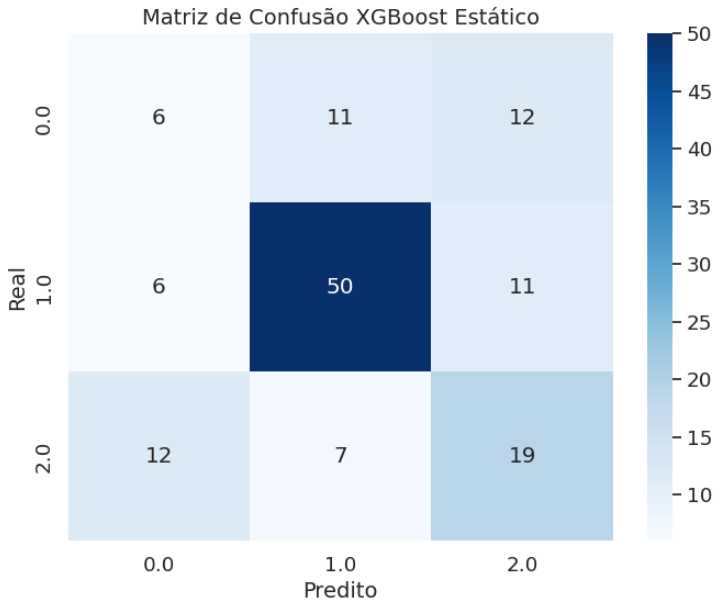
Classificador	Acurácia	*F1-Score*
RF	0.5741	0.5518
KNN	0.5296	0.5231
XGB	0.6109	0.5877
Baseline	0.3398	0.2977

Fonte: Produzida pela própria autora.

Os resultados da *baseline* foram significativamente mais baixos quando comparados aos outros algoritmos, e podem ser atribuídos à limitação da *baseline*, que se baseia exclusivamente na eficácia de ataque, defesa e no fator de jogar em casa. A *baseline*, ao depender de um número reduzido de atributos, enfrenta desafios ao abordar uma tarefa tão complexa como a predição de resultados de partidas de futebol.

Os algoritmos *Random Forest* e *XGBoost* apresentaram os melhores resultados, com uma ligeira vantagem para o XGB. Uma possível explicação para essa diferença pode estar relacionada à natureza da profundidade máxima da árvore dos algoritmos RF e XGB, uma vez que o XGB é construído com árvores de decisão de profundidade reduzida (Chen; Guestrin, 2016). Árvores mais profundas tendem a se ajustar bem aos dados conhecidos, mas podem ter desempenho inferior em dados desconhecidos. A Figura 4.2 exibe a matriz de confusão do *XGBoost*, destacando uma dificuldade do modelo em prever empates (classe 0).

Figura 4.2 – Matriz de Confusão do XGB estático



Fonte: Autora

Essa matriz indica uma tendência forte de vitória dos times mandantes, no entanto, essa precisão não foi igualmente refletida nas previsões das outras classes. A classe dos empates, por exemplo, teve uma taxa de acerto de apenas 35%.

Os resultados considerando o cenário incremental foram semelhantes para todos os algoritmos aplicados, com exceção do KNN, que apresentou uma redução significativa no rendimento. A Tabela 4.3 exibe os melhores resultados dos classificadores.

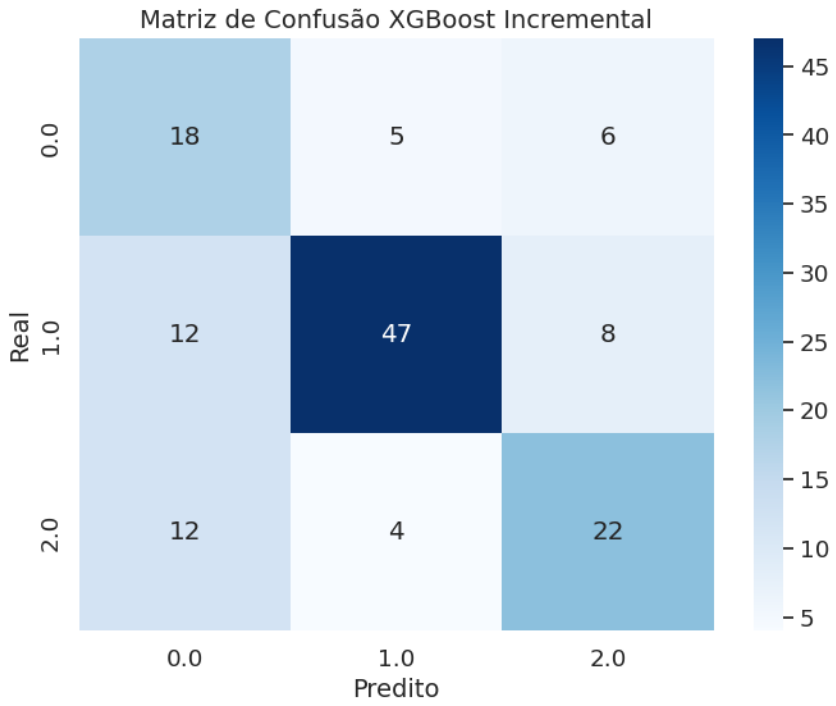
Tabela 4.3 – Melhores Resultados dos classificadores com o método incremental

Classificador	Acurácia	*F1-Score*
RF	0.6044	0.6150
KNN	0.3731	0.3640
XGB	0.6343	0.6518

Fonte: Produzida pela própria autora.

