DADOS ESTATÍSTICOS NO FUTEBOL: aplicativo para análise estatística no futebol

TEIXEIRA, João Vitor [a]; BAIA, Joás Wesley [b] TREVIZANO, Waldir Andrade[2]; FREITAS, Kenedy Antônio de[2]

[a] Discente do curso de Ciência da Computação – Unifagoc.
[b] Docente Curso de Ciência da Computação - Unifagoc.

joaovit52@hotmail.com
joas.baia@unifagoc.edu.br

RESUMO

Este artigo apresenta o desenvolvimento de um protótipo de software preditivo para resultados de partidas de futebol, aplicando algoritmos de aprendizado de máquina para análise estatística com foco no Campeonato Brasileiro de 2023. O objetivo central é fornecer uma ferramenta inicial e robusta que possa futuramente auxiliar técnicos, analistas e apostadores na previsão de vitórias, empates e derrotas, considerando variáveis como saldo de gols, posição na tabela e desempenho recente dos times. O método de desenvolvimento incluiu a coleta de dados públicos, tratamento e transformação das informações para adequação ao protótipo, e o uso do algoritmo XGBoost, reconhecido pela sua capacidade de capturar padrões complexos em dados desbalanceados. Para aprimorar o desempenho do protótipo, foram implementadas técnicas de otimização de hiperparâmetros e ponderação de classes, além de validação cruzada k-fold para garantir a generalização. As previsões foram avaliadas com métricas de acurácia, precisão, recall e F1-Score, com destaque para as vitórias de mandantes, onde o protótipo demonstrou maior precisão. Apesar dos resultados promissores, o estudo identifica áreas de melhoria no protótipo, especialmente na previsão de empates e vitórias de visitantes. Em versões futuras, o projeto base poderá ser aprimorado com a inclusão de variáveis contextuais, como clima e histórico de lesões, para oferecer uma análise mais abrangente e confiável. Esse trabalho contribui para a aplicação inicial de aprendizado de máquina no esporte, apresentando um recurso prototípico com potencial para evoluir em uma ferramenta robusta de análise de desempenho e tomada de decisões no futebol.

Palavras-chave: Aprendizado de máquina. XGBoost. Previsão de resultados. Ferramenta para analistas esportivos.

INTRODUÇÃO

O futebol é o esporte mais popular no Brasil e em muitos outros países, contando com uma grande base de fãs ao redor do mundo. Estima-se que o futebol reúna bilhões de seguidores, representando uma parcela significativa da população mundial. Grandes eventos, como a Copa do Mundo da FIFA, atraem uma audiência global expressiva, com as finais frequentemente alcançando milhões de espectadores (BBC, 2018). Essa popularidade posiciona o futebol como um motor econômico importante, gerando receitas substanciais por meio de diversas fontes, como bilheterias, salários dos jogadores, contratos comerciais e audiência em múltiplos meios de comunicação (FIFA, 2006).

Com o aumento da competitividade a cada temporada, a busca por informações sobre o desempenho de jogadores e equipes tornou-se essencial para os clubes. A análise estatística, portanto, é uma ferramenta fundamental para comissões

técnicas no futebol, ajudando na conquista de resultados positivos. O uso de dados estatísticos permite identificar padrões, tendências e estratégias que podem ser decisivas em jogos competitivos.

Na prática, comissões técnicas utilizam a estatística para diversas finalidades, como análise de desempenho individual e coletivo, prevenção de lesões e desenvolvimento de táticas. Clubes de elite, como Liverpool e Manchester City, mantêm equipes de analistas de dados que utilizam métricas avançadas para avaliar o desempenho dos jogadores em tempo real, permitindo ajustes táticos durante a partida. Essas análises de desempenho podem incluir dados como passes completados, distância percorrida em campo e número de finalizações, entre outros (Low *et al.*, 2019; Sarmento *et al*., 2018; Coutts, 2014).

Além disso, empresas de apostas esportivas têm se beneficiado amplamente da análise de dados estatísticos. Elas utilizam algoritmos que analisam uma vasta quantidade de dados históricos e atuais para calcular probabilidades e definir "odds" de apostas de maneira mais precisa. Plataformas como Bet365 e Pinnacle utilizam dados detalhados sobre desempenho de equipes e jogadores, histórico de confrontos, condições climáticas e até fatores psicológicos dos atletas para oferecer previsões mais assertivas aos usuários (Labce!, 2024). O futebol gera dados continuamente, a cada partida e movimentação de atletas entre clubes. Essas informações, junto com o valor das transações financeiras envolvidas, tornam os resultados das partidas economicamente e socialmente impactantes. A análise dos dados gerados pelo futebol possui o potencial de influenciar decisões futuras, como a escalação de jogadores, contratações de novos atletas e estratégias de patrocínio (Assunção, 2024; Vendite; Moraes; Vendite, 2015).

Apesar dos benefícios, o acesso aos dados estatísticos no futebol enfrenta desafios, incluindo a dispersão das fontes e a falta de padronização na coleta e disponibilização das informações. A complexidade da análise estatística também pode ser um obstáculo para entusiastas do esporte, que muitas vezes se sentem sobrecarregados pela interpretação desses dados. Além disso, o potencial viés na apresentação dos dados pode levar à promoção seletiva de certas estatísticas, afetando a percepção pública sobre o desempenho no futebol (Capanema, 2024; Datagnostic, 2024).

A questão que se coloca, então, é: como prever o resultado de uma partida de futebol? Este trabalho tem como objetivo geral desenvolver um protótipo de um software para previsões de resultados de partidas de futebol. Os objetivos específicos incluem o estudo do algoritmo de árvore de decisão através do XGBoost, o desenvolvimento de um algoritmo baseado em árvore de decisão e assim o teste e avaliação do modelo obtido.

REFERENCIAL TEÓRICO

A análise de dados no esporte, especialmente no futebol, tem se tornado cada vez mais relevante com o avanço da tecnologia e a disponibilidade de grandes volumes de dados. Essa prática envolve a coleta, processamento e interpretação de dados para auxiliar nas tomadas de decisão, tanto dentro quanto fora do campo. Segundo Davenport (2014), a análise de dados pode fornecer uma vantagem competitiva significativa para as equipes esportivas, permitindo uma melhor

compreensão dos fatores que influenciam o desempenho. No entanto, essa prática não está isenta de críticas e desafios.

