Aplicação de Machine Learning para Predição de Partidas Baseada em Dados do Campeonato Brasileiro de Futebol

1st Enzo Reis de Oliveira
Instituto de Ciência e Tecnologia
Universidade Federal de São Paulo
São José dos Campos, Brasil
enzo.oliveira@unifesp.br

2nd Renan Augusto de Souza Leite
Instituto de Ciência e Tecnologia
Universidade Federal de São Paulo
São José dos Campos, Brasil
rleite@unifesp.br

ABSTRACT

Este trabalho visa desenvolver um modelo de previsão para o Campeonato Brasileiro de Futebol, explorando o crescimento significativo do mercado de apostas esportivas no Brasil. Com o futebol sendo o esporte mais popular para apostas, representando 93% do total, há uma oportunidade para criar sistemas de previsão mais sofisticados. O objetivo é analisar parâmetros específicos de cada partida e gerar uma previsão do vencedor, perdedor ou empate. O modelo será treinado e ajustado usando dados de campeonatos passados, de 2015 até 2023. Utilizando uma combinação de várias técnicas, incluindo um algoritmo especializado de aprendizado de máquina e o uso de uma rede MLP (Multilayer Perceptron) como o modelo de aprendizagem. A avaliação dos modelos será baseada na acurácia, utilizando uma matriz de confusão. O melhor modelo será utilizado para prever resultados de partidas futuras no Brasileirão de 2023. Os desafios incluem a natureza imprevisível do futebol, o desbalanceamento das classes de resultados e a coleta manual de dados de entrada para cada partida.

I. Introdução

O mercado de apostas esportivas cresceu significativamente no Brasil nos últimos anos. A prática, que sempre teve muita popularidade, foi reinventada com o avanço da tecnologia. Agora, os torcedores podem fazer suas previsões através de sites e aplicativos, sem sair de casa. O funcionamento é simples: as empresas apresentam aos apostadores quais as probabilidades de concretização de cada fato e qual valor será pago nas apostas de sucesso. Depois, basta acompanhar as modalidades e contar com a sorte.

Segundo o "UOL", o futebol está no topo das apostas esportivas, respondendo por 93% do número total de apostas. Em seguida vêm o basquete (31%) e os eSports (29%). Mais da metade dos apostadores de futebol no país afirmaram que costumam apostar em jogos do Campeonato Brasileiro (56%). A Copa Libertadores e Copa do Brasil também se destacam, com 44% e 41% das apostas, respectivamente.

Segundo o "Estadão", desde 2018, o valor destinado às apostas esportivas no país passou a girar em torno de 120-150 bilhões de reais por ano. É interessante relacionar os altos números vinculados ao futebol com a cultura brasileira, que o torna o esporte mais popular e com o maior número de torcedores no país.

O crescimento do mercado de apostas esportivas no Brasil, especialmente no futebol, como ilustrado pelas estatísticas do "UOL" e "Estadão", inspira a necessidade de desenvolver sistemas de previsão mais sofisticados. O papel predominante do futebol na cultura brasileira torna o desenvolvimento de um simulador de campeonato para seleções brasileiras, que tentará prever próximas partidas no Campeonato Brasileiro de Futebol, vulgo Brasileirão, particularmente relevante.

Para atingir esse objetivo, foram adotadas diversas técnicas de aprendizado de máquina, desde a preparação dos dados, incluindo a incorporação de rankings da CBF, valores de mercado das equipes e desempenho recente, até o treinamento e ajuste de um modelo MLP para realizar as previsões (Schumaker et al., 2010).

II. Trabalhos relacionados

A aplicação de machine learning no contexto de predição de partidas em esportes é muito específica, já que existem diversos autores com diferentes ideias de como aplicar aprendizado de máquina para diferentes problemas. Isso torna a pesquisa em relação ao assunto desafiadora, uma vez que torna-se difícil encontrar, em meio a tantas referências, uma boa contribuição Por outro lado, essa dimensão de trabalhos também é inspiradora, dado que é possível aplicar diversos métodos diferentes para encontrar aquele que melhor se ajusta aos nossos ideais. Assim, as referências não se limitam a poucos modelos, auxiliando na acurácia desse trabalho.

Dito isso, é válido mencionar quais foram os trabalhos externos analisados e quais foram as suas contribuições a esse projeto. O artigo "The use of machine learning in sport outcome prediction: A review" fala sobre aplicações de machine learning em esportes como um todo. Nele, a importância da manipulação de dados é ratificada a partir da análise de

mais de 100 artigos sobre aplicação de machine learning em esportes em geral. A conclusão é que, na maioria deles, alguma forma de feature selection e feature extraction é utilizada antes do algoritmo de machine learning entrar. Essas técnicas serão implementadas com o intuito de encontrar os melhores atributos para nosso conjunto de dados.

Durante a realização desse projeto, houve um dilema a respeito da melhor forma para fazer as predições de cada partida: um algoritmo que chegará ao resultado da partida pelos gols, como visto no trabalho de Gianluca Baio e Marta Blangiardo "Bayesian hierarchical model for the prediction of football results" e no trabalho de João Marcos Amorim dos Santos "Previsões de Resultados em Partidas do Campeonato Brasileiro de Futebol" (neste trabalho mostra diferentes modelos que podem prever a quantidade de gols, juntamente com benefícios e malefícios de cada modelo) ou a partir dos pontos por vitória, derrota e empate, como visto no trabalho de C. Liti, V. Piccialli e M. Sciandrone "Predicting soccer match outcome using machine learning algorithm".

Após uma análise extensa de ambos artigos, foi determinado que seria melhor tratar esse problema através da classificação, uma vez que os resultados tendem a ser mais próximos da realidade. Além disso, através dessa solução, haverá menos variáveis para a consideradas quando comparadas com o resultado por gols. Outro artigo que irá nortear bastante esse projeto foi o trabalho desenvolvido por Werner Dubitzky, Philippe Lopes, Jesse Davis e Daniel Berrar conhecido como "The Open International Soccer Database for machine learning". Nele, os pesquisadores mostraram as grandes dificuldades de predizer resultados de futebol. Foram estudados diversos artigos, mostrando como diferentes autores tiveram diferentes abordagens para a predição de partidas de futebol. Além disso, foi demonstrado um extenso banco de dados e possíveis ideias para sua manipulação. Pensando nessa solução, foi decidida uma nova abordagem ao projeto: trazer um banco de dados que analisará partida por partida, de cada campeonato (em nosso caso ao invés de analisar campeonatos diferentes, será analisado temporadas diferentes do mesmo campeonato).

Tendo conhecimento que o futebol, como todo esporte, é imprevisível, sabe-se que os resultados podem contrariar totalmente expectativas, então vale ressaltar que é extremamente complicado determinar fatores serão decisivos para um algoritmo predizer. Porém, existe algo que contribui bastante para que uma equipe tenha uma maior probabilidade de vitória em relação a outra: o elenco. Um time com um elenco superior está mais propício a vencer uma partida e por isso, esse é um atributo que não pode faltar. Entretanto, é extremamente complicado definir a disparidade entre elenco, pois isso pode ser relativo. O artigo feito por Anthony Costa Constantinou e Norman Elliott Fenton, intitulado de "Determining the level of ability of football teams by dynamic ratings based on the relative discrepancies in scores between adversaries" e o trabalho de estudantes da UFRJ chamado de "Futebol Consciente" criaram uma solução interessante para essa problemática. Eles calculam/classificam a força de equipes através de duas maneiras. O primeiro utiliza uma ideia mais geral, que pode ser aplicado em diferentes times de diferentes campeonatos. Já o segundo é mais específico para o futebol brasileiro, especificamente o campeonato Brasileiro (campeonato base do estudo).

Outra literatura que contribui muito para esse projeto foi um TCC feito por João Henrique Martins Lima, com o título de "APLICAÇÃO DE MACHINE LEARNING PARA APOSTAS ESPORTIVAS: uso de Regressão Logística, SVM, Árvore de Decisão e Naive Bayes". Nesse TCC, o autor, além de fornecer um data frame com as ODDS de diversas partidas de diversas temporadas diferentes do Brasileirão, mostrou o comportamento de diferentes modelos diferentes quando aplicado no conjunto de dados. Ademais, foram demonstradas como as comparações entre modelos podem ser evidenciadas e como cada modelo pode ser implementado. Por isso, esse artigo se mostrou extremamente importante para a realização do nosso trabalho.

Por fim, outras duas literaturas que irão influenciar esse projeto é o de Ígor Barbosa de Costa em seu trabalho de "Modelagem e Predição de Resultados de Futebol Antes e Durante as Partidas Usando Aprendizagem de Máquina" que utiliza diversos modelos para realizar predições de partidas, com grande enfoque de predição de partidas durante o jogo. Esse trabalho irá contribuir para o desenvolvimento do projeto, pois dentre os modelos utilizados pelo autor está o de séries temporais. Acredita-se que séries temporais serão muito utilizadas nesse trabalho, então possuir uma literatura que utiliza-se desse modelo, ainda mais voltada com predições em partidas de futebol, irá facilitar o trabalho como um todo. E uma última literatura que irá contribuir no decorrer deste trabalho é um artigo feito por Edward Wheatcroft e Ewelina Sienkiewicz chamado "Calibration and hyper parameter tuning in football forecasting with Machine Learning" no artigo é muito falado sobre a importância dos hiperparâmetros para ajustar modelos para realizar predições mais precisas. As técnicas, dicas e problemas que podem surgir quando hiperparâmetros entrarem nos modelos, terá uma grande base nas informações escritas nesse artigo.

III. Conceitos fundamentais

Para melhor compreensão de 'jargões' que serão utilizados ao decorrer desse projeto, é bom defini-los adequadamente:

- "Machine Learning", ou aprendizado de máquina, é um ramo da inteligência artificial que utiliza algoritmos para permitir que os computadores reconheçam padrões em dados massivos e façam previsões (análise preditiva). Esse aprendizado permite que os computadores executem determinadas tarefas de forma autônoma, ou seja, sem programação.
- "Data Frame", ou quadro de dados, é uma estrutura tabular que organiza os dados em linhas e colunas. É uma maneira conveniente de armazenar dados estruturados, onde cada coluna representa uma variável e cada linha uma observação ou registro. No contexto deste trabalho, os "data frames"são usados para armazenar o histórico

de jogo de cada equipe e outras informações relevantes para análise e previsão.