Novamente, o *XGBoost* demonstrou o melhor desempenho, apresentando uma leve melhora que pode ser atribuída à constante atualização do conjunto de treinamento, tornando o modelo mais adaptado. A Figura 4.3 apresenta a matriz de confusão desse cenário.

Figura 4.3 – Matriz de Confusão do XGB incremental



Fonte: Autora

A matriz do cenário incremental, apesar de apresentar uma maior dificuldade em acertar as vitórias dos times mandantes, revelou um equilíbrio para as demais classes. Essa capacidade de equilibrar as previsões entre diferentes resultados indica uma melhoria na adaptabilidade do modelo ao longo do tempo.

Com apenas 134 jogos de amostra, foi possível obter desempenho comparável aos outros experimentos, o que indica que a engenharia de *features* foi bem-sucedida. Ao comparar os três métodos, o cenário incremental apresentou os melhores resultados nas técnicas baseadas em *XGBoost* e RF, enquanto o baseado KNN destacou-se com o conjunto de dados menor. A superioridade do cenário incremental no RF pode ser atribuída à atualização contínua do conjunto de treinamento, permitindo que o modelo se ajuste melhor às nuances e mudanças ao longo do tempo, resultando em aprimoramentos graduais no desempenho. No caso do XGB, sua robustez em lidar com dados desbalanceados, comuns no cenário incremental a cada nova partida inserida, contribuiu para obter o melhor resultado. Por outro lado, o desempenho do KNN foi afetado negativamente, uma vez que a distância entre instâncias pode aumentar a cada iteração, prejudicando sua eficácia em um contexto dinâmico.

Os resultados destacam os desafios inerentes à tarefa de prever resultados de jogos de futebol. A acurácia média próxima a 60% ilustra a imprevisibilidade característica desses eventos esportivos e está de acordo com resultados obtidos por outros trabalhos (Schneider, 2020), (Leitner; Zeileis; Hornik, 2020).

A interpretação combinada da acurácia, *F1-score* evidenciam uma consistência nos desempenhos dos classificadores.

5 Considerações Finais

Neste capítulo, serão feitas as considerações finais sobre o trabalho, além de abordar o que pode ser explorado em trabalhos futuros.

5.1 Conclusão

O presente estudo abordou a predição de resultados de jogos do Campeonato Brasileiro Feminino de futebol, por meio de uma abordagem fundamentada em Aprendizado de Máquina. Foram inferidos classificadores e utilizados em dois experimentos distintos: um estático e outro incremental. Na abordagem estática, os modelos foram treinados e avaliados considerando conjuntos de dados fixos, enquanto no experimento incremental, o estudo buscou capturar e incorporar dinamicamente as mudanças ao longo do tempo. Essas abordagens permitiram avaliar a estabilidade do modelo e sua capacidade de se adaptar a novos padrões e informações que surgiram ao longo do campeonato.

Os resultados revelaram, em termos gerais, uma similaridade nas métricas de acurácia e F1-score para a maioria dos algoritmos. No entanto, destaca-se a variação nos resultados entre os classificadores (*Random Forest*, KNN e *XGBoost*) e a *baseline* no cenário estático, indicando que a previsão de resultados em partidas de futebol feminino vai além da consideração apenas de taxas de ataque, defesa e vantagem de jogar em casa. Além disso, o KNN apresentou variações significativas no cenário incremental em comparação com os outros dois classificadores, evidenciando a dificuldade do KNN em lidar com o desbalanceamento contínuo do conjunto de treino.

O *XGBoost* destacou-se como o melhor classificador em ambos os cenários. No primeiro, considerando o conjunto de treino estático, a acurácia e o F1-Score foram de 61,09% e 58,77%, respectivamente. No segundo cenário, em que o conjunto de treino foi incrementado a cada iteração, a acurácia e o F1-Score atingiram 63,43% e 65,18%, respectivamente. A diferença entre os dois pode ser explicada pela eficácia do *XGBoost* em lidar com dados desbalanceados.

Apesar das variações, a consistência nos desempenhos dos classificadores, evidenciada pela interpretação combinada da acurácia e do *F1-Score*, sugere uma robustez nas abordagens adotadas, uma vez que alcançaram resultados equivalentes ou superiores à literatura analisada e à *baseline*. Isso pode indicar a capacidade dos modelos em capturar padrões, mesmo diante da complexidade intrínseca dos eventos esportivos.