A Evolução da Análise Estatística e do Uso de Dados no Esporte

A análise estatística no esporte teve seu início documentado no beisebol, conforme relatam estudos de Sampaio e Janeira (2001), e evoluiu ao longo dos anos para englobar diversas modalidades. A aplicação sistemática de estatísticas para avaliar o desempenho dos atletas e identificar padrões de jogo surgiu em 1936 com Reep e Benjamin. Eles introduziram uma abordagem analítica no futebol ao registrar ações técnico-táticas durante as partidas, como passes, chutes e gols, buscando identificar padrões no jogo. Essa análise pioneira serviu como base para o desenvolvimento de táticas e estratégias fundamentadas em dados no esporte (Reep; Benjamin, 1968).

No futebol, que hoje é o esporte mais popular do mundo em termos de participação e espectadores, a análise estatística evoluiu significativamente. Devido à sua grande audiência e à constante competitividade das partidas, essa modalidade tem sido objeto de investigações científicas que buscam aprimorar o entendimento de suas dinâmicas. Estudos como o de Borrie *et al.* (2002) demonstram como a análise de dados é usada para avaliar tanto o desempenho individual quanto o coletivo, auxiliando os técnicos a tomarem decisões mais precisas. A pesquisa de Carpita *et al.* (2016) ilustra esse avanço ao explorar as ligações entre os resultados dos jogos (vitória, derrota ou empate) e variáveis específicas do jogo, como a criação de oportunidades de gol pela equipe mandante e as ações defensivas de ambas as equipes. Ao examinar quatro temporadas da liga italiana de futebol, Carpita e seus colegas concluíram que as estatísticas, quando aplicadas de forma criteriosa, permitem aos treinadores desenvolver estratégias mais eficazes para atacar e defender, ajustando-as conforme a dinâmica da partida.

Além de contribuir para a compreensão das variáveis que influenciam diretamente os resultados, a análise estatística se consolidou como uma ferramenta essencial para o desenvolvimento de táticas de jogo. Com o avanço da tecnologia, o uso de dados se tornou ainda mais sofisticado, permitindo que equipes como o TSG 1899 Hoffenheim, da Alemanha, criem estratégias altamente personalizadas para cada adversário (Miller, 2019). O clube alemão é conhecido por seu uso extensivo de softwares de análise que possibilitam a identificação de fraquezas nos adversários, oferecendo à equipe técnica dados valiosos que aumentam suas chances de vitória. Esse uso estratégico dos dados é capaz de explorar brechas no esquema tático dos oponentes, promovendo uma preparação mais meticulosa e adaptada a cada partida.

Contudo, o uso intensivo de dados para desenvolver táticas também apresenta desafios e limitações. A implementação de estratégias baseadas em análises detalhadas depende da capacidade dos jogadores de compreender e executar essas orientações, o que pode variar de acordo com o nível de habilidade e a experiência do elenco (Carling; Williams; Reilly, 2008). Além disso, o foco excessivo em dados e métricas pode tornar a equipe previsível para os adversários, levando a uma rigidez tática que impede a flexibilidade necessária durante o jogo.

Assim, embora a análise de dados ofereça uma vantagem competitiva considerável, seu uso deve ser equilibrado com a adaptabilidade e a espontaneidade que o futebol requer, ressaltando a importância de combinar a intuição dos jogadores e técnicos com as informações fornecidas pelas estatísticas (Anderson; Sally, 2013).

Essa evolução no uso da análise estatística e de dados reflete a crescente importância da ciência esportiva na busca pelo sucesso competitivo. A aplicação desses métodos no futebol e em outros esportes destaca como a união de ciência e esporte transforma a maneira como jogos são compreendidos, estratégias são elaboradas e decisões são tomadas, promovendo uma evolução contínua tanto nas dinâmicas esportivas quanto na experiência dos fãs (Patel; Shah; Shah, 2020).

Principais Críticas e Desafios

Entre as principais críticas à análise de dados no futebol está a desumanização do esporte. O futebol é um jogo de paixão, criatividade e imprevisibilidade, características que o tornam único e cativante para torcedores em todo o mundo (Anderson; Sally, 2013) e a análise de dados pode reduzir o jogo a números e estatísticas, perdendo a essência do esporte. Outra crítica é a questão da privacidade dos atletas. A coleta de dados detalhados sobre os jogadores pode levantar preocupações sobre a privacidade e o uso ético dessas informações (Norton *et al*., 2017). Um dos desafios mais significativos é a qualidade dos dados. Dados incompletos, imprecisos ou mal interpretados podem levar a conclusões erradas e decisões inadequadas. Além disso, a integração de dados de diferentes fontes e a necessidade de expertise para analisar esses dados representam obstáculos adicionais. A resistência à mudança por parte de técnicos e jogadores que preferem métodos tradicionais também pode limitar a adoção e a eficácia da análise de dados (Rathke, 2017).

Futuro da Análise de Dados no Futebol

O futuro da análise de dados no futebol promete ser ainda mais inovador com o advento de novas tecnologias como a inteligência artificial (IA) e o aprendizado de máquina. Essas tecnologias têm o potencial de fornecer insights ainda mais profundos e precisos sobre todos os aspectos do jogo. Segundo um estudo de Silva *et al.* (2020), a aplicação de IA pode transformar a análise tática e estratégica, permitindo a simulação de cenários e a criação de modelos preditivos altamente sofisticados. No entanto, a implementação bem-sucedida dessas tecnologias dependerá da superação dos desafios mencionados.

XGboost para Árvore de Decisão

O XGBoost (Extreme Gradient Boosting) é uma implementação otimizada da técnica de boosting, na qual múltiplas árvores de decisão são criadas de forma sequencial, com cada nova árvore tentando corrigir os erros cometidos pelas anteriores. Essa abordagem baseada em gradiente descendente permite que o XGBoost ajuste continuamente suas previsões e minimize os erros residuais a cada iteração, o que aumenta significativamente sua precisão, especialmente em

conjuntos de dados complexos e desbalanceados. Esse processo sequencial contrasta com o método Random Forest, onde as árvores são geradas de forma independente e os resultados são agregados (Breiman, 2001).