- Modelo de campeonato por "pontos corridos": Cada equipe se enfrenta uma vez em casa (jogando em seu estádio) e uma vez fora de casa (jogando no estádio do time adversário). O resultado de cada partida é tradicionalmente pontuado com 3 pontos para a vitória, 1 ponto para o empate e 0 pontos para a derrota
- Elenco são os jogadores e a equipe técnica de cada time, incluindo o treinador e sua comissão técnica. Os jogadores são divididos em diferentes posições, cada um com sua responsabilidade específica dentro de campo São eles: goleiro, zagueiro, meio-campista e atacante, cada um com suas responsabilidades específicas dentro de campo. O técnico é responsável por definir a estratégia de jogo, escalar os jogadores e orientar a equipe.
- "Cartola FC" é um jogo virtual de futebol muito popular no Brasil, onde os participantes montam times fictícios com jogadores reais do campeonato brasileiro. O desempenho nas partidas reais de cada jogador gera pontos, que são calculados levando em consideração critérios como gols, assistências, defesas, entre outros. No contexto deste trabalho, o Cartola FC é utilizado como fonte de dados para análise das estatísticas individuais dos jogadores e para avaliar a força dos elencos das equipes.
- ODD (ou odds), são as probabilidades atribuídas a um determinado resultado em uma partida ou evento esportivo. Elas representam a expectativa de um determinado resultado ocorrer e são usadas no mercado de apostas esportivas para determinar os pagamentos. As odds são geralmente expressas como números decimais, fracionários ou como linhas de dinheiro. Por exemplo, uma odd de 2.50 significa que, se você apostar em um determinado resultado e ganhar, receberá duas vezes e meia o valor apostado.
- "Força do time"é uma métrica de performance de um time que leva em conta dados como gols feitos, gols sofridos, gols esperados e qualidade das chances criadas por um time. A partidr desses dados é gerado um rating ofensivo e defensivo e depois um geral, que classifica a força de um time
- Um jogador titular é aquele selecionado pelo treinador para iniciar uma partida. Esses jogadores são considerados os mais qualificados, em melhores condições físicas e técnicas, e são escolhidos para começar o jogo. Geralmente, os jogadores titulares são os mais confiáveis e têm um papel importante nas estratégias e táticas da equipe. Eles são escalados desde o início da partida e desempenham um papel fundamental na tentativa de alcançar o sucesso dentro de campo.
- Temporada, em futebol, é um termo que refere-se a um período específico de competição que geralmente segue o calendário anual. Em ligas de pontos corridos, como o Campeonato Brasileiro, a temporada normalmente abrange vários meses, durante os quais todas as equipes se enfrentam em jogos de ida e volta. A temporada é usada como uma unidade de tempo para analisar o desempenho das equipes ao longo do campeonato e para acompanhar os resultados e estatísticas dos jogos disputados nesse período.

Esse algoritmo é especializado em prever resultados do Campeonato Brasileiro de Futebol, com dados abrangendo os anos de 2015 a 2023.

O cerne do nosso estudo é a análise de dados que englobam informações sobre cada partida realizada no Campeonato Brasileiro durante o período indicado, combinadas com informações adicionais como os rankings das equipes fornecidos pela CBF e valores de mercado das equipes. Estes dados, quando processados, formam a base sobre a qual o nosso algoritmo de predição opera.

Adotamos uma estratégia de treinamento e teste, que é uma prática padrão em aprendizado de máquina. Esta estratégia envolve dividir nosso conjunto de dados em dois subconjuntos: 70% dos dados são usados para treinar nosso modelo, enquanto os 30% restantes são reservados para testar a precisão do modelo. As variáveis consideradas para a predição abrangem um amplo espectro, desde odds de apostas até o desempenho recente dos times, passando por classificações de cada equipe.

Utilizamos uma rede neural Multilayer Perceptron (MLP) para o nosso modelo de predição. Esta escolha de modelo se deve à capacidade das MLPs de lidar efetivamente com a alta dimensionalidade e complexidade inerentes aos nossos dados. A estrutura da nossa MLP envolve várias camadas totalmente conectadas, com um mecanismo de dropout incorporado para evitar o overfitting e melhorar a generalização do modelo.

Durante a fase de treinamento, os dados são divididos em batches, o que permite um treinamento mais eficiente e menos intensivo em termos de memória. Um aspecto interessante do nosso treinamento é que focamos na retropropagação para modelos com alta entropia, ou seja, modelos nos quais a rede neural expressa incerteza sobre a classificação. Isso nos ajuda a concentrar nosso esforço de treinamento onde é mais necessário. O modelo que demonstra a melhor acurácia durante a fase de teste é selecionado como nosso modelo final e é preservado para referência futura.

Após a fase de treinamento e seleção do modelo, realizamos uma série de análises para avaliar o desempenho do nosso modelo. Geramos diversas métricas de desempenho, incluindo Acurácia Balanceada, Precisão, Revocação e F1 Score, que fornecem uma visão holística do desempenho do nosso modelo.

Adicionalmente, também produzimos várias visualizações gráficas, como a evolução da função de perda ao longo das épocas de treinamento e um histograma das classes previstas, oferecendo uma visão clara da distribuição das previsões. Uma matriz de confusão é gerada para cada equipe individualmente, permitindo-nos examinar o desempenho do modelo para cada equipe de maneira mais detalhada.

Finalmente, para ilustrar a estabilidade do nosso modelo, geramos um gráfico que exibe a acurácia do modelo ao longo de várias janelas deslizantes. Este gráfico nos permite

visualizar a consistência do nosso modelo ao longo do tempo, evidenciando a eficácia do nosso algoritmo de predição.

IV. Objetivos

Este trabalho visa desenvolver um modelo de previsão para o Campeonato Brasileiro de Futebol, usando aprendizado de máquina para prever resultados de partidas que já ocorreram e que ainda estão por acontecer, de acordo com a metodologia descrita na seção de trabalhos relacionados (Kumar et al., 2020; Baio e Blangiardo, 2010; Santos, 2015; Liti et al., 2020; Constantinou e Fenton, 2012; Lima, 2022; Costa, 2022; Wheatcroft e Sienkiewicz, 2022).

Ou seja, nosso trabalho tem como objetivo analisar uma partida específica, seus parâmetros de entrada, e gerar uma saída o ganhador, perdedor ou o empate. Como em nosso banco de dados, existem partidas de longa data, iremos dividir nossa predição por campeonatos passados. Então, por exemplo, iremos analisar todas as partidas do Campeonato Brasileiro de 2015, e iremos comparar os resultados gerados, com os resultados verdadeiros. Então, iremos treinar o modelo, deixar ele o mais apropriado possível.

Depois de ter sido gerado o melhor modelo, iremos tentar aplicar esse modelo com o Campeonato Brasileiro de 2016. Caso não haja um resultado muito bom, iremos melhorar o modelo para que encaixe da melhor forma para o campeonato daquele ano, e fazer uma comparação de resultados. A comparação será entre o resultado com o modelo de 2050, resultado obtido com o novo modelo e o resultado real. Iremos fazer isso para todos os anos, até 2023. Depois disso, ao adquirir o melhor modelo para o Brasileirão de 2023 iremos utilizar o algoritmo para predizer a próxima partida, que ainda não aconteceu efetivamente no momento do teste.

V. Metodologia Experimental

A. Diagrama de bloco

Foi feito um diagrama de bloco para resumir visualmente nosso fluxo de trabalho, desde a coleta e processamento dos dados até a análise dos resultados gerados pelo nosso modelo.

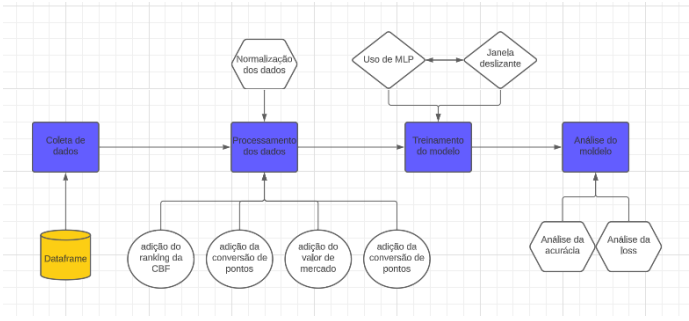


Figura 1. Diagrama de bloco

B. Metodologia

O desafio central deste projeto está em determinar os melhores parâmetros de entrada para treinar o modelo, considerando a natureza imprevisível do futebol (Dubitzky et al., 2020) [12]. No contexto deste estudo, cada linha de nosso conjunto de dados representa uma partida específica de campeonatos ocorridos entre 2015 e 2023, tornando a coleta de dados uma tarefa complexa, uma vez que não existem conjuntos de dados atualmente disponíveis que considerem todas as variáveis que optamos por incluir.

Iniciamos o projeto coletando o máximo possível de entradas que acreditamos influenciar o resultado de uma partida. Entre estas entradas, incluímos as ODDS das partidas, o histórico de jogos entre as duas equipes em questão, informações sobre jogos anteriores em casa e fora, e a composição da equipe no momento da partida, entre outras. A influência de cada entrada foi analisada de maneira criteriosa, e os detalhes desta análise são apresentados na seção 'Base de Dados' deste estudo.

Após a montagem do conjunto de dados, com todas as entradas consideradas importantes, um modelo foi treinado para todas as partidas de cada temporada especificada. Devido à natureza temporal das partidas de um campeonato, adotamos uma metodologia que inclui a implementação de séries temporais e janelas deslizantes para a análise. Foi utilizado o método de Redes Neurais de Múltiplas Camadas (MLP) para a modelagem, e vários modelos foram testados.