Em síntese, os resultados apresentados indicam que, embora a predição de resultados de jogos de futebol seja desafiadora, abordagens baseadas em aprendizado de máquina têm potencial para fornecer conhecimentos valiosos. Por exemplo, a avaliação dos atributos das jogadoras e

levar em consideração o no cenário incremental, devido a capacidade de um time aprimorar sua performance ao longo do campeonato, ressaltando a importância de uma abordagem dinâmica e adaptativa na modelagem preditiva. A compreensão dos limites e pontos fortes de cada método é crucial para aprimorar ainda mais a precisão e utilidade desses modelos em contextos esportivos.

5.2 Trabalhos Futuros

Os resultados desta pesquisa abrem perspectivas para diversas oportunidades de estudos futuros, com o intuito de aprimorar e expandir o entendimento na predição de resultados em jogos de futebol feminino.

Uma abordagem promissora para pesquisas subsequentes envolve a integração de eventos específicos durante as partidas, tais como passes no campo de ataque e a avaliação da taxa de expectativa de gols. Além disso, considerar elementos extra-campo, como lesões de jogadoras, condições climáticas e mudanças de técnicos, pode agregar profundidade aos modelos analíticos. Dado que a predição de resultados por si só é algo já dominado pelas casas de apostas, a tendência é que a ciência de dados aplicada ao futebol se concentre cada vez mais em eventos específicos da partida que podem influenciar no resultado.

Além disso, uma linha de pesquisa interessante seria a avaliação de estratégias para combinar os resultados de diferentes classificadores, a fim de gerar previsões mais robustas e precisas.

Outro ponto relevante para investigações futuras consiste na busca por plataformas que forneçam conjuntos de dados mais completos e abrangentes. Este desafio emergiu como uma consideração crucial durante o desenvolvimento deste trabalho, e superá-lo pode representar um avanço significativo para a pesquisa nesse domínio.

Referências

BATISTA, G. E. d. A. P. *et al.* **Pré-processamento de dados em aprendizado de máquina supervisionado**. Tese (Doutorado) — Universidade de São Paulo, 2003.

BAYES, T. Naive bayes classifier. **Article Sources and Contributors**, p. 1–9, 1968.

BREIMAN, L. Random forests. **Machine learning**, Springer, v. 45, n. 1, p. 5–32, 2001.

BREIMAN, L. *et al.* **Classification and Regression Trees**. 1st. ed. New York: Routledge, 1984. 368 p. Disponível em: <https://doi.org/10.1201/9781315139470>.

BURMAN, P. A comparative study of ordinary cross-validation, v-fold cross-validation and the repeated learning-testing methods. **Biometrika**, Oxford University Press, v. 76, n. 3, p. 503–514, 1989.

CHEN, T.; GUESTRIN, C. Xgboost: A scalable tree boosting system. In: **Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining**. New York, NY, USA: Association for Computing Machinery, 2016. (KDD '16), p. 785–794. ISBN 9781450342322. Disponível em: <https://doi.org/10.1145/2939672.2939785>.

COVER, T.; HART, P. Nearest neighbor pattern classification. **IEEE transactions on information theory**, IEEE, v. 13, n. 1, p. 21–27, 1967.

COX, D. R. Regression models and life-tables. **Journal of the Royal Statistical Society. Series B (Methodological)**, [Royal Statistical Society, Wiley], v. 34, n. 2, p. 187–220, 1972. ISSN 00359246. Disponível em: <http://www.jstor.org/stable/2985181>.

DETONI, H. O. Futebol feminino brasileiro e as dificuldades encontradas nesse subcampo esportivo. Universidade Interamericana do Brasil, 2022. Disponível em: <https://repositorio.uninter.com/handle/1/1081>.

DIETTERICH, T. G. Ensemble methods in machine learning. **Multiple Classifier Systems**, Springer, v. 1857, p. 1–15, 2000.

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

DUARTE, R.; COPPINI, J. Utilizando abordagens de aprendizado de máquina para prever resultados de jogos: o caso da liga nacional de futsal. **Revista Brasileira de Futsal e Futebol**, v. 13, n. 53, p. 275–283, 2021.

ESTADAO. **Futebol feminino tem crescimento de mais de 130% em apostas esportivas em um ano**. 2022. Estadão website. Acesso em 15 de agosto de 2023. Disponível em <https://www.estadao.com.br/esportes/futebol/apostas-esportivas-em-competicoes-de-futebol-feminino-crescem-mais-de-130-em-um-ano/>.

EXAME. **Interesse pelo Futebol Feminino no Brasil cresceu 34% nos últimos cinco anos**. 2024. Acesso em 24 de Janeiro de 2024. Disponível em: <https://exame.com/esporte/interesse-pelo-futebol-feminino-no-brasil-cresceu-34-nos-ultimos-cinco-anos/>.

FACELI, K. *et al.* Inteligência artificial: uma abordagem de aprendizado de máquina. 2021.