Uma das características que diferencia o XGBoost é a introdução de técnicas de regularização. A regularização penaliza a complexidade do modelo, o que ajuda a evitar o overfitting — um problema comum em modelos com grande capacidade de aprendizado. Essas técnicas permitem que o XGBoost equilibre entre a precisão do treinamento e a capacidade de generalização para novos dados, tornando-o altamente eficiente em tarefas preditivas e ideal para competições de machine learning, onde a alta performance é essencial. Além disso, o XGBoost se destaca por sua eficiência computacional. Ele implementa a paralelização de árvores de decisão, acelera o tempo de treinamento e otimiza o uso da memória, o que o torna adequado para grandes volumes de dados e aplicações em larga escala. Suas otimizações internas, como a divisão eficiente dos nós das árvores e o suporte para processamento distribuído, garantem que o XGBoost seja uma escolha preferida para analistas de dados que buscam alta performance e flexibilidade em seus modelos (Geeksforgeeks, 2024; Data Geeks, 2024).

Síntese dos Principais Pontos e Relação com o Estudo

A análise de dados no futebol tem se mostrado vantajosa não apenas para melhorar o desempenho dos jogadores, mas também para fornecer suporte prático a técnicos e analistas. Esse estudo contribui para a aplicação prática do aprendizado de máquina no futebol, com previsões que podem embasar decisões táticas e estratégicas, como a adaptação de estilos de jogo para diferentes adversários. Em comparação com outras abordagens, como regressões logísticas tradicionais e análise de movimento em vídeo, o modelo XGBoost oferece um entendimento detalhado de padrões históricos e desempenhos recentes, ajudando os profissionais do futebol a obterem vantagem competitiva com base em dados concretos. Contudo, enfrenta críticas relacionadas à desumanização do esporte e desafios como a qualidade e interpretação dos dados, além de questões éticas sobre a privacidade dos atletas. Apesar dessas críticas e desafios, o potencial transformador da análise de dados é inegável, especialmente com o avanço das tecnologias de IA e aprendizado de máquina.

Este estudo se propõe a desenvolver um software para previsões de resultados de partidas de futebol, utilizando algoritmos de árvore de decisão. A abordagem busca unir a paixão pelo futebol com a precisão da análise de dados, visando fornecer ferramentas que auxiliem tanto clubes quanto apostadores a tomar decisões mais informadas. Ao abordar tanto os benefícios quanto as limitações da análise de dados, este trabalho pretende

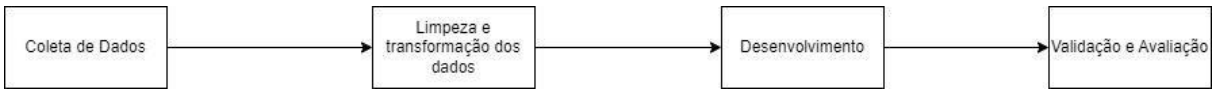
contribuir para uma compreensão mais equilibrada e crítica da aplicação dessa tecnologia no contexto futebolístico.

MÉTODO DE DESENVOLVIMENTO DO TRABALHO

Para desenvolvimento deste protótipo, algumas etapas fundamentais foram realizadas: coleta de dados; limpeza e transformação dos dados; desenvolvimento

do algoritmo de árvore de decisão; validação e avaliação.

Figura 1 – Fluxo de desenvolvimento do projeto



Fonte: elaborada pelo autor.

Coleta de Dados

A análise estatística no futebol para prever resultados futuros será realizada a partir de uma base de dados pública disponível gratuitamente na plataforma Kaggle [11].

Essa base contém informações detalhadas sobre partidas do Campeonato Brasileiro de 2023, incluindo resultados, estatísticas de jogadores e eventos das partidas (posições, pontuações, vitórias, empates, derrotas, etc.). A escolha desse banco de dados foi feita devido à sua relevância e riqueza de informações que cobrem diferentes aspectos críticos de uma partida, proporcionando uma base sólida para a modelagem preditiva.

Limpeza e transformação dos dados

Após a coleta, os dados foram submetidos a um processo rigoroso de limpeza e preparação, garantindo que estivessem adequados para o modelo de aprendizado de máquina. Primeiramente, foi realizada uma verificação criteriosa para identificar e corrigir possíveis erros e inconsistências, como valores ausentes, duplicados ou discrepantes, evitando distorções nos resultados. Para lidar com valores nulos, foi utilizada a técnica de imputação de valores, conforme descrito por Batista e Monard (2003), garantindo que dados incompletos não afetassem negativamente a análise. Em seguida, foi aplicada
normalização z-score, conforme sugerido por Han, Kamber e Pei (2012), padronizando as variáveis numéricas para evitar que magnitudes diferentes influenciassem desproporcionalmente o protótipo. Também houve a conversão de variáveis categóricas, como o mando de campo e o histórico dos últimos jogos, para representações numéricas adequadas, utilizando técnicas como one-hot encoding ou variáveis ordinais. Esse processo foi crucial para garantir que o algoritmo de aprendizado de máquina pudesse processar corretamente todas as informações. A preparação cuidadosa dos dados foi essencial para evitar vieses e garantir a precisão dos modelos, como destacou Miller (2019).

Desenvolvimento do algoritmo XGBoost

O protótipo do modelo preditivo foi desenvolvido utilizando a linguagem Python, com o apoio das bibliotecas Pandas, Scikit-Learn, XGBoost e Numpy, que permitiram a manipulação de dados, a implementação do algoritmo de aprendizado de máquina e a execução de operações matemáticas. O algoritmo selecionado foi o

[1] 1 Disponível em: https://www.kaggle.com/datasets/davidantonioteixeira/brasileiro-2023. Acesso em: 26 fev. 2024

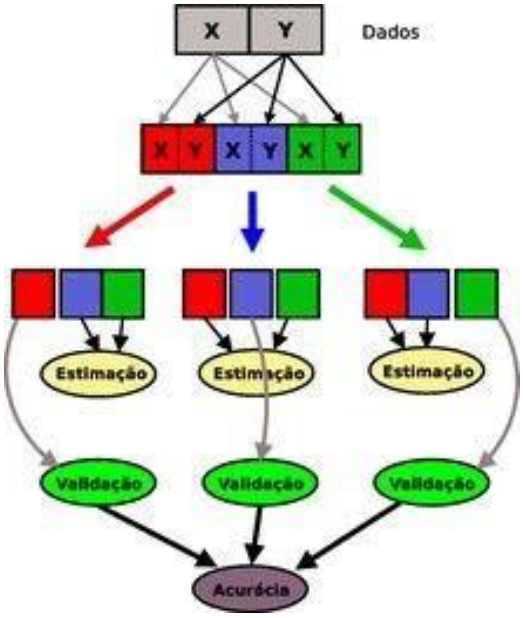
XGBoost, um modelo avançado de árvores de decisão baseado em gradiente. Esse algoritmo foi escolhido por sua robustez, especialmente em lidar com dados desbalanceados, e por sua capacidade de capturar padrões complexos nos dados, o que o tornou altamente eficiente para a tarefa de prever os resultados de partidas de futebol. Outras alternativas, como Random Forest e Redes Neurais, foram consideradas, mas o XGBoost mostrou-se mais adequado para o escopo do projeto devido à sua alta performance em classificações.