A seleção do melhor modelo foi feita com base no desempenho no conjunto de validação, considerando métricas de avaliação de classificação, como acurácia balanceada, precisão, revocação e a pontuação F1 (Costa, 2022) [18].

Uma vez selecionado o melhor modelo, este foi usado para fazer predições e gerar métricas para a avaliação de desempenho, incluindo acurácia balanceada, precisão, revocação e a pontuação F1. Foram gerados gráficos ilustrando a evolução da perda do conjunto de validação ao longo das épocas, além de um histograma de classes para exibir a distribuição das previsões de classes do modelo.

Para um entendimento mais aprofundado, foram geradas matrizes de confusão para cada time individualmente, juntamente com a acurácia balanceada. Adicionalmente, criamos gráficos que representam a acurácia do campeonato com janelas deslizantes de tamanhos variando de 1 a 21, permitindo uma análise mais aprofundada dos resultados ao longo do tempo.

Por fim, fizemos uma análise comparativa entre diferentes temporadas e modelos, possibilitando uma avaliação mais detalhada do desempenho geral do nosso método.

C. Base de Dados

Cada linha do Data Frame desse projeto é uma partida de futebol. Dito isso, entende-se que quanto maiores as informações prévias de uma partida, mais preciso será o algoritmo. Nosso Data Frame é uma junção de outros Data Frames, além de algumas informações colocadas manualmente. Existirá um Data Frame por temporada (estamos analisando os campeonatos

Brasileiros de 2015 - 2023), sendo que cada um deles contará com os 20 times participantes e as 380 partidas disputadas no campeonato (com exceção de 2016 que teve uma partida anulada, entre Atlético Mineiro e Chapecoense devido ao acidente que o time da Chapecoense sofreu de avião).

Primeiramente, o Data Frame principal é proveniente do site Football-Data.co.uk Nele, temos as seguintes colunas: Country, League, Season, Date, Time, Home, Away, HG, AG, Res, PH, PD, PA, MaxH, MaxD, MaxA, AvgH, AvgD, AvgA. Abaixo está explicado o que cada coluna significa.

- Country: País em que a partida está acontecendo (no caso só temos partidas do Brasil)
- League: A liga em que o campeonato está ocorrendo (também só foi adquirido informações referentes aos Campeonatos Brasileiros).
- Season: Referente a que temporada o jogo ocorreu.
- Date: A data em que o jogo foi jogado.
- Time: O horário do jogo
- Home: O time que jogou a respectiva partida como mandante (então o time que recebeu o time adversário em seu estádio)
- Away: O time que jogou a respectiva partida como visitante (então o time que não jogou no próprio estádio)
- HG: Número de gols marcados pelo time que jogou como mandante
- AG: Número de gols marcados pelo time que jogou como visitante
- Res: O resultado da partida. Sendo H = vitória do time que jogou como mandante, D = empate entre as equipes, A = vitória do time que jogou como visitante
- PH: Probabilidade (ODD) de vitória do time que jogou como mandante, segunda a Pinnacle (site de apostas)
- PA: Probabilidade (ODD) de vitória do time que jogou como visitante, segunda a Pinnacle (site de apostas)
- MaxH: Probabilidade (ODD) máxima de vitória do time mandante
- MaxD: Probabilidade (ODD) máxima de empate do time mandante
- MaxA: Probabilidade (ODD) máxima de vitória do time visitante
- AvgH: Probabilidade (ODD) média de vitória do time mandante
- AvgD: Probabilidade (ODD) média de empate
- AvgA: Probabilidade (ODD) média de vitória do time visitante

Esses dados serão levados em consideração para a previsão de partidas. Porém, foram adicionadas, manualmente, mais algumas informações. Tais como o VH, VA, Rank_Home e Rank_Away:

- VH: valor de todo o elenco do time mandante
- VA: valor de todo o elenco do time visitante
- Rank_Home: posição no ranking feito pela CBF, dos melhores times brasileiros, do time mandante
- Rank_Away: posição no ranking feito pela CBF, dos melhores times brasileiros, do time visitante

Vale ressaltar que, no mundo inteiro, existem as chamadas "janelas de transferências", que são períodos em que as equipes liberam/vendem jogadores e compram outros jogadores. Há duas janelas de transferências por ano, então, para os valores de VH e VA consideramos um valor para as equipes antes e depois dessa janela de transferência, para aquele respectivo ano.

Além disso, em outro Data Frame disponibilizado também pelo Transfermarkt, foi adquirida a colocacao_man e colocacao_vis, que são definidas da seguinte forma

- colocacao_man: Refere-se a posição da equipe mandante na rodada em que aconteceu a partida. Cada rodada, as posições dos times mudam, então para cada jogo, contém a informação de que posição o time está, no momento daquela partida, na tabela geral.
- colocacao_vis: Refere-se a posição da equipe visitante na rodada em que aconteceu a partida. Cada rodada, as posições dos times mudam, então para cada jogo, contém a informação de que posição o time está, no momento daquela partida, na tabela geral.

Infelizmente, esse Data Frame tem dados limitados até 2020, logo, para os Campeonatos Brasileiros de 2021 em diante, as posições de cada time por rodada foram adquiridas no site da CBF e adicionadas manualmente.

Os dados do trabalho, por enquanto, englobam como as pessoas acham que a partida resultará (ODDs), a disparidade de elencos (valor dos elencos) e disparidade entre posições das equipes (ranking CBF e posição na rodada). Vale ressaltar que o desempenho atual das equipes não está sendo levado em consideração, uma vez que é algo relativo. Então, como uma tentativa de quantificar e trazer ao projeto, foram adicionadas 4 novas colunas: Last_X_Pts_Ratio_Home, Last_X_Pts_Ratio_Away, Last_X_Home_Pts_Ratio, Last_X_Away_Pts_Ratio.

- Last_X_Pts_Ratio_Home: Porcentagem de conversão de pontos nas últimas X partidas disputadas pelo time mandante. Para calcular está sendo usada a seguinte lógica (pontos convertidos (nas últimas X partidas)/pontos totais (nas últimas X partidas)). Assim será considerado quantos pontos cada time converter nas últimas partidas, em porcentagem.
- Last_X_Pts_Ratio_Away: Porcentagem de conversão de pontos nas últimas X partidas disputadas pelo time visitante. Para calcular está sendo usada a seguinte lógica (pontos convertidos (nas últimas X partidas)/pontos totais (nas últimas X partidas)). Assim será considerado quantos pontos cada time converter nas últimas partidas, em porcentagem.
- Last_X_Home_Pts_Ratio: Porcentagem de conversão de pontos nas últimas X partidas do time mandante como time mandante. Basicamente, o cálculo será feito analisando as últimas X partidas do time mandante jogando como mandante. O cálculo é feito da seguinte maneira (pontos convertidos como time mandante (nas últimas X partidas como mandante)/pontos totais em partidas

(nas últimas X partidas como mandante)). Assim será considerado o desempenho do time mandante jogando como mandante, em porcentagem.

- Last_X_Away_Pts_Ratio: Porcentagem de conversão de pontos nas últimas X partidas do time visitante como time visitante. Basicamente, o cálculo será feito analisando as últimas X partidas do time visitante jogando como visitante. O cálculo é feito da seguinte maneira (pontos convertidos como time visitante (nas últimas X partidas como visitante)/pontos totais em partidas (nas últimas X partidas como visitante)). Assim será considerado o desempenho do time visitante jogando como visitante, em porcentagem.

É possível que, ao começo do campeonato, os times não tenham disputado X partidas ou não tenham disputado X partidas como visitante ou mandante. Então, para as primeiras X partidas, para as colunas Last_X_Pts_Ratio_Home Last_X_Pts_Ratio_Away, foi adicionado um valor padrão de 0.5 (50%), ou seja, 50% de vitória/derrota, para ficar um valor padrão. Depois que X partidas foram disputadas, então passa-se a calcular as porcentagens. Para as colunas Last_X_Home_Pts_Ratio Last_X_Away_Pts_Ratio, segue-se a mesma lógica, para as primeiras X partidas como mandante/visitante é adicionado um valor padrão de 0.5 (50%), e depois que X partidas forem disputadas pelos times seja como mandante ou visitante, então o algoritmo calcula a porcentagem de cada um.

Essas são todas as colunas que serão consideradas no trabalho. Entretanto, algumas omo 'Season', 'Country', 'League', 'HG', 'AG' serão retiradas para o treino do algoritmo (pois no caso das 3 primeiras colunas não são informações relevantes, e no caso dos gols, não se tem informação de quantos gols foram feitos antes da partida acontecer, então essa informação precisa ser retirada). Por fim, vale ressaltar que, caso o modelo como um todo seja treinado, o algoritmo irá encontrar padrões em partidas de times diferentes, o que pode não ser verdade. Para arrumar isso, o Data Frame total será separado, para cada temporada do Campeonato, em 20 Data Frames que representam os 20 times que irão disputar aquela respectiva temporada do Brasileirão. Cada um deles terá todas as partidas disputadas e as informações dessas partidas. Portanto, nosso modelo será especializado para cada equipe individualmente, melhorando, provavelmente, ainda mais na precisão como um todo.

D. Protocolo de Validação

Estamos abordando a parte de validação do nosso projeto, trabalhando com dados de uma série temporal, ou seja, os jogos de futebol que ocorrem em uma sequência específica. Para tais dados, adotamos uma estratégia de validação específica baseada no conceito de janelas deslizantes, ao invés de usar a técnica de validação cruzada K-Fold.

As séries temporais são uma sequência de pontos de dados, coletados em intervalos de tempo sucessivos. No contexto do futebol, cada jogo ocorre em um ponto específico no tempo, criando uma sequência de jogos ao longo do tempo. Os dados de séries temporais possuem uma característica importante: eles são dependentes do tempo. Isso significa que a ordem dos dados importa e que os padrões nos dados mudam com o tempo. Esta natureza temporal é o que requer técnicas de análise especializadas, como a estratégia de janelas deslizantes.