FISHER, R. A. The use of multiple measurements in taxonomic problems. **Annals of eugenics**, Wiley Online Library, v. 7, n. 2, p. 179–188, 1936.

FRIEDMAN, J. H. Stochastic gradient boosting. **Computational statistics & data analysis**, Elsevier, v. 38, n. 4, p. 367–378, 2002.

GANHOR, J. P. *et al.* Predição de resultados para partidas de futebol: um olhar para a produção científica em periódicos nacionais. **Revista Brasileira de Futsal e Futebol**, v. 18, n. 72, p. 58–67, 2020. ISSN 1984-4956. Disponível em: <http://www.rbff.com.br>.

GEURTS, P.; ERNST, D.; WEHENKEL, L. Extremely randomized trees. **Machine learning**, Springer, v. 63, p. 3–42, 2006.

HAYKIN, S. **Neural networks and learning machines, 3/E**. [S.l.]: Pearson Education India, 2009.

KARLIS, D.; NTZOUFRAS, I. Analysis of sports data by using bivariate poisson models. **Journal of the Royal Statistical Society: Series D (The Statistician)**, Wiley Online Library, v. 52, n. 3, p. 381–393, 2003.

KOTSIANTIS, S. B. *et al.* Supervised machine learning: A review of classification techniques. **Emerging artificial intelligence applications in computer engineering**, Amsterdam, v. 160, n. 1, p. 3–24, 2007.

LEITNER, C.; ZEILEIS, A.; HORNIK, K. A hybrid random forest approach for predicting the results of the fifa women's world cup 2019. **Journal of Applied Statistics**, Taylor & Francis, v. 47, n. 9, p. 1875–1893, 2020.

MACKENZIE, R.; ESPORTE, E. F. Ciência e tecnologia aplicada à melhoria do desempenho esportivo. **Revista Mackenzie de Educação Física e Esporte**, v. 11, n. 1, p. 143–157, 2012.

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

MATOS, P. F. *et al.* Relatório técnico "métricas de avaliaçao". **Universidade Federal de Sao Carlos**, 2009.

MIKHAIL, E. M.; ACKERMANN, F. E. Observations and least squares. University Press of America, 1976.

NILSSON, N. J. Introduction to machine learning, 1997. **Standford University**, 1997.

NUNES, M. **O que mudou no futebol feminino do Brasil desde a última Copa?** 2023. Dibradoras website. Acesso em 17 de julho de 2023. Disponível em <https://dibradoras.com.br/2023/03/22/o-que-mudou-no-futebol-feminino-do-brasil-desde-a-ultima-copa/>.

PECONICK, L. D. F. Inteligência artificial aplicada à previsão de jogos de futebol. **Monografia de Graduação**, Universidade de Brasília, 2018. Disponível em: <http://repositorio.unb.br/handle/10482/33708>.

PEDREGOSA, F. *et al.* Scikit-learn: Machine learning in python. **the Journal of machine Learning research**, JMLR. org, v. 12, p. 2825–2830, 2011.

QUINLAN, J. R. Induction of decision trees. **Machine learning**, Springer, v. 1, p. 81–106, 1986.

ROSA, P. H. C. Mineração de dados aplicada a previsão de resultados de jogos de futebol. Pontifícia Universidade Católica de Goiás, 2022.

RUMELHART, D. David e. rumelhart, geoffrey e. hinton, and ronald j. williams learning representations by back-propagating errors nature 323: 533-536. **nature**, v. 323, p. 533–536, 1986.

SAITO, T.; REHMSMEIER, M. The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets. **PloS one**, Public Library of Science San Francisco, CA USA, v. 10, n. 3, p. e0118432, 2015.

SCHNEIDER, C. F. Aplicação de técnicas de machine learning para previsão de resultados de partidas de futebol. Universidade Federal do Rio Grande do Sul, 2020. Projeto de diplomação em Engenharia Elétrica. Disponível em: <http://hdl.handle.net/10183/209064>.

SILVA, A. B. L.; BARROS, K. N. N. de O.; ALBUQUERQUE, M. A. Modelagem via árvore de decisão para previsão de jogos de futebol. **Research, Society and Development**, v. 9, n. 9, p. e204996869–e204996869, 2020.

SILVA, B. M. d. Modelo preditivo aplicado ao futebol brasileiro. **RBFF - Revista Brasileira de Futsal e Futebol**, v. 14, n. 58, p. 291–297, nov. 2022. Disponível em: <http://www.rbff.com.br/index.php/rbff/article/view/1265>.

SILVA, B. M. da. Multiple linear regression applied to football/ regressao linear multipla aplicada ao futebol. **Revista Brasileira de Futsal e Futebol**, Instituto Brasileiro de Pesquisa e Ensino em Fisiologia do Exercicio. IBPEFEX, v. 10, n. 38, p. 262, 2018. ISSN 1984-4956.