Para garantir a eficácia do protótipo, os dados foram divididos em dois conjuntos: 80% para treinamento e 20% para teste, como sugerido por Han, Kamber e Pei (2012). O desbalanceamento natural das classes, devido à maior frequência de vitórias de mandantes, foi tratado utilizando a técnica de ponderação de classes. Pesos foram atribuídos dinamicamente com base na frequência de cada classe no conjunto de treinamento, garantindo que erros em classes minoritárias, como empates e vitórias de visitantes, fossem penalizados proporcionalmente. Essa abordagem proporcionou previsões mais equilibradas entre as diferentes categorias, sem a necessidade de oversampling.

Validação e Avaliação

A Figura 2 ilustra a aplicação da técnica de validação cruzada com k-fold cross- validation, conforme descrito por Kohavi (1995). Esse método divide os dados em k subconjuntos (ou folds), como mostrado na imagem, permitindo treinar e testar o protótipo em diferentes combinações. Esse processo assegura uma avaliação mais robusta, reduzindo o risco de overfitting e fornecendo uma estimativa mais confiável do desempenho do modelo.

Figura 3 – Validação Cruzada com K-fold



Fonte: Wikipedia.

Para a avaliação do protótipo, foram utilizadas métricas como acurácia, precisão, recall e F1-Score. Dentre essas, a F1-Score desempenhou um papel fundamental em cenários de desbalanceamento das classes, como neste caso, em que as vitórias do mandante foram mais frequentes. A inclusão da F1-Score garantiu que o protótipo não fosse avaliado apenas pela acurácia, evitando resultados enviesados por prever majoritariamente a classe mais comum. Essa abordagem promoveu uma avaliação equilibrada do desempenho do modelo em todas as classes.

Por fim, o protótipo passou por uma análise detalhada dos resultados, identificando seus pontos fortes e áreas de melhoria. Foi feita uma comparação com outros modelos de aprendizado de máquina, como Random Forest e Regressão Logística, para validar a eficácia do XGBoost. Além disso, possíveis estratégias de aprimoramento foram exploradas, como o ajuste de hiperparâmetros (ex: learning rate, max_depth, n_estimators) e a inclusão de novasvariáveis preditivas, como dados sobre condições climáticas e lesões de jogadores, para melhorar ainda mais a precisão das previsões.

RESULTADOS E DISCUSSÃO

Os resultados deste protótipo são baseados na aplicação do algoritmo de aprendizado de máquina XGBoost para prever os resultados de partidas de futebol. O protótipo foi desenvolvido para testar a viabilidade de previsões no futebol utilizando aprendizado de máquina, com ênfase em variáveis como saldo de gols, posição na tabela e vitórias consecutivas. A otimização dos hiperparâmetros foi realizada utilizando o RandomizedSearchCV, e os melhores parâmetros encontrados foram aplicados para maximizar o desempenho do protótipo nesta fase inicial.

Para compreender os fatores considerados pelo modelo no processo de previsão dos resultados das partidas, foi elaborado um conjunto de features representativas, descrito na Tabela 1. Essas features foram escolhidas com base na relevância estatística e no impacto potencial no desempenho das equipes, abrangendo aspectos como desempenho acumulado, momentos específicos e tendências recentes.

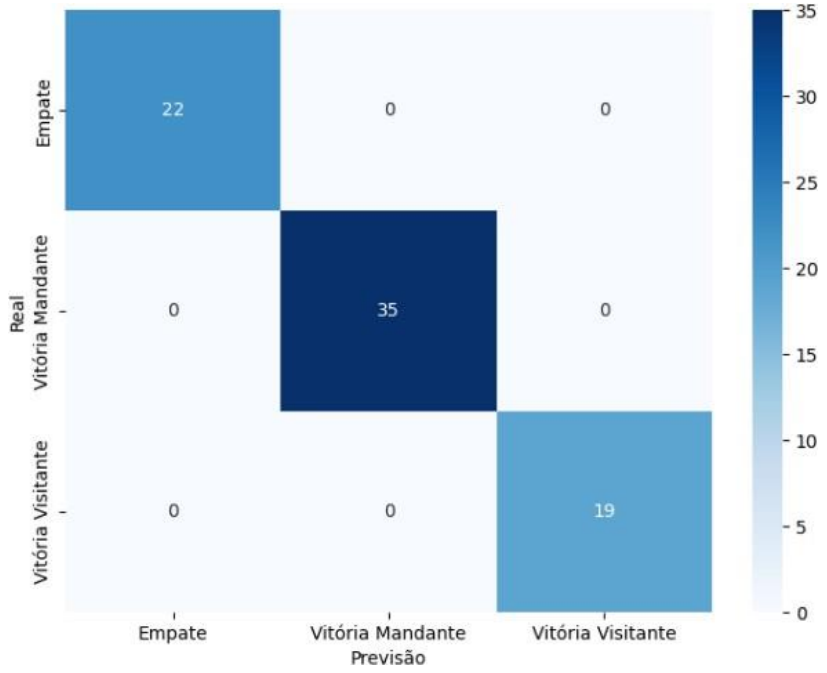
Tabela 1 – Conjunto de Features

Feature	Descrição
Mandante	Representação numérica do time mandante.
Visitante	Representação numérica do time visitante.
Posição	Posição do time na tabela de classificação.
Pontos	Pontos acumulados no campeonato até a rodada atual.
Vitória	Número total de vitórias do time até a rodada atual.
Derrota	Número total de derrotas do time até a rodada atual.
Empate	Número total de empates do time até a rodada atual.
Saldo de Gols	Diferença entre gols marcados e sofridos pelo time.
Vitórias Consecutivas	Número de vitórias consecutivas alcançadas pelo time.
Derrotas Consecutivas	Número de derrotas consecutivas sofridas pelo time.
Jogos Marcando	Número de partidas consecutivas marcando ao menos um gol.
Jogos Sem Sofrer Gols	Número de partidas consecutivas sem sofrer gols.