A técnica de janelas deslizantes é comumente usada em séries temporais onde uma "janela", de um determinado tamanho, é escolhida, e o modelo é treinado nesta janela de dados. A janela então "desliza"ao longo do tempo, permitindo que o modelo seja treinado em diferentes segmentos dos dados. No nosso caso, estamos usando uma janela deslizante de tamanho 3 para treinar o modelo em um número de jogos e então testar o modelo no próximo jogo.

Inicialmente, dividimos o conjunto de dados de cada equipe em conjuntos de treino e teste, utilizando a função train_test_split da biblioteca Scikit-learn com um split de 70% dos dados para treinamento e os restantes 30% para teste. É importante ressaltar que sempre mantemos a ordem cronológica dos jogos. Esse procedimento é comum para evitar o sobreajuste ao modelo e garantir que nosso modelo seja capaz de generalizar bem para dados não vistos, já que nosso algoritmo é preditivo.

Neste contexto, garantimos a reprodutibilidade do código estabelecendo uma semente fixa (random seed), que assegura que todas as operações aleatórias irão produzir os mesmos resultados cada vez que o código é executado.

Dentro de cada divisão, implementamos a janela deslizante, selecionando uma janela de dados para treinamento e usando o próximo ponto de dados como validação. Este processo é repetido, com a "janela"se movendo um jogo por vez. Essa abordagem é adequada para o nosso conjunto de dados, pois leva em conta a natureza sequencial e temporal dos jogos de futebol.

Este protocolo de validação garante que estamos sempre treinando em dados do passado para prever o futuro, o que é a abordagem correta ao lidar com dados de séries temporais. Ele melhora a capacidade do nosso modelo de generalizar para novos dados e evita vazamento de dados, onde o modelo tem acesso a informações futuras que não estariam disponíveis na vida real.

E. Treinamento do modelo

No experimento realizado, o algoritmo de Aprendizado de Máquina foi treinado e testado através de um processo iterativo conhecido como modelo de épocas. O processo começa com o conjunto de treinamento, que foi cuidadosamente normalizado para assegurar que nenhuma característica tivesse mais influência do que outra.

O conjunto de treinamento foi então dividido em partes menores, conhecidas como 'batches', cada uma contendo 32 valores. A utilização de batches é uma estratégia eficiente para treinar o algoritmo, pois permite o processamento de uma pequena parcela dos dados de cada vez, o que otimiza o uso da memória e acelera o processo de aprendizado.

Durante o treinamento, foi adotada uma janela deslizante com tamanho igual a 3. Essa técnica de janela deslizante permite que o modelo aprenda sequências de dados ao longo do tempo, levando em conta não apenas o valor atual, mas também os valores anteriores. Isso é particularmente útil em problemas temporais, onde a sequência dos dados tem importância.

Em cada época, ou seja, em cada ciclo completo de processamento do conjunto de treinamento, o algoritmo gera previsões chamadas de 'outputs'. Esses outputs são valores que o algoritmo calcula com base nas características dos dados de treinamento, representando as previsões de classe feitas pelo modelo.

Através da função softmax, esses outputs são transformados em probabilidades, onde cada probabilidade representa a chance do dado pertencer a uma determinada classe. No entanto, quando o algoritmo tem dificuldades para identificar a verdadeira classe de um dado, as probabilidades das diferentes classes tornam-se bastante semelhantes. Isso é indesejável, pois reflete a incerteza do algoritmo na classificação.

Para quantificar essa incerteza, calculamos a entropia das previsões. A entropia, que é uma medida de variabilidade ou incerteza, é alta quando as probabilidades das classes são semelhantes. Uma alta entropia é prejudicial para o desempenho do algoritmo, pois indica que o modelo não está confiante em suas previsões.

Quando a entropia é superior a 0.5, ou seja, quando a incerteza é alta, o algoritmo calcula a perda, também conhecida como 'loss'. Essa perda é uma medida do erro do modelo, ou seja, da diferença entre as previsões do algoritmo e os verdadeiros valores de classe.

A perda é então usada em um processo chamado retropropagação, que ajusta os pesos do modelo. A retropropagação é um algoritmo que atualiza os pesos do modelo com o objetivo de minimizar a perda. A ideia é que, ao ajustar os pesos de acordo com a perda, o modelo aprenda a fazer previsões mais precisas e, assim, reduza a incerteza em suas previsões futuras. Essa atualização de pesos é realizada apenas para previsões com alta entropia, pois é nesses casos que a atualização pode ser mais eficaz para melhorar o desempenho do modelo.

Em seguida, o desempenho do modelo é avaliado em cada época O conjunto de teste, dividido em batches de 16 valores, é usado para esta avaliação. A perda do conjunto de validação é calculada e a acurácia balanceada é determinada para cada teste. A acurácia balanceada é uma métrica que leva em conta o desequilíbrio da classe, fornecendo uma visão mais justa do desempenho do modelo.

Ao final de cada época, o modelo com a melhor acurácia balanceada é salvo. Este modelo será utilizado para fazer previsões em dados novos ou não vistos, representando a versão mais precisa do nosso modelo de aprendizado de máquina após essa época específica de treinamento e teste.

VI. Resultados finais esperados

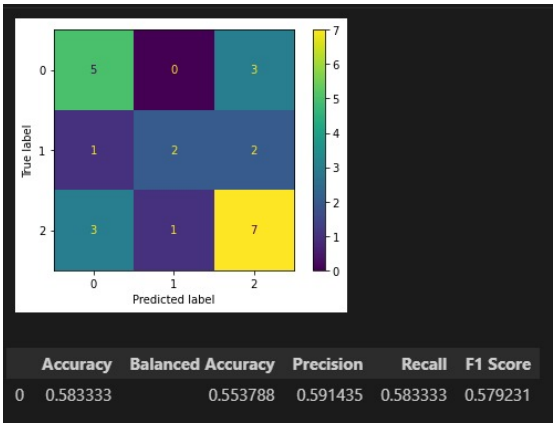
O resultado esperado será, por temporada, gerar uma matriz de confusão de cada time, e se possível, segundo aquele modelo, como a tabela do campeonato seria em comparação com o real. Além disso, será comparado acurácias e modelos entre os campeonatos, qual a diferença entre modelos de campeonato, uma possível explicação do porque determinado modelo é mais apropriado para aquela temporada...

Além disso, caso tudo seja cumprida com certa antecedência e fique com resultados relativamente aceitáveis, então iremos tentar procurar o melhor modelo para o Campeonato Brasileiro de 2023, e tentar predizer o resultado, neste campeonato, de partidas que ainda não aconteceram, esperar um tempo até elas acontecerem, e depois comparar os resultados obtidos e os reais. Como nossas entradas consideram coisas "momentâneas" para determinada partida, como desfalques, titulares, as ODDS, seria difícil predizer campeonatos inteiros, mas partidas que estão para acontecer, que os parâmetros já estão determinados, nosso algoritmo tentaria predizer o resultado.

VII. Resultados e discussões

Nós conduzimos testes estatísticos para verificar a significância das diferenças observadas. Nossa abordagem inicial foi a validação cruzada com janela deslizante para lidar com a natureza temporal dos dados. No entanto, simplificamos o protocolo de validação e agora usamos apenas uma janela deslizante, sem validação cruzada. Essa abordagem pode ser revisada e ajustada no futuro com base nos resultados obtidos.

Os resultados dos nossos experimentos iniciais são apresentados nas tabelas e gráficos a seguir. Fornecemos uma matriz de confusão que ilustra a performance do modelo nas diferentes classes de resultado (vitória, empate e derrota), e gráficos de perda durante o treinamento e a validação do modelo.



	Accuracy	Balanced Accuracy	Precision	Recall	F1 Score
0	0.583333	0.553788	0.591435	0.583333	0.579231

Figura 2. Primeira matriz de confusão de 2015

Os primeiros resultados indicam que o modelo precisa de ajustes para lidar com o overfitting e o desequilíbrio das classes. Planejamos abordar essas questões nos próximos

experimentos, explorando diferentes configurações de hiperparâmetros e técnicas para reduzir o overfitting. Dependendo dos resultados desses ajustes, podemos também considerar a mudança no protocolo de validação para incluir a validação cruzada sem janelas deslizantes.

Após a primeira tentativa, uma nova foi feita para verificar se a acurácia melhorava. O resultado está no gráfico abaixo, que, como mostrado, a acurácia não melhorou, até piorou em valores numéricos. Em testes posteriores ela continuou com os mesmos valores mostrados na matriz de confusão 2.**??**

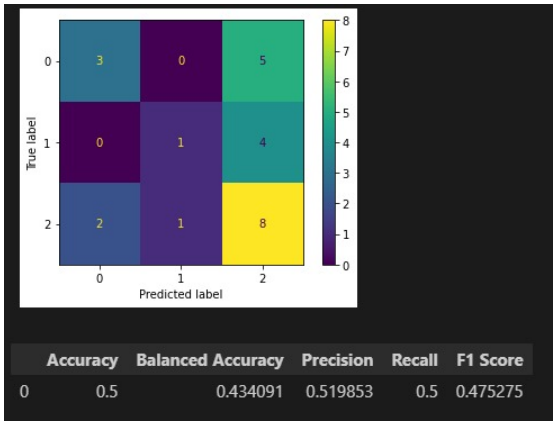


Figura 3. Segunda matriz de confusão de 2015 após novo teste

Foi plotado também um gráfico de curva de aprendizado que será mostrado abaixo. A cada época, ou seja, a cada passagem completa pelo conjunto de dados de treinamento, calculamos a perda do modelo. A figura a seguir mostra o gráfico de Loss, onde podemos observar a evolução da perda durante o treinamento.