TURING, A. M. Computing machinery and intelligence. **Mind**, Oxford University Press, LIX, n. 236, p. 433–460, 1950.

VAPNIK, V. **The nature of statistical learning theory**. [S.l.]: Springer science & business media, 1995.

ZAKI, M. J.; MEIRA, W. **Data mining and analysis: fundamental concepts and algorithms**. [S.l.]: Cambridge University Press, 2020.

Apêndices

APÊNDICE A – Atributos utilizados na predição

Tabela A.1 – Descrição dos Dados

Coluna	Descrição
aerial_duels_won_per_game_overall_home_forward	Número médio de duelos aéreos ganhos por jogo para jogadores de ataque da equipe da casa.
red_cards_overall_home_forward	Número total de cartões vermelhos recebidos por jogadores de ataque da equipe da casa.
average_rating_overall_home_forward	Avaliação média de desempenho geral para jogadores de ataque da equipe da casa.
min_per_goal_overall_away_midfielder	Média de minutos por gol marcado para meio-campistas da equipe visitante.
red_cards_overall_away_midfielder	Número total de cartões vermelhos recebidos por meio-campistas da equipe visitante.
yellow_cards_overall_away_midfielder	Número total de cartões amarelos recebidos por meio-campistas da equipe visitante.
conceded_overall_away_midfielder	Número total de gols sofridos pela equipe visitante enquanto os meio-campistas estão em campo.
clean_sheets_overall_away_midfielder	Número total de partidas sem sofrer gols pela equipe visitante enquanto os meio-campistas estão em campo.
assists_overall_away_midfielder	Número total de assistências fornecidas por meio-campistas da equipe visitante.
games_started_home_midfielder	Número de jogos iniciados por meio-campistas da equipe da casa.

games_subbed_in_home_midfielder	Número de jogos em que meio-campistas da equipe da casa foram substituídos.
games_subbed_out_home_midfielder	Número de jogos em que meio-campistas da equipe da casa foram substituídos.
shots_total_overall_home_midfielder	Número total de chutes a gol de meio-campistas da equipe da casa.
aerial_duels_won_per_game_overall_home_midfielder	Número médio de duelos aéreos ganhos por jogo para meio-campistas da equipe da casa.
xg_per_game_overall_home_midfielder	Média de expectativa de gols (xG) por jogo para meio-campistas da equipe da casa.
passes_per_game_overall_home_midfielder	Média de passes por jogo para meio-campistas da equipe da casa.
assists_per_game_overall_home_midfielder	Média de assistências por jogo para meio-campistas da equipe da casa.
average_rating_overall_home_midfielder	Avaliação média de desempenho geral para meio-campistas da equipe da casa.
average_rating_overall_away_midfielder	Avaliação média de desempenho geral para meio-campistas da equipe visitante.
xg_per_game_overall_home_forward	Média de expectativa de gols (xG) por jogo para jogadores de ataque da equipe da casa.
assists_per_game_overall_away_midfielder	Média de assistências por jogo para meio-campistas da equipe visitante.
yellow_cards_overall_home_midfielder	Número total de cartões amarelos recebidos por meio-campistas da equipe da casa.
min_per_goal_overall_home_forward	Média de minutos por gol marcado para jogadores de ataque da equipe da casa.
min_per_goal_overall_home_midfielder	Média de minutos por gol marcado para meio-campistas da equipe da casa.
yellow_cards_overall_home_forward	Número total de cartões amarelos recebidos por jogadores de ataque da equipe da casa.

conceded_overall_home_forward	Número total de gols sofridos pela equipe da casa enquanto os jogadores de ataque estão em campo.
clean_sheets_overall_home_forward	Número total de partidas sem sofrer gols pela equipe da casa enquanto os jogadores de ataque estão em campo.
assists_overall_home_forward	Número total de assistências fornecidas por jogadores de ataque da equipe da casa.
assists_overall_home_midfielder	Número total de assistências fornecidas por meio-campistas da equipe da casa.
clean_sheets_overall_home_midfielder	Número total de partidas sem sofrer gols pela equipe da casa enquanto os meio-campistas estão em campo.
games_started_away_midfielder	Número de jogos iniciados por meio-campistas da equipe visitante.
games_subbed_in_away_midfielder	Número de jogos em que meio-campistas da equipe visitante foram substituídos.
games_subbed_out_away_midfielder	Número de jogos em que meio-campistas da equipe visitante foram substituídos.
shots_total_overall_away_midfielder	Número total de chutes a gol de meio-campistas da equipe visitante.
aerial_duels_won_per_game_overall_away_midfielder	Número médio de duelos aéreos ganhos por jogo para meio-campistas da equipe visitante.
xg_per_game_overall_away_midfielder	Média de expectativa de gols (xG) por jogo para meio-campistas da equipe visitante.
conceded_overall_home_midfielder	Número total de gols sofridos pela equipe da casa enquanto os meio-campistas estão em campo.
passes_per_game_overall_away_midfielder	Média de passes por jogo para meio-campistas da equipe visitante.
min_per_conceded_overall_away_defender	Média de minutos por gol concedido para defensores da equipe visitante.