Fonte: elaborada pelo autor.

A matriz de confusão (Figura 3) demonstra a capacidade do modelo em prever os resultados das partidas. Observa-se alta precisão na previsão de "Vitórias de Mandantes" e "Empates", com 35 e 22 acertos, respectivamente. No entanto, a previsão de 'Vitórias de Visitantes' apresentou maior dificuldade, refletindo o desbalanceamento de classes nos dados utilizados para treinamento. Esse viés indica que o modelo pode estar inclinado a prever as classes mais frequentes, o que sugere a necessidade de ajustes para melhor balanceamento das previsões. Análises detalhadas dos erros por classe ajudarão a identificar ajustes necessários para reduzir o viés e melhorar o desempenho em cenários sub-representados.

Figura 3 - Matriz de Confusão gerada pelo algoritmo



Fonte: elaborada pelo autor.

O F1-Score ponderado de 1.00 obtido pelo protótipo sugere um excelente equilíbrio entre precisão e recall, especialmente na previsão de vitórias dos mandantes. No entanto, esse valor excepcionalmente alto pode ser um indício de *overfitting*, ou seja, o modelo pode estar excessivamente ajustado aos dados de treinamento, comprometendo sua capacidade de generalizar para novos cenários. Para mitigar o problema de *overfitting* e aprimorar a robustez do modelo, podem ser adotadas diversas estratégias. Uma abordagem importante é o ajuste dos hiperparâmetros de regularização, como lambda (L2 regularization) e alpha (L1 regularization), que penalizam a complexidade excessiva do modelo, ajudando-o a generalizar melhor para dados não vistos e reduzindo a sensibilidade a ruídos no conjunto de treinamento. Além disso, limitar a profundidade das árvores de decisão, ajustando o parâmetro max_depth, pode evitar que o modelo capture padrões específicos dos dados de treinamento que dificultem a generalização.

Outra estratégia eficaz é a modificação da taxa de aprendizado (learning_rate), combinada com o aumento do número de estimadores (n_estimators). Essa

combinação permite ao modelo aprender padrões gradualmente, evitando ajustes excessivos durante o treinamento. Para assegurar uma avaliação mais abrangente e precisa do desempenho do modelo, a validação cruzada com múltiplos folds (k-fold cross-validation) pode ser implementada, reduzindo o risco de métricas infladas devido a divisões específicas dos dados. Além disso, o balanceamento das classes pode ser aprimorado. O desbalanceamento natural nas classes, como a maior frequência de vitórias dos mandantes, pode levar a previsões enviesadas. Nesse caso, estratégias como oversampling, undersampling ou técnicas avançadas, como o Synthetic Minority Oversampling Technique (SMOTE), podem melhorar a representatividade das classes minoritárias no modelo.

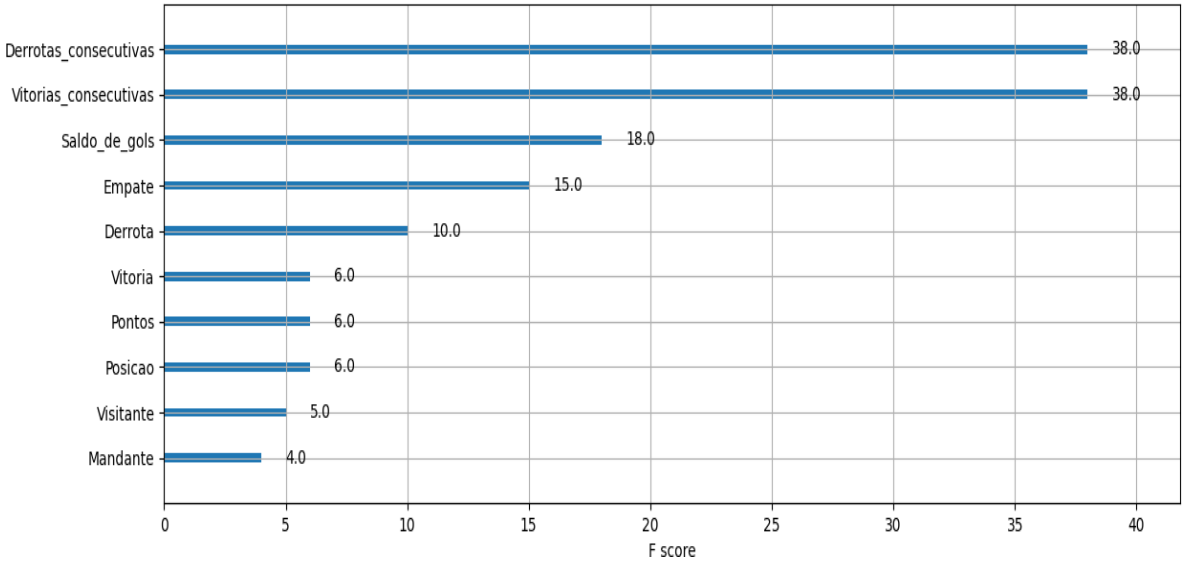
O desempenho do protótipo também pode ser enriquecido pela inclusão de variáveis contextuais, como condições climáticas, histórico de lesões e dados de deslocamento das equipes visitantes, que capturam nuances contextuais relevantes para os resultados das partidas. Essas estratégias, combinadas com ajustes contínuos nos hiperparâmetros, podem aumentar a capacidade do XGBoost de generalizar para cenários reais, reduzindo o risco de overfitting e aprimorando o equilíbrio nas previsões para classes majoritárias e minoritárias. Adicionalmente, explorar a combinação do XGBoost com outros algoritmos, como Random Forest e Redes Neurais, pode oferecer ganhos no desempenho preditivo e na robustez do modelo.

A análise da importância das variáveis (Figura 4) mostrou que as métricas vitórias consecutivas e derrotas consecutivas têm o maior impacto no modelo, ambas alcançando um F-score de 38. Isso evidencia que o histórico recente de desempenho das equipes é essencial para prever os resultados. Além disso, o saldo de gols e a ocorrência de empates também se destacaram, indicando que tanto a capacidade ofensiva quanto o equilíbrio nas partidas são fatores relevantes.