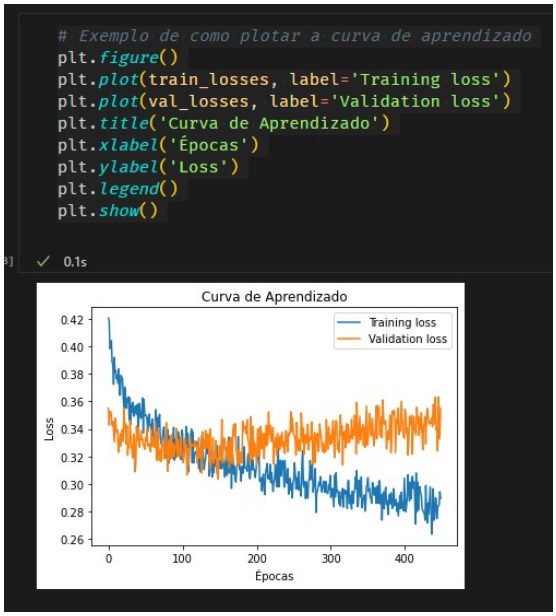


Figura 4. Gráfico da curva de aprendizado de 2015

Ao examinar a curva de aprendizado, observamos um aumento da perda de validação, indicando um overfitting ao conjunto de treino. O histograma das previsões também mostra um desequilíbrio na predição das classes, com o modelo tendendo a prever mais frequentemente uma das classes.

Nossa análise dos resultados mostra que a técnica MLP se saiu bem em nossos dados, mas acreditamos que ajustes adicionais nos hiperparâmetros e a exploração de outras técnicas podem melhorar ainda mais a acurácia e a perda de nosso sistema de previsão. Continuaremos a refinar nosso modelo para obter melhores resultados em futuras previsões do Campeonato Brasileiro.

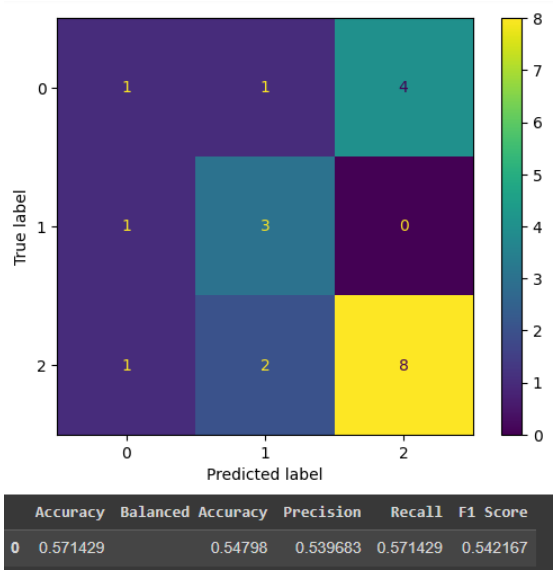
Os resultados de nossos experimentos iniciais são promissores, mas acreditamos que há espaço para melhorias. Abaixo, apresentamos uma tabela que resume os resultados de nosso estudo.

Figura 5. Histograma de 2015

A. Resultados da predição por campeonatos

Neste relatório, analisaremos os resultados de um modelo de aprendizado de máquina para prever o resultado de partidas de futebol por ano, onde as classes representam: 0 = vitória do time mandante, 1 = empate e 2 = vitória do time visitante. Os resultados abrangem os anos de 2015 a 2023. Abaixo iniciamos a análise pelo campeonato de 2015.

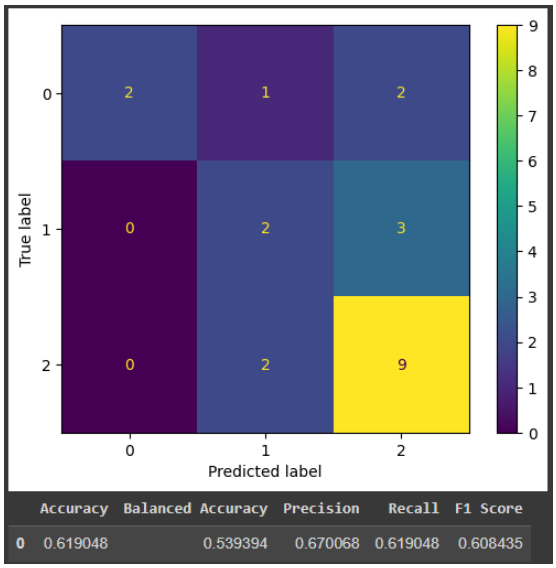
Em 2015 o modelo teve uma precisão de 57,14%, uma precisão balanceada de 54,80%, precisão de 53,97%, recall de 57,14% e F1 Score de 54,22%. As maiores dificuldades foram encontradas ao prever a vitória do time visitante, com 4 erros.



	Accuracy	Balanced Accuracy	Precision	Recall	F1 Score
0	0.571429	0.54798	0.539683	0.571429	0.542167

Figura 6. Matriz de confusão de 2015 definitiva

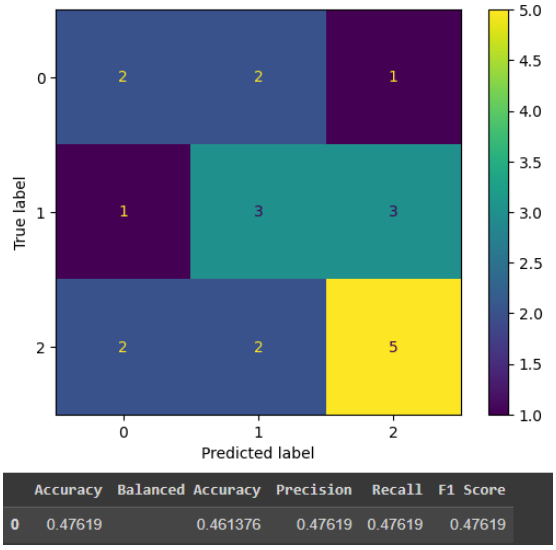
Em 2016 houve um aumento na precisão para 61,90%, mas uma ligeira queda na precisão balanceada para 53,94%. A precisão subiu para 67,01%, enquanto o recall permaneceu constante em 61,90% e o F1 Score subiu para 60,84%. Mais uma vez, a maior dificuldade foi prever a vitória do time visitante.



	Accuracy	Balanced Accuracy	Precision	Recall	F1 Score
0	0.619048	0.539394	0.670068	0.619048	0.608435

Figura 7. Matriz de confusão de 2016

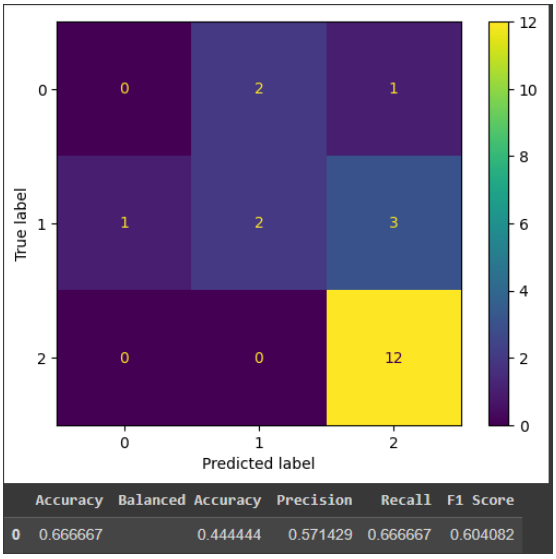
No ano de 2017 viu-se uma queda na precisão para 47,62% e todos os outros índices também caíram para 46,14% (precisão balanceada), 47,62% (precisão), 47,62% (recall) e 47,62% (F1 Score). As predições foram distribuídas de maneira relativamente igual entre todas as classes.



	Accuracy	Balanced Accuracy	Precision	Recall	F1 Score
0	0.47619	0.461376	0.47619	0.47619	0.47619

Figura 8. Matriz de confusão de 2017

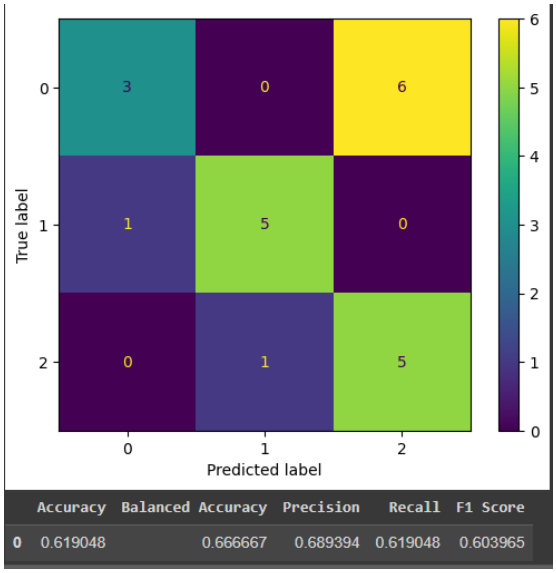
Em 2018 a precisão aumentou para 66,67%, mas a precisão balanceada caiu para 44,44%. A precisão foi de 57,14%, o recall foi de 66,67% e o F1 Score foi de 60,41%. Neste ano, o modelo teve um desempenho extremamente bom na previsão de vitórias do time visitante, acertando todas as 12 partidas.



	Accuracy	Balanced Accuracy	Precision	Recall	F1 Score
0	0.666667	0.444444	0.571429	0.666667	0.604082

Figura 9. Matriz de confusão de 2018

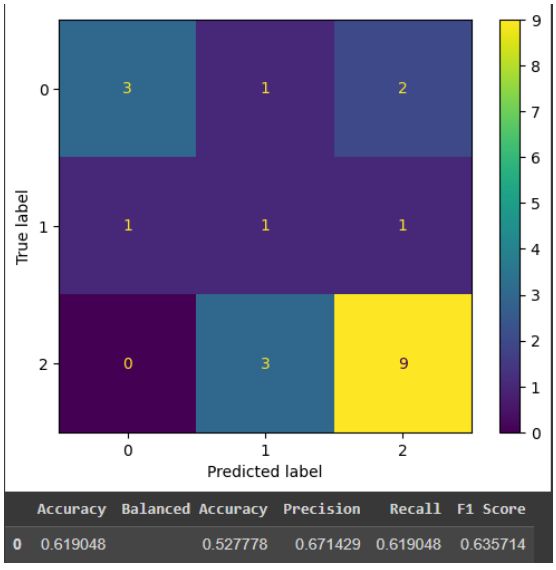
O ano de 2019 viu uma ligeira queda na precisão para 61,90%, mas um aumento na precisão balanceada para 66,67%. A precisão foi de 68,94%, o recall foi de 61,90% e o F1 Score foi de 60,40%. O modelo teve uma dificuldade particular em prever vitórias do time da casa, errando 6 vezes.