red_cards_overall_home_midfielder	Número total de cartões vermelhos recebidos por meio-campistas da equipe da casa.
min_per_goal_overall	Média de minutos por gol marcado para todos os jogadores.
xg_per_game_overall	Média de expectativa de gols (xG) por jogo para todos os jogadores.
minutes_played_overall_away_goalkeeper	Número total de minutos jogados por goleiros da equipe visitante.
passes_per_game_overall	Média de passes por jogo para todos os jogadores.
assists_per_game_overall	Média de assistências por jogo para todos os jogadores.
age_away_goalkeeper	Idade dos goleiros da equipe visitante.
clean_sheets_overall_away_defender	Número total de partidas sem sofrer gols pela defesa da equipe visitante.
conceded_overall_away_defender	Número total de gols sofridos pela defesa da equipe visitante.
clean_sheets_overall_away_goalkeeper	Número total de partidas sem sofrer gols pelos goleiros da equipe visitante.
conceded_overall_away_goalkeeper	Número total de gols sofridos pelos goleiros da equipe visitante.
min_per_goal_overall_away_defender	Média de minutos por gol marcado para defensores da equipe visitante.
yellow_cards_overall_away_goalkeeper	Número total de cartões amarelos recebidos pelos goleiros da equipe visitante.
yellow_cards_overall_away_defender	Número total de cartões amarelos recebidos pelos defensores da equipe visitante.
red_cards_overall_away_defender	Número total de cartões vermelhos recebidos pelos defensores da equipe visitante.
clean_sheets_overall_home_defender	Número total de partidas sem sofrer gols pela defesa da equipe da casa.
yellow_cards_overall_home_defender	Número total de cartões amarelos recebidos pelos defensores da equipe da casa.

conceded_overall_home_defender	Número total de gols sofridos pela defesa da equipe da casa.
leading_at_half_time_percentage_home	Percentual de vezes que a equipe da casa estava liderando no intervalo.
clean_sheet_half_time_percentage	Percentual de vezes que a equipe da casa não sofreu gols no primeiro tempo.
minutes_per_goal_conceded_away	Média de minutos por gol sofrido pela equipe visitante.
wins_away	Número de vitórias da equipe visitante.
goals_scored_half_time_home	Número de gols marcados pela equipe da casa no primeiro tempo.
leading_at_half_time_percentage_away	Percentual de vezes que a equipe visitante estava liderando no intervalo.
minutes_per_goal_scored_home	Média de minutos por gol marcado pela equipe da casa.
shots_home	Número total de chutes a gol da equipe da casa.
goals_conceded_per_match_half_time_home	Média de gols sofridos pela equipe da casa no primeiro tempo por partida.
under45_percentage_away	Percentual de jogos em que a equipe visitante marcou menos de 4,5 gols.
over15_percentage_home	Percentual de jogos em que a equipe da casa marcou mais de 1,5 gols.
goals_scored_per_match_home	Média de gols marcados pela equipe da casa por partida.
interceptions_total_overall_home_forward	Número total de interceptações feitas pelos atacantes da equipe da casa.
goals_scored_per_match_half_time_home	Média de gols marcados pela equipe da casa no primeiro tempo por partida.
over25_count_half_time_home	Número de vezes que a equipe da casa marcou mais de 2,5 gols no primeiro tempo.
fts_half_time_home	Número de vezes que a equipe da casa não marcou no primeiro tempo.
draw_percentage_home	Percentual de empates da equipe da casa.
interceptions_total_overall_away_forward	Número total de interceptações feitas pelos atacantes da equipe visitante.

over15_count_half_time_away	Número de vezes que a equipe visitante marcou mais de 1,5 gols no primeiro tempo.
corners_per_match_home	Média de escanteios por partida da equipe da casa.
league_position_home	Posição da equipe da casa na liga.
over15_percentage_away	Percentual de jogos em que a equipe visitante marcou mais de 1,5 gols.
league_position_away	Posição da equipe visitante na liga.
shots_off_target_home	Número total de chutes para fora do alvo da equipe da casa.
draw_percentage_away	Percentual de empates da equipe visitante.
goals_scored_half_time_away	Número de gols marcados pela equipe visitante no primeiro tempo.
pen_committed_total_overall_home_midfielder	Número total de penalidades cometidas pelos meio-campistas da equipe da casa.
under45_percentage_home	Percentual de jogos em que a equipe mandante marcou menos de 4,5 gols.
corners_total_home	Número total de escanteios da equipe da casa.
loss_percentage_away	Percentual de derrotas da equipe visitante.
wins_home	Número de vitórias da equipe da casa.
under25_count_home	Número de jogos em que a equipe mandante marcou menos de 2,5 gols.
corners_total_away	Número total de escanteios da equipe visitante.
over55_count_away	Número de jogos em que foram marcados mais de 5,5 gols pela equipe visitante.
over05_percentage_home	Percentual de jogos em que a equipe mandante marcou pelo menos 1 gol.
goals_conceded_half_time_home	Número de gols sofridos pela equipe da casa no primeiro tempo.
cards_total_away	Número total de cartões recebidos pela equipe visitante.