Variáveis como derrotas, vitórias e pontos apresentam importância intermediária, refletindo o desempenho geral das equipes na temporada. Por outro lado, características como posição na tabela, visitante e mandante tiveram menor peso, mas ainda fornecem contribuições importantes para captar nuances do contexto competitivo.

Esses resultados corroboram a literatura existente, que aponta o momento competitivo das equipes como um dos fatores mais preditivos no futebol. Variáveis como séries de derrotas e vitórias consecutivas capturam não apenas o desempenho recente, mas também o impacto psicológico e motivacional das equipes, fatores cruciais para entender os resultados esportivos.

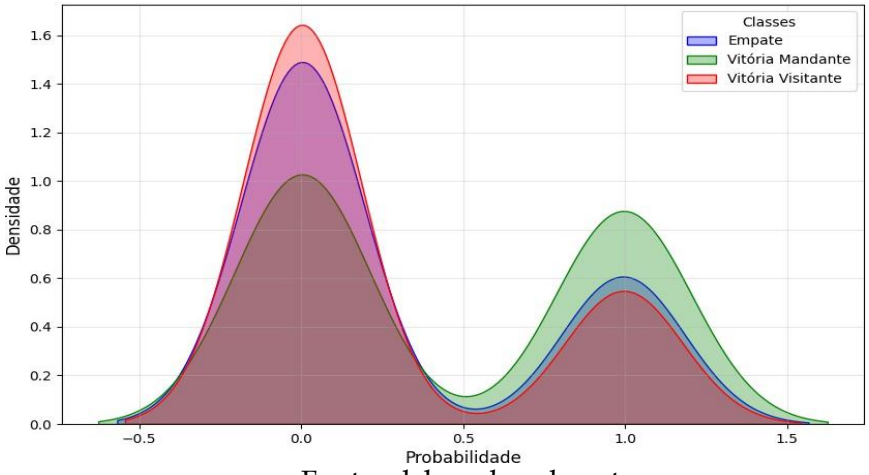
Figura 4 - Gráfico de importância das Variáveis para o protótipo



Fonte: elaborada pelo autor.

A Figura 5 apresenta a distribuição das probabilidades previstas pelo modelo XGBoost para os três possíveis resultados de uma partida: vitória do mandante, empate e vitória do visitante. Observa-se que o modelo demonstra alta confiança ao prever vitórias do time mandante, com uma grande concentração de probabilidades próximas de 1 para essa classe. Esse padrão reflete uma forte tendência do modelo em reconhecer a vantagem histórica dos mandantes no futebol. Por outro lado, as distribuições associadas às classes de empate e vitória do visitante mostram-se mais dispersas e com densidades menos acentuadas. Isso sugere que o modelo apresenta maior incerteza ao prever esses cenários, indicando um nível de confiança reduzido em relação a essas classes. Essa disparidade pode ser resultado de um desequilíbrio no conjunto de dados ou de uma tendência inerente ao futebol, onde os empates e vitórias dos visitantes ocorrem com menor frequência. Portanto, os resultados evidenciam que, embora o modelo consiga capturar bem a predominância de vitórias dos mandantes, há espaço para aprimorar o balanceamento das previsões. Esforços adicionais podem ser direcionados para aumentar a precisão nas classes menos favorecidas, como empates e vitórias dos visitantes, potencialmente por meio de ajustes no treinamento do modelo ou inclusão de novas features que capturem melhor os fatores que influenciam esses resultados.

Figura 5 - Gráfico de Probabilidades gerada pelo protótipo



Fonte: elaborada pelo autor.

A Tabela 2 complementa a análise ao apresentar a comparação entre os resultados reais e os resultados previstos pelo modelo para cinco partidas específicas. Nela, observa-se uma correspondência perfeita entre as previsões do modelo e os resultados ocorridos em campo, com acertos para as categorias "Vitória do Mandante", "Empate" e "Vitória do Visitante". Por exemplo, no jogo entre Palmeiras e Atlético-MG, tanto o resultado real quanto o previsto foram "Vitória Visitante". Essa consistência reforça a confiabilidade do modelo nas situações analisadas e indica que ele está bem ajustado para captar padrões associados aos resultados mais frequentes, como a vitória do mandante. Contudo, a análise anterior sobre as distribuições de probabilidades destaca que há incertezas nas previsões de empates e vitórias dos visitantes, indicando que, embora preciso nestes casos específicos, o modelo pode ser refinado para melhorar sua performance em cenários menos recorrentes.

Tabela 2 – Resultado Real x Resultado Previsto

Resultado Real x Resultado Previsto			
Mandante	**Visitante**	**Resultado Real**	**Resultado Previsto**
Palmeiras	Atletico-MG	Vitória Visitante	Vitória Visitante
Coritiba	Cuiabá	Vitória Visitante	Vitória Visitante
Bahia	Internacional	Vitória Mandante	Vitória Mandante
Corinthians	Fortaleza	Empate	Empate
São Paulo	Internacional	Vitória Mandante	Vitória Mandante

Fonte: elaborada pelo autor.

Embora o protótipo tenha apresentado bons resultados na previsão de vitórias de mandantes, ele enfrenta limitações ao prever empates e vitórias de visitantes devido à ausência de variáveis contextuais. Outros fatores, como condições climáticas, lesões de jogadores-chave e psicologia das equipes visitantes, foram apontados como relevantes na literatura, mas não foram incorporados a este protótipo devido à indisponibilidade desses dados na base utilizada. Estratégias de melhoria incluem a coleta e análise de dados mais abrangentes para essas variáveis e

o desenvolvimento de protótipo com abordagens híbridas, combinando algoritmos baseados em árvores com redes neurais para lidar melhor com a complexidade e variabilidade dos dados de futebol. Entretanto, a performance do protótipo na previsão de empates e vitórias de visitantes mostrou-se inferior, destacando uma área importante para melhorias futuras. Essa dificuldade pode ser atribuída a alguns fatores característicos do futebol que não foram contemplados no protótipo atual. A ausência de variáveis contextuais como condições climáticas e histórico de lesões foi identificada como uma limitação no protótipo desenvolvido. Essas variáveis podem desempenhar um papel significativo na previsão dos resultados das partidas, já que capturam elementos do contexto externo que influenciam diretamente o desempenho das equipes.