	Accuracy	Balanced Accuracy	Precision	Recall	F1 Score
0	0.619048	0.666667	0.689394	0.619048	0.603965

Figura 10. Matriz de confusõ de 2019

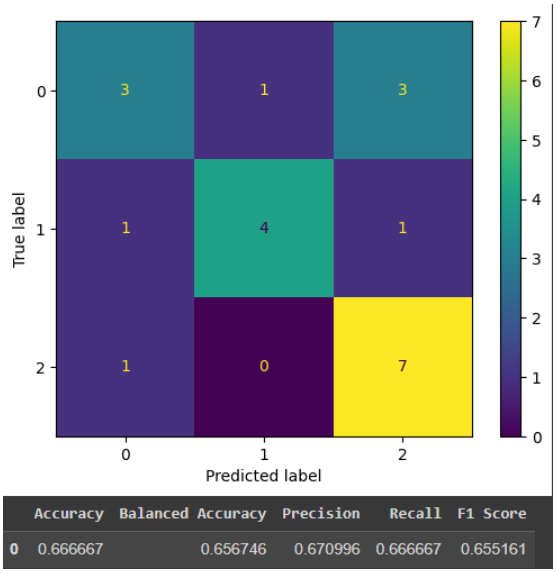
Em 2020 a precisão se manteve estável em 61,90%, mas a precisão balanceada caiu para 52,78%. A precisão subiu para 67,14%, o recall se manteve constante em 61,90% e o F1 Score subiu para 63,57%. Novamente, a maior dificuldade foi prever a vitória do time da casa.



	Accuracy	Balanced Accuracy	Precision	Recall	F1 Score
0	0.619048	0.527778	0.671429	0.619048	0.635714

Figura 11. Matriz de confusão de 2020

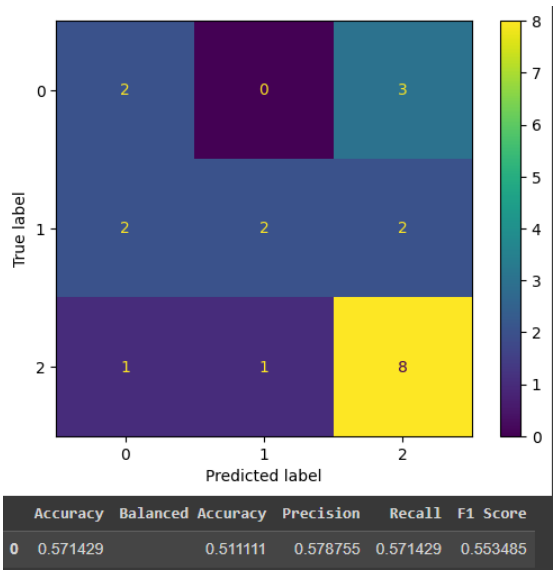
Em 2021 viu-se um aumento na precisão para 66,67% e na precisão balanceada para 65,67%. A precisão foi de 67,10%, o recall foi de 66,67% e o F1 Score foi de 65,52%. O modelo teve uma dificuldade particular em prever empates, errando duas vezes.



	Accuracy	Balanced Accuracy	Precision	Recall	F1 Score
0	0.666667	0.656746	0.670996	0.666667	0.655161

Figura 12. Matriz de confusão de 2021

Em 2022 houve uma queda na precisão para 57,14% e na precisão balanceada para 51,11%. A precisão foi de 57,88%, o recall foi de 57,14% e o F1 Score foi de 55,35%. Neste ano, o modelo teve dificuldades semelhantes em prever todas as classes de resultados.



	Accuracy	Balanced Accuracy	Precision	Recall	F1 Score
0	0.571429	0.511111	0.578755	0.571429	0.553485

Figura 13. Matriz de confusão de 2022

Até o momento, o modelo alcançou a perfeição em 2023, com todos os índices atingindo 100%. No entanto, é importante ressaltar que até agora houve apenas 9 partidas em 2023 e todas foram vitórias do time visitante. Isso pode não ser representativo do desempenho futuro do modelo.

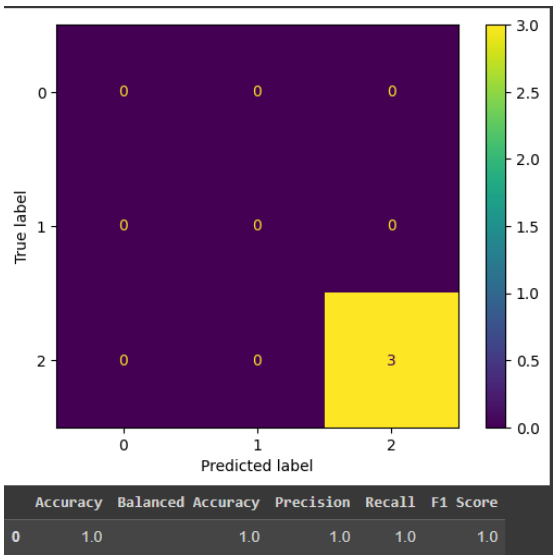


Figura 14. Matriz de confusão de 2023

De forma geral, o modelo mostrou uma tendência de melhora na precisão de 2015 a 2021, com um pequeno declínio em 2022. A precisão balanceada variou um pouco mais, atingindo seu pico em 2019 e seu ponto mais baixo em 2018.

A precisão do modelo também melhorou em geral, atingindo um pico em 2019. No entanto, o recall se manteve relativamente estável ao longo dos anos, com pequenas variações. O F1 Score seguiu uma tendência similar à da precisão.

É notável que o modelo apresenta um desafio maior na previsão de vitórias do time da casa (classe 0) ou empates (classe 1), comparativamente ao desempenho na previsão de vitórias do time visitante (classe 2). Em outras palavras, o algoritmo parece ter uma facilidade maior em identificar quando o time visitante irá vencer, ao contrário dos cenários de vitória do time da casa ou de empates.

Vale ressaltar que o modelo está projetado para prever resultados de partidas em um campeonato de futebol, não necessariamente para prever resultados de um time específico. Então, quando nos referimos a ”partidas em casa”, estamos considerando as partidas onde qualquer time, na posição de anfitrião no campeonato, joga em seu próprio estádio.

Por este motivo, quando mencionamos a dificuldade do modelo em prever vitórias ”em casa”, não estamos analisando a performance de um time específico dentro do seu estádio, mas sim a performance geral dos times da casa em todas as partidas do campeonato.

Quanto aos empates, a previsão se torna ainda mais complexa, visto que pequenas variações no jogo podem facilmente transformar um empate em uma vitória para qualquer um dos lados. Isso representa um desafio considerável para o modelo.

O desempenho mais eficiente na previsão de vitórias do time visitante pode ser um reflexo do desbalanceamento das classes nos dados. Se existem mais vitórias fora de casa no conjunto de dados utilizado para treinamento, o modelo terá mais exemplos para aprender e poderá ser mais preciso nesses casos.

Essa análise indica que estratégias de balanceamento de classes, como técnicas de oversampling, undersampling ou geração sintética de dados, podem ser úteis para melhorar a precisão do modelo na previsão de vitórias em casa e empates, proporcionando um desempenho mais balanceado entre as classes.

Em resumo, o modelo tem mostrado melhoria constante com o tempo, mas ainda há espaço para otimizações, especialmente na previsão da vitória do time da casa. É aconselhável continuar treinando e ajustando o modelo para melhorar sua performance.

Abaixo temos um gráfico que mostra a evolução de cada métrica de avaliação ao longo dos anos.

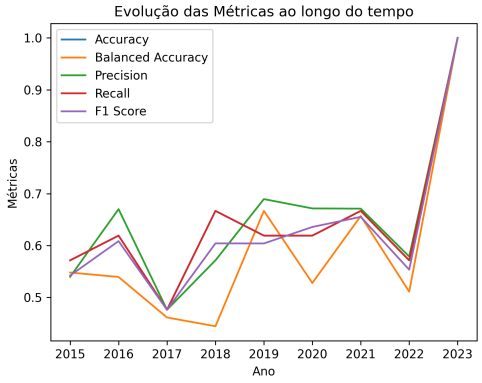


Figura 15. Evolução das métricas ao longo dos anos

1) Análise da variação de janelas: 2015: Começamos com uma acurácia de 0.47 na janela 1, tendo um crescimento razoável até a janela 7, com uma acurácia de 0.57. A partir daqui, notamos algumas oscilações até a janela 15, que chega a uma acurácia impressionante de 0.78. No entanto, não podemos tirar conclusões precipitadas, pois as janelas 16 a 21 caem para 0, provavelmente devido à falta de dados nessas janelas.

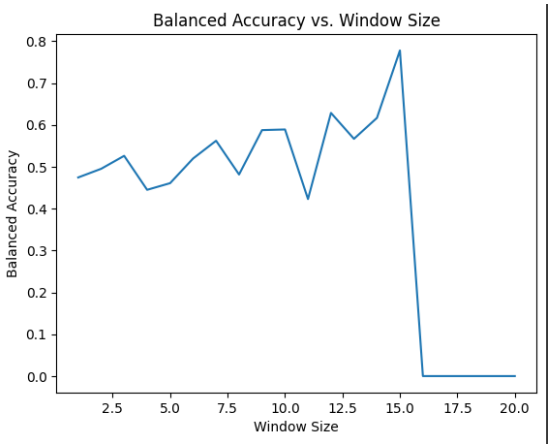


Figura 16. Variação de janelas para 2015

2016: A acurácia inicia em 0.44 na janela 1 e se mantém relativamente estável até a janela 8. A partir daí, há uma melhoria constante até a janela 13, alcançando 0.70. Após isso, os valores oscilam um pouco, mas ainda se mantêm relativamente altos até a janela 15. Mais uma vez, as janelas 16 a 21 caem para 0.