btts_half_time_percentage_home	Percentual de jogos em que ambas as equipes marcaram no primeiro tempo na equipe da casa.
over25_half_time_percentage	Percentual de jogos em que foram marcados mais de 2,5 gols no primeiro tempo.
pen_save_percentage_overall_away_defender	Percentual de penalidades salvas pelos defensores da equipe visitante.
leading_at_half_time_away	Número de vezes que a equipe visitante estava liderando no intervalo.
goals_conceded_per_match_half_time_away	Média de gols sofridos pela equipe visitante no primeiro tempo por partida.
clean_sheet_half_time_percentage_home	Percentual de vezes que a equipe da casa não sofreu gols no primeiro tempo.
under55_percentage_home	Percentual de jogos em que a equipe da casa marcou menos de 5,5 gols.
draws_away	Número de empates da equipe visitante.
first_team_to_score_percentage_home	Percentual de vezes que a equipe da casa foi a primeira a marcar.
fts_percentage_away	Percentual de jogos em que a equipe visitante não marcou gols.
over25_percentage_away	Percentual de jogos em que a equipe visitante marcou mais de 2,5 gols.
under15_count_away	Número de jogos em que a equipe visitante marcou menos de 1,5 gols.
under45_count_home	Número de jogos em que a equipe da casa marcou menos de 4,5 gols.
under35_percentage_away	Percentual de jogos em que a equipe visitante marcou menos de 3,5 gols.
fts_half_time_away	Número de vezes que a equipe visitante não marcou no primeiro tempo.
over05_count_home	Número de jogos em que a equipe mandante marcou pelo menos 1 gol.
over45_count_away	Número de jogos em que a equipe visitante marcou mais de 4,5 gols.

over15_count_half_time_home	Número de vezes que a equipe da casa marcou mais de 1,5 gols no primeiro tempo.
goals_scored_per_match_away	Média de gols marcados pela equipe visitante por partida.
clean_sheet_half_time_away	Número de vezes que a equipe visitante não sofreu gols no primeiro tempo.
cards_total_home	Número total de cartões recebidos pela equipe da casa.
under35_percentage_home	Percentual de jogos em que a equipe da casa marcou menos de 3,5 gols.
shots_per_goal_scored_overall_away_midfielder	Média de chutes por gol marcado pelos meio-campistas da equipe visitante.
xg_for_avg_away	Média de expectativa de gols (xG) para a equipe visitante.
first_team_to_score_percentage_away	Percentual de vezes que a equipe visitante foi a primeira a marcar.
over15_count_home	Número de vezes que a equipe da casa marcou mais de 1,5 gols.
over45_count_home	Número de jogos em que a equipe da casa marcou mais de 4,5 gols.
under55_percentage_away	Percentual de jogos em que a equipe visitante marcou menos de 5,5 gols.
over05_count_away	Número de jogos em que a equipe visitante marcou pelo menos 1 gol.
draw_at_half_time_home	Número de vezes que a equipe da casa empatou no intervalo.
age_home_forward	Idade dos atacantes da equipe da casa.
hit_woodwork_total_overall_home_forward	Número total de vezes que os atacantes da equipe da casa acertaram a trave.
under25_count_away	Número de jogos em que a equipe visitante marcou menos de 2,5 gols.
shots_off_target_away	Número total de chutes para fora do alvo da equipe visitante.
over25_percentage_home	Percentual de jogos em que a equipe da casa marcou mais de 2,5 gols.
minutes_per_goal_conceded_home	Média de minutos por gol sofrido pela equipe da casa.

goals_overall_away_forward	Número total de gols marcados pelos atacantes da equipe visitante.
xg_for_avg_home	Média de expectativa de gols (xG) para a equipe da casa.
over_25_percentage_pre_match	Percentual de jogos em que foram marcados mais de 2,5 gols antes da partida.
over05_percentage_away	Percentual de jogos em que a equipe visitante marcou pelo menos 1 gol.
xg_against_avg_home	Média de expectativa de gols (xG) contra a equipe da casa.
over65_corners_percentage	Percentual de jogos em que a equipe da casa teve mais de 6,5 escanteios.
under05_percentage	Percentual de jogos em que foram marcados menos de 0,5 gols.
average_total_goals_per_match	Média total de gols por partida.
interceptions_total_overall_away_midfielder	Número total de interceptações feitas pelos meio-campistas da equipe visitante.
over05_half_time_percentage_home	Percentual de jogos em que a equipe da casa marcou pelo menos 1 gol no primeiro tempo.
home_team_name	Nome da equipe da casa.
age_home_goalkeeper	Idade dos goleiros da equipe da casa.
hit_woodwork_total_overall_home_goalkeeper	Número total de vezes que os goleiros da equipe da casa acertaram a trave.
over_45_percentage_pre_match_away	Percentual de jogos em que foram marcados mais de 4,5 gols antes da partida pela equipe visitante.
minutes_played_overall_home_forward	Número total de minutos jogados por jogadores de ataque da equipe da casa.
minutes_played_overall_away_forward	Número total de minutos jogados por jogadores de ataque da equipe visitante.
shots_total_overall_home_forward	Número total de chutes a gol de jogadores de ataque da equipe da casa.
assists_per_game_overall_home_forward	Média de assistências por jogo para jogadores de ataque da equipe da casa.