Futuramente, a inclusão dessas variáveis poderia ser feita por meio de fontes complementares de dados. Por exemplo, informações sobre o clima podem ser obtidas por meio de APIs meteorológicas, como a OpenWeather, que fornecem dados em tempo real e históricos sobre temperatura, umidade e condições de chuva durante as partidas. Esses fatores são particularmente importantes em jogos ao ar livre, onde condições adversas podem afetar o desempenho físico e técnico dos jogadores. Da mesma forma, o histórico de lesões poderia ser incorporado a partir de bases de dados esportivas especializadas que monitoram a saúde e disponibilidade dos atletas. Dados sobre a frequência e gravidade das lesões, combinados com informações sobre o tempo de recuperação e a forma física dos jogadores, podem ajudar a modelar o impacto da ausência ou limitação de jogadores-chave nas equipes.

Além dessas variáveis, outros fatores como distância de deslocamento das equipes visitantes, histórico de confrontos entre as equipes e estilo tático dos treinadores também poderiam ser adicionados ao modelo. Essas informações poderiam ser integradas por meio de técnicas de feature engineering, utilizando representações numéricas adequadas para variáveis categóricas e contínuas. A inclusão dessas variáveis exigiria a ampliação da base de dados utilizada, possivelmente combinando múltiplas fontes para capturar o cenário completo antes de cada jogo.

A integração de variáveis contextuais ajudaria a enriquecer o modelo, fornecendo uma visão mais abrangente dos fatores que influenciam os resultados das partidas. Isso permitiria ao protótipo gerar previsões mais precisas e adaptadas a cenários específicos, aumentando sua aplicabilidade tanto para clubes quanto para analistas esportivos e apostadores.

Outro ponto relevante é o possível overfitting do protótipo. Embora o F1-Score ponderado tenha sido excepcionalmente alto (1.00), isso pode indicar que o protótipo está excessivamente ajustado aos dados de treino, limitando sua capacidade de generalização para novos cenários (Kohavi, 1995). Para mitigar esse risco, seria interessante explorar estratégias de ajuste de hiperparâmetros, como a redução da profundidade das árvores de decisão ou o aumento da regularização, visando um equilíbrio melhor entre precisão e generalização (Géron, 2019). Além disso, o desbalanceamento das classes é um fator que influencia o desempenho do protótipo.

A maior frequência de vitórias de mandantes no futebol gera um viés natural, que pode comprometer as previsões para empates e vitórias de visitantes. Para tratar esse desbalanceamento, foi utilizada a técnica de ponderação de classes, que atribui

pesos proporcionais às amostras de classes minoritárias durante o treinamento.

Embora essa abordagem tenha melhorado o equilíbrio nas previsões, desafios persistem na captura de características que levam a resultados menos frequentes, como empates e vitórias dos visitantes. Futuras melhorias podem incluir o uso combinado de algoritmos de classificação baseados em ensemble e otimizações específicas para tratar o desbalanceamento residual das classes. Adicionalmente, a inclusão de variáveis contextuais, como histórico tático dos treinadores e condições de jogo, poderia enriquecer o modelo e aumentar sua capacidade de prever cenários sub-representados.

Por fim, é importante reconhecer que o uso de algoritmos de aprendizado de máquina para prever resultados de futebol, embora promissor, enfrenta desafios conceituais. O futebol, por sua natureza, envolve alta dose de imprevisibilidade e fatores subjetivos, como motivação e tomadas de decisão instantâneas, que são difíceis de quantificar e modelar (Norton et al., 2017). Embora o XGBoost tenha se mostrado eficaz na captura de padrões históricos, ele ainda não é capaz de incorporar completamente a complexidade e a aleatoriedade do jogo. Portanto, futuras melhorias no software poderiam incluir a ampliação da base de dados com variáveis contextuais, como clima, lesões e histórico tático de treinadores, além de testes em outras ligas e temporadas, permitindo validar a robustez do protótipo em diferentes contextos (Silva *et al*., 2020). O refinamento do protótipo com o uso de técnicas avançadas de balanceamento de dados e ajustes de hiperparâmetros para evitar o overfitting pode ampliar a capacidade preditiva do software e proporcionar previsões mais precisas, tornando-o uma ferramenta ainda mais valiosa para técnicos, analistas de futebol e apostadores esportivos (Davenport, 2014).

Embora o XGBoost tenha sido o algoritmo escolhido para este protótipo, modelos alternativos foram considerados, como Random Forest e Redes Neurais. O XGBoost foi escolhido devido à sua capacidade de generalizar bem em dados desbalanceados e seu desempenho computacional superior em tarefas de classificação. Comparado ao Random Forest, o XGBoost possui uma implementação de gradiente que otimiza continuamente o desempenho com cada iteração, o que o torna uma escolha mais robusta para este protótipo. Em futuras versões, testes adicionais poderão incluir Random Forest e Redes Neurais como alternativas, explorando a combinação desses modelos para avaliar qual técnica melhor se adapta aos dados do futebol brasileiro.

CONCLUSÃO

Este estudo utilizou uma abordagem rigorosa para o desenvolvimento de um software preditivo de resultados de partidas de futebol, aplicando algoritmos de aprendizado de máquina, com destaque para o XGBoost. A análise detalhada das variáveis, como saldo de gols, desempenho recente das equipes e posição na tabela, demonstrou que tais fatores são determinantes para a previsão de resultados, confirmando sua relevância prática.

Os resultados indicam que o modelo preditivo é eficaz em captar tendências e padrões históricos de vitórias, especialmente para mandantes. Contudo, identificou-se uma área de aprimoramento significativa na previsão de empates e vitórias de visitantes, devido à complexidade e imprevisibilidade natural do futebol. Essa limitação aponta para a necessidade de incluir variáveis adicionais, como clima,

lesões e dados táticos, que podem enriquecer o protótipo e torná-lo mais abrangente.

Este trabalho contribui para o avanço da análise de dados aplicada ao futebol, oferecendo uma ferramenta prática e embasada que pode ser utilizada por técnicos, analistas e até apostadores que busquem uma compreensão mais profunda do esporte. Futuros estudos poderiam explorar a aplicação desse protótipo em diferentes contextos, como outras ligas e temporadas, e a inclusão de dados dinâmicos, para aumentar a capacidade de generalização e precisão das previsões. Ao validar o uso de aprendizado de máquina no contexto esportivo, este estudo estabelece uma base sólida para que novas tecnologias e abordagens possam ser integradas, ampliando as possibilidades de análise e tomada de decisão no futebol.