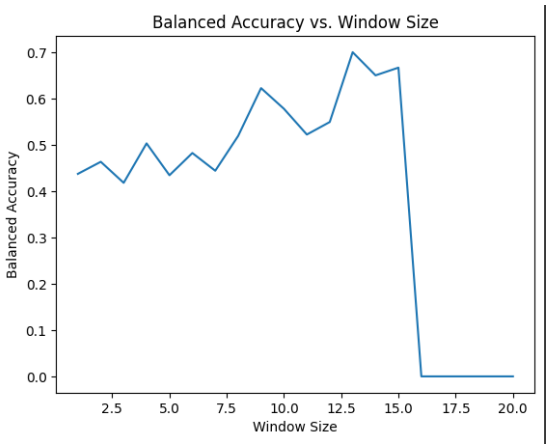


Figura 17. Variação de janelas para 2016

2017: Começando com uma acurácia de 0.43 na janela 1, a acurácia permanece bastante constante até a janela 10, variando principalmente entre 0.40 e 0.53. A partir da janela 11, há uma melhora geral na acurácia, culminando na janela 15 com um pico de 0.86. As janelas 16 a 21 novamente caem para 0.

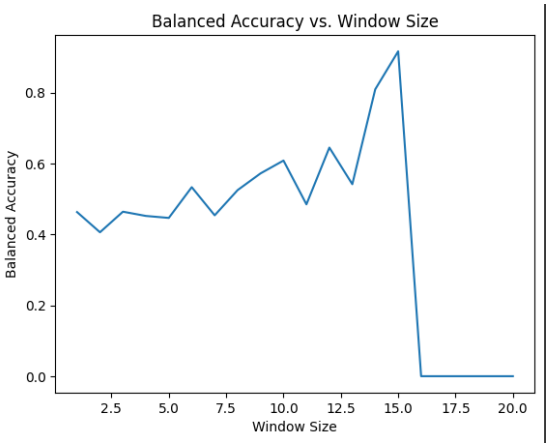


Figura 18. Variação de janelas para 2017

2018: Iniciando em 0.50 na janela 1, a acurácia oscila entre 0.42 e 0.53 até a janela 9. Então, começa a aumentar constantemente até a janela 15, alcançando 0.76. As janelas 16 a 21 caem para 0.

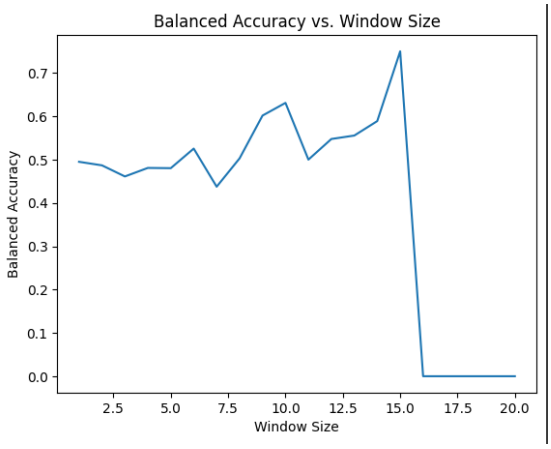


Figura 19. Variação de janelas para 2018

2019: Começando com uma acurácia de 0.44 na janela 1, o modelo tem um desempenho estável até a janela 10. A partir da janela 11, a acurácia começa a aumentar, chegando a um pico de 0.87 na janela 15. As janelas 16 a 21 caem para 0.

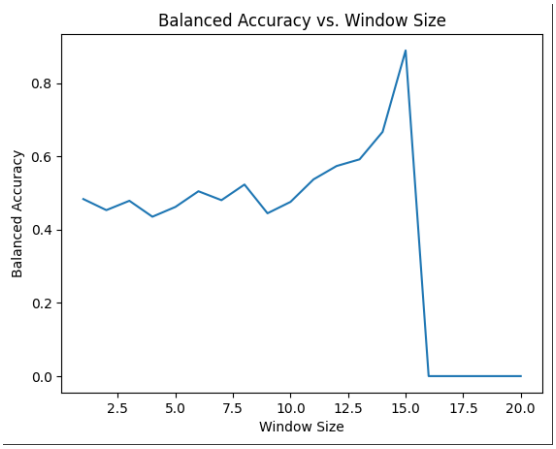


Figura 20. Variação de janelas para 2019

2020: Iniciando com uma acurácia de 0.45 na janela 1, o modelo oscila levemente até a janela 12, quando atinge um pico de 0.74. Após isso, há uma queda, mas ainda assim, o modelo consegue se manter relativamente alto até a janela 15. As janelas 16 a 21 caem para 0.

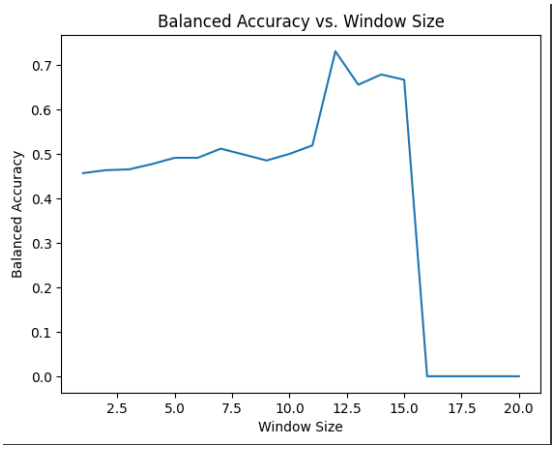


Figura 21. Variação de janelas para 2020

2021: Iniciando com uma acurácia de 0.42 na janela 1, o modelo se mantém relativamente estável até a janela 10. A partir daí, há uma subida até a janela 15, onde atinge 0.86. As janelas 16 a 21 caem para 0.

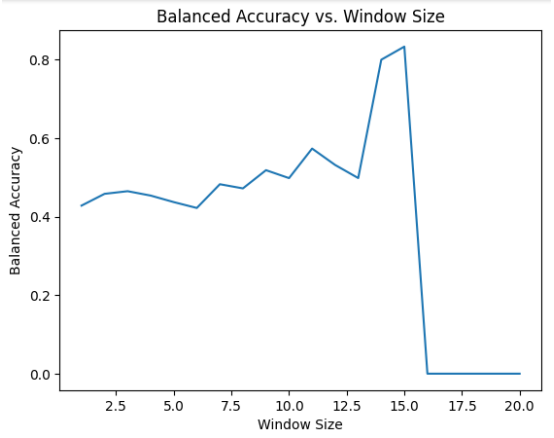


Figura 22. Variação de janelas para 2021

2022: Começando com uma acurácia de 0.43 na janela 1, a acurácia oscila até a janela 10. A partir daí, vemos um aumento geral até a janela 15, que chega a 0.81. As janelas 16 a 21 caem para 0.

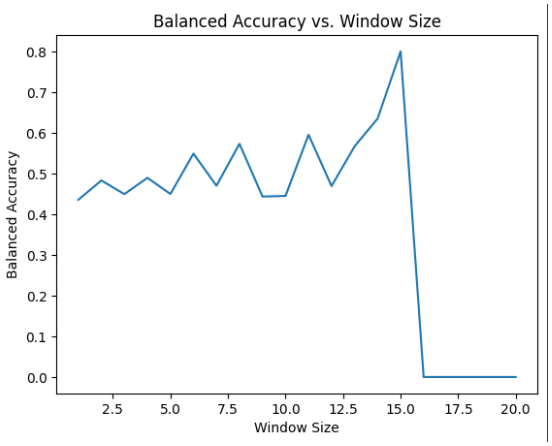


Figura 23. Variação de janelas para 2022

2023: A nossa base de dados a respeito do campeonato brasileiro de 2023 é bem pequena, contendo apenas as 9 primeiras rodadas, então, pelo baixo número de jogos, não foi possível fazer uma análise das janelas do campeonato desse ano, que inclusive ainda está em andamento.

2) Análise Coletiva: Ao analisar todos os anos juntos, podemos ver alguns padrões comuns. Em todos os anos, a acurácia tende a ser mais baixa no início (janelas 1 a 6), apresenta alguma oscilação no meio (janelas 7 a 10), e então melhora notavelmente em direção à janela 15. Isso sugere que o modelo pode se beneficiar de uma quantidade maior de dados históricos para fazer previsões mais precisas.

Também é interessante notar que, em todos os anos, a acurácia cai para 0 nas janelas 16 a 21. Isso provavelmente se deve à falta de dados nessas janelas, o que impede o modelo de fazer previsões precisas.

Em resumo, o desempenho do modelo tende a melhorar à medida que a janela de dados aumenta, pelo menos até a janela 15. Além disso, em muitos dos anos, há um aumento notável na acurácia ao passar da janela 10 para a janela 15. Isso sugere que essa pode ser uma faixa crucial para a melhoria do desempenho do modelo.

B. Resultados da predição por times

Os dados apresentados correspondem à classificação final dos times nos respectivos campeonatos de 2015 a 2022, juntamente com a acurácia balanceada de cada ano. A acurácia balanceada é uma métrica de avaliação que leva em consideração o desempenho do modelo em todas as classes, sendo mais indicada quando temos um desbalanceamento entre as classes. Vamos analisar os dados ano a ano. Abaixo temos um mapa de calor com os dados das acurácias balanceadas e a média de cada ano.

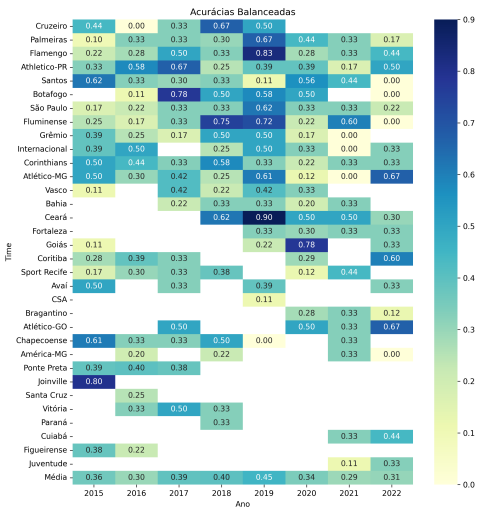


Figura 24. Mapa de calor com as acurácias individuais de cada time por ano

Em 2015, o Corinthians teve uma acurácia balanceada de 0,5, indicando que o modelo teve uma performance média ao prever os resultados deste time. O Santos, por outro lado, teve uma acurácia balanceada de 0,62, sugerindo que o modelo foi mais eficaz ao prever os resultados deste time.