average_rating_overall_away_forward	Avaliação média de desempenho geral para jogadores de ataque da equipe visitante.
xg_per_game_overall_away_forward	Média de expectativa de gols (xG) por jogo para jogadores de ataque da equipe visitante.
passes_per_game_overall_away_forward	Média de passes por jogo para jogadores de ataque da equipe visitante.
assists_per_game_overall_away_forward	Média de assistências por jogo para jogadores de ataque da equipe visitante.
yellow_cards_overall_away_forward	Número total de cartões amarelos recebidos por jogadores de ataque da equipe visitante.
conceded_overall_away_forward	Número total de gols sofridos pela equipe visitante enquanto os jogadores de ataque estão em campo.
clean_sheets_overall_away_forward	Número total de partidas sem sofrer gols pela equipe visitante enquanto os jogadores de ataque estão em campo.
games_started_home_forward	Número de jogos iniciados por jogadores de ataque da equipe da casa.
games_subbed_in_home_forward	Número de jogos em que jogadores de ataque da equipe da casa foram substituídos.
games_subbed_out_home_forward	Número de jogos em que jogadores de ataque da equipe da casa foram substituídos.
min_per_conceded_overall_home_defender	Média de minutos por gol concedido para defensores da equipe da casa.
minutes_played_overall_home_goalkeeper	Número total de minutos jogados por goleiros da equipe visitante.
clean_sheets_overall_home_goalkeeper	Número total de partidas sem sofrer gols pelos goleiros da equipe visitante.
conceded_overall_home_goalkeeper	Número total de gols sofridos pelos goleiros da equipe visitante.
min_per_goal_overall_home_defender	Média de minutos por gol marcado para defensores da equipe visitante.

yellow_cards_overall_home_goalkeeper	Número total de cartões amarelos recebidos pelos goleiros da equipe visitante.
red_cards_overall_home_defender	Número total de cartões vermelhos recebidos pelos defensores da equipe visitante.
fts_percentage_away	Percentual de jogos em que a equipe visitante não marcou gols.
goals_conceded_half_time_away	Número de gols sofridos pela equipe visitante no primeiro tempo.
shots_on_target_per_match_home	Média de chutes no alvo por partida da equipe da casa.
goals_conceded_half_time_percentage_home	Percentual de gols sofridos pela equipe da casa no primeiro tempo.
goals_conceded_per_match_away	Média de gols sofridos pela equipe visitante por partida.
goals_scored_half_time_percentage_home	Percentual de gols marcados pela equipe da casa no primeiro tempo.
shots_on_target_per_match_away	Média de chutes no alvo por partida da equipe visitante.
corners_per_match_half_time_away	Média de escanteios por partida no primeiro tempo para a equipe visitante.
fts_percentage_home	Percentual de jogos em que a equipe da casa não marcou gols.
shots_off_target_per_match_away	Média de chutes para fora do alvo por partida da equipe visitante.
over15_count_away	Número de jogos em que a equipe visitante marcou mais de 1,5 gols.
btts_half_time_percentage_away	Percentual de jogos em que ambas as equipes marcaram no primeiro tempo na equipe visitante.
shots_off_target_per_match_home	Média de chutes para fora do alvo por partida da equipe da casa.
over15_half_time_percentage	Percentual de jogos em que foram marcados mais de 1,5 gols no primeiro tempo.
goals_conceded_per_match_home	Média de gols sofridos pela equipe da casa por partida.

clean_sheet_half_time_percentage_away	Percentual de vezes que a equipe visitante não sofreu gols no primeiro tempo.
corners_per_match_half_time_home	Média de escanteios por partida no primeiro tempo para a equipe da casa.
penalties_won_total_home	Número total de pênaltis vencidos pela equipe da casa.
over35_count_away	Número de jogos em que a equipe visitante marcou mais de 3,5 gols.
btts_percentage_away	Percentual de jogos em que ambas as equipes marcaram na equipe visitante.
over45_percentage_away	Percentual de jogos em que a equipe visitante marcou mais de 4,5 gols.
minutes_per_goal_scored_away_forward	Média de minutos por gol marcado por atacantes da equipe visitante.
over35_count_home	Número de jogos em que a equipe da casa marcou mais de 3,5 gols.