REFERÊNCIAS

ANDERSON, C.; SALLY, D. **The numbers game**: why everything you know about soccer is wrong. London: Penguin Books, 2013.

ASSUNÇÃO, D. S. A. **Análise e gestão de dados em futebol:** social network analysis. Disponível em: https://comum.rcaap.pt/handle/10400.26/35976. Acesso em: 27 abr. 2024.

BATISTA, G. E. A. P. A.; MONARD, M. C. An analysis of four missing data treatment methods for supervised learning. **Applied Artificial Intelligence**, v. 17, n. 5-6, p. 519-533, 2003.

BBC. **World Cup final 2018**: TV audience of one billion watched France win. Disponível em: https://www.bbc.com/sport/football/44805012. Acesso em: 27 abr. 2024.

BORRIE, A.; JONSSON, G.; MAGNUSSON, M. Temporal pattern analysis and its applicability in sport: an explanation and exemplar data. **Journal of Sports Sciences**, v. 20, n. 10, p. 845– 852, 2002. DOI: 10.1080/026404102320675675.

BREIMAN, L. Random forests. **Machine Learning**, v. 45, n. 1, p. 5-32, 2001. DOI: https://doi.org/10.1023/A:1010933404324.

CAPANEMA, Daniel. **Análise de desempenho e previsões em esportes coletivos**: uma abordagem estatística. 2024. 200 f. Tese (Doutorado em Estatística) – Centro Federal de Educação Tecnológica de Minas Gerais, Belo Horizonte, 2024. Disponível em: https://sig-arquivos.cefetmg.br/arquivos/ 2024175216a39c5223304d644476efec7/DoutoradoFinalDanielCapanema.pdf. Acesso em: 24 ago. 2024.

CARLING, C.; WILLIAMS, A. M.; REILLY, T. **Handbook of soccer match analysis**: a systematic approach to improving performance. New York: Routledge, 2008.

CARPITA, M.; SANDRI, M.; SIMONETTO, A.; ZUCCOLOTTO, P. Discovering the drivers of football match outcomes with data mining. **Quality Technology and Quantitative Management**, v. 12, n. 4, p. 561–577, 2016. DOI: 10.1080/16843703.2015.11673436.

COUTTS, A. Evolution of football match analysis research. **Journal of Sports Sciences**, v. 32, p. 1829-1830, 2014. Disponível em: https://doi.org/10.1080/02640414.2014.985450. Acesso em: 18 nov. 2024.

DATAGNOSTIC. **Sobre o uso de análise de dados no futebol –** Parte 1. 2 fev. 2024. Disponível em: https://datagnostic.com.br/2024/02/02/sobre-o-uso-de-analise-de-dados-no-futebol- parte-1/. Acesso em: 24 ago. 2024.

DAVENPORT, T. H.. **Big data at work:** dispelling the myths, uncovering the opportunities. Boston: Harvard Business Review Press, 2014.

FIFA. **FIFA Big Count 2006**: 270 million people active in football. Disponível em: https://digitalhub.fifa.com/m/eda1e4e4c1003f7/original/xpbl3kvgkjgvjydc5w2w-pdf.pdf. Acesso em: 27 abr. 2024.

GABBETT, T. J. The training—injury prevention paradox: should athletes be training smarter and harder? **British Journal of Sports Medicine**, v. 50, n. 5, p. 273-280, 2016.

GEEKSFORGEEKS. **Difference between Random Forest and XGBoost**. Disponível em: https://www.geeksforgeeks.org/difference-between-random-forest-vs-xgboost/. Acesso em: 15 out. 2024.

GÉRON, Aurélien. **Mãos à obra:** aprendizado de máquina com Scikit-Learn & TensorFlow. 1. ed. Rio de Janeiro: Alta Books, 2019.

HAN, J.; KAMBER, M.; PEI, J. **Data mining**: concepts and techniques. 3. ed. San Francisco: Morgan Kaufmann, 2012.

KOHAVI, R. **A study of cross-validation and bootstrap for accuracy estimation and model selection**. *In:* IJCAI, v. 14, n. 2, p. 1137-1145, 1995.

LOW, B. *et al*. A systematic review of collective tactical behaviors in football using positional data. **Sports Medicine**, v. 50, p. 343-385, 2019. Disponível em: https://doi.org/10.1007/s40279-019-01194-7. Acesso em: 18 nov. 2024.

MILLER, R. The use of analytics in professional soccer: an overview. **Soccer & Society**, v. 20, n. 4, p. 467-487, 2019.

MONARD, Maria Carolina; BARANAUSKAS, José Augusto. Indução de regras e árvores de decisão. **Sistemas inteligentes: fundamentos e aplicações**, v. 1, p. 115-139, 2003.

NIELSEN, J. **Usability engineering**. San Francisco: Morgan Kaufmann, 1994.

NORTON, T.; NORTON, C.; SADLER, A. Data Protection and Privacy Issues in Sport. Journal of Legal Aspects of Sport, v. 27, n. 1, p. 32-50, 2017.

PATEL, D.; SHAH, D.; SHAH, M. The intertwine of brain and body: a quantitative analysis on how big data influences the system of sports. **Annals of Data Science**, v. 7, p. 1–16, 2020. DOI: 10.1007/s40745-019-00239-y.

POWERS, D. M. W. Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation. **Journal of Machine Learning Technologies**, v. 2, n. 1, p. 37-63, 2011.

QUINLAN, J. R. **C4.5:** programs for machine learning. San Francisco: Morgan Kaufmann, 1993.

REEP, C.; BENJAMIN, B. Skill and chance in association football. **Journal of the Royal Statistical Society**. Series A (General), v. 131, n. 4, p. 581–585, 1968. DOI: https://doi.org/10.2307/2343728.

SMITH, A. Data science and sports betting: an analysis of the industry's leading practices. **Journal of Gambling Studies**, v. 36, n. 1, p. 53-70, 2020.

VENDITE, Laercio Luiz; MORAES, Antonio Carlos de; VENDITE, Carolina Coluccio. Scout no futebol: uma análise estatística. **Conexões**, Campinas, SP, v. 1, n. 2, p. 183–194, 2015. DOI: 10.20396/conex.v1i2.8638024. Disponível em: https://periodicos.sbu.unicamp.br/ojs/index.php/ conexoes/article/view/8638024. Acesso em: 27 abr. 2024.