Em 2016, o Palmeiras e o Santos tiveram uma acurácia balanceada de 0,33, o que indica uma performance inferior à do ano anterior. No entanto, o Athletico-PR apresentou uma acurácia balanceada de 0,58, mostrando uma melhora na capacidade do modelo de prever os resultados deste time.

No ano de 2017, a acurácia balanceada do Corinthians caiu para 0,33, enquanto que a do Santos manteve-se em 0,3. No entanto, o Botafogo teve uma acurácia balanceada impressionante de 0,78, sugerindo que o modelo teve uma alta capacidade de prever corretamente os resultados deste time.

Em 2018, o Palmeiras manteve sua acurácia balanceada em 0,3, enquanto o Flamengo melhorou para 0,33. Contudo, o Fluminense teve uma alta acurácia balanceada de 0,75, o que indica que o modelo foi particularmente eficaz ao prever os resultados deste time.

Em 2019, o Flamengo apresentou uma acurácia balanceada muito alta de 0,83, sugerindo que o modelo foi extremamente eficaz ao prever os resultados deste time. No entanto, o Santos teve uma acurácia balanceada muito baixa de 0,11, o que indica que o modelo teve dificuldades para prever corretamente os resultados deste time.

No ano de 2020, o Flamengo teve uma queda significativa na acurácia balanceada para 0,28, enquanto o Internacional teve uma acurácia balanceada de 0,33. No entanto, o Goiás teve uma alta acurácia balanceada de 0,78, sugerindo uma alta capacidade do modelo de prever corretamente os resultados deste time.

Em 2021, nenhum time teve uma acurácia balanceada acima de 0,6, o que pode indicar uma dificuldade geral do modelo em prever corretamente os resultados neste ano.

Finalmente, em 2022, a acurácia balanceada do Palmeiras caiu para 0,17, enquanto a do Internacional manteve-se em 0,33. Contudo, o Atlético-MG teve uma alta acurácia balanceada de 0,67, indicando uma alta capacidade do modelo de prever corretamente os resultados deste time.

Estes resultados mostram que a acurácia balanceada variou consideravelmente entre os times e ao longo dos anos. Isso pode ser devido a uma série de fatores, como mudanças nos times, na competição e nos dados de treinamento disponíveis para o modelo.

Os anos de 2015 a 2019 mostram um aumento gradual na acurácia média, alcançando o pico em 2019 com uma acurácia média de 0.45. Este pode ser um indicativo de que, durante esse período, os times ou o modelo utilizado para determinar a acurácia podem ter sido mais eficazes.

No entanto, a partir de 2020, observamos uma queda na acurácia média. O declínio acentuado em 2021 e 2022 para 0.29 e 0.31, respectivamente, pode indicar que os times ou o modelo de previsão estavam menos eficazes durante esses anos.

Os resultados mostram que a acurácia das previsões pode variar de ano para ano, talvez devido a mudanças no desempenho dos times ou na eficácia dos modelos de previsão.

Em resumo, parece haver uma tendência geral de aumento na acurácia das previsões de 2015 a 2019, seguida por uma queda em 2020 a 2022. No entanto, seria útil investigar mais para entender as causas dessas mudanças na acurácia das previsões. Por exemplo, pode ser interessante analisar se houve mudanças significativas na estratégia ou desempenho dos times durante esses períodos.

Sem dúvida, é importante destacar que o modelo foi inicialmente projetado para prever resultados de partidas em um campeonato como um todo, e não para prever resultados de partidas de times específicos. Esta é uma distinção importante, pois o desempenho individual dos times pode variar significativamente devido a uma série de fatores, incluindo a habilidade dos jogadores, as estratégias do time, as condições do jogo, entre outros.

Como resultado, quando tentamos aplicar o modelo a times individuais, é natural que possamos observar uma diminuição na acurácia. Cada time tem seu próprio conjunto de características únicas, incluindo estratégias de jogo, força da equipe, condição física dos jogadores, entre outros, que podem não ser totalmente capturados pelo modelo. Portanto, as previsões podem não ser tão precisas quando comparadas àquelas feitas para o campeonato como um todo.

Isso, no entanto, não diminui a utilidade do modelo. Ele ainda pode fornecer uma visão geral útil de como os times podem se sair em uma competição e ajudar a entender as tendências gerais dentro de um campeonato. No entanto, para previsões mais precisas a nível de time, podem ser necessários modelos mais sofisticados ou específicos, que levem em conta as características individuais de cada time.

VIII. Referências

Referências

[1] MARTINS, Marcus Vinicius Carvalho. *Um modelo para previsão de resultados em partidas do Campeonato Brasileiro de Futebol*. [S. l.], 2018. Disponível em: https://www.cc.ufrrj.br/wpcontent/uploads/2021/12/Marcus_Vinicius___TCC_UFRRJ1.pdf. Acesso em: 10/05/2023.

[2] SANTOS, João Marcos Amorim dos. *Previsões de Resultados em Partidas do Campeonato Brasileiro de Futebol*. Dissertação (mestrado em modelagem matemática) - Programa de Pós-Graduação em Matemática aplicada, FGV, 2019. Disponível em: https://bibliotecadigital.fgv.br/dspace/bitstream/handle/10438/27672/joao_marcos_amorim_dos_santos.pdf?sequence=1&isAllowed=y. Acesso em: 10/05/2023.

[3] RODRIGUES, Fátima; PINTO, Ângelo. *Prediction of football match results with Machine Learning*. Previsões de Resultados em Partidas do Campeonato Brasileiro de Futebol, [S. l.], p. 464-479, 30 maio 2023. Disponível em: https://www.sciencedirect.com/science/article/pii/S1877050922007955.Acesso em: 10/05/2023.

[4] YADA, Vaishnavi. *How to Use Python and Machine Learning to Predict Football Match Winners*. [S. l.], 18 jan. 2023. Disponível em: https://www.kdnuggets.com/2023/01/python-machine-learning-predict-football-match-winners. Acesso em: 29/05/2023.

[5] BOICE, Jay. *How Our Club Soccer Predictions Work*. 1.3. [S. l.], 2 jul. 2020. Disponível em: https://fivethirtyeight.com/methodology/how-our-club-soccer-predictions-work/. Acesso em: 29/05/2023.

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

[7] FARIAS, F. F. *Análise e previsão de resultados de partidas de futebol*. Departamento de Métodos Estatísticos, Universidade Federal do Rio de Janeiro, 2008.

[8] ARTUSO, A. R. *Distribuição gaussiana dos resultados do campeonato brasileiro de futebol: um modelo para estimar classificações em campeonatos de modalidades coletivas*. Revista Brasileira de Ciências do Esporte, v. 30, n. 1, 2008.

[9] LIMA, João Henrique Martins. *Aplicação de machine learning para apostas esportivas: uso de regressão logística, SVM, árvore de decisão e Naive Bayes*. 2022. Trabalho de Conclusão de Curso (Ciências Atuariais) - Universidade Federal de Pernambuco, Recife, 2022. Disponível em: https://repositorio.ufpe.br/handle/123456789/47481. Acesso em: 06/06/2023

[10] PICCIALLI, Veronica; LITI, Chiara; SCIANDRONE, Marco. *Predicting soccer match outcome using machine learning algorithms*. MathSport International 2017 Conference Proceedings, [s. l.], p. 229-237, 2017. Disponível em: https://eprints.kingston.ac.uk/id/eprint/39162/1/MathSport2017Proceedings.pdf#page=234. Acesso em: 6/06/2023.

[11] HORVAT, Tomislav; JOB, Josip. *The use of machine learning in sport outcome prediction: A review*. WIREs data mining and knowledge discovery, [S. l.], v. 12, p., 30 jun. 2020. Disponível em: https://doi.org/10.1002/widm.1380. Acesso em: 6/06/2023.

[12] DUBITZKY, W., LOPES, P., DAVIS, J. et al. *Open International Soccer Database for machine learning*. Mach Learn 108, 9–28 (2019). Disponível em: https://doi.org/10.1007/s10994-018-5726-0. Acesso em: 6/06/2023

[13] BERRAR, D., LOPES, P. & DUBITSKY, W. *Incorporating domain knowledge in machine learning for soccer outcome prediction*. Mach Learn 108, 97–126 (2019). Disponível em: https://doi.org/10.1007/s10994-018-5747-8. Acesso em: 6/06/2023

[14] KAGGLE. *Kaggle*. Disponível em: https://www.kaggle.com/. Acesso em: 14/06/2023

[15] FOOTBALL-DATA.CO.UK. *Football-Data.co.uk*. Disponível em: https://football-data.co.uk/. Acesso em: 14/06/2023

[16] TRANSFERMARKT. *Transfermarkt*. Disponível em: https://www.transfermarkt.com.br/. Acesso em: 14/06/2023

[17] SCIKIT-LEARN. *scikit-learn*. Disponível em: https://scikit-learn.org/stable/. Acesso em: 12/06/2023

[18] CONSTANTINOU, Anthony Costa; FENTON, Norman Elliott. *"Determining the level of ability of football teams by dynamic ratings based on the relative discrepancies in scores between adversaries"* Journal of Quantitative Analysis in Sports, v. 9, no. 1, 2013, p. 37-50. Disponível em: https://doi.org/10.1515/jqas-2012-0036. Acesso em: 14/06/2023

[19] BAIO, Gianluca; BLANGIARDO, Marta. *Bayesian hierarchical model for the prediction of football results*. Journal of Applied Statistics, v. 37, no.2, 2010, p. 253-264, DOI: 10.1080/02664760802684177