# Avaliação de Modelos Preditivos em Estatísticas Esportivas Ao Vivo: Uma Abordagem de Dados em Tempo Real

**Nathan C. de M. Gomes[1], Yuri de A. M. Barbosa [1], Thaís G. do Rego[1]**

[1]Centro de Informática – Universidade Federal da Paraíba (UFPB)

João Pessoa – PB – Brazil

nathan.gomes@estudantes.ufpb.br, yuri@ci.ufpb.br, gaudenciothais@gmail.com

***Abstract.*** *Sports prediction has expanded remarkably with the advancement of technology and artificial intelligence, which encourages the search for more refined methods of data collection and pattern recognition. This study focuses on the analysis of models and techniques such as Neural Networks and AutoML to understand the occurrence of events in live matches, adopting a large-scale data approach. It was observed that feature engineering and defining time windows before events are crucial for model performance.*

***Resumo.*** *A previsão esportiva tem se expandido notavelmente com o avanço da tecnologia e da inteligência artificial, o que estimula a busca por métodos mais refinados de coleta de dados e reconhecimento de padrões. Este estudo foca na análise de modelos e técnicas como Redes Neurais e AutoML para compreender a ocorrência de eventos em partidas ao vivo, adotando uma abordagem de dados em larga escala. Observou-se que a engenharia de atributos e a definição de janelas de tempo antes dos eventos são cruciais para o desempenho do modelo.*

**Catalogação na publicação**
**Seção de Catalogação e Classificação**

G633a Gomes, Nathan Carlos de Macena.
Avaliação de modelos preditivos em estatísticas esportivas ao vivo: uma abordagem de dados em tempo real / Nathan Carlos de Macena Gomes. - João Pessoa, 2024.
31 f. : il.

Orientação: Yuri de Almeida Malheiros Barbosa.
Coorientação: Thais Gaudencio do Rego.
TCC (Graduação) - UFPB/CI.

1. Previsão esportiva. 2. Inteligência artificial. 3. Análise de dados. 4. Redes Neurais. 5. AutoML. 6. Reconhecimento de padrões. 7. Engenharia de atributos. 8. Análise em tempo real. 9. Modelagem preditiva. 10. Análise de larga escala. I. Barbosa, Yuri de Almeida Malheiros. II. Rego, Thais Gaudencio do. III. Título.

UFPB/CI CDU 004.8

Elaborado por Michelle de Kássia Fonseca Barbosa - CRB-738

# 1. Introdução

A obtenção de informações em jogos esportivos para análise e previsões começou de maneira rudimentar e se desenvolveu significativamente ao longo do tempo com o avanço da tecnologia. Inicialmente, as informações eram coletadas manualmente por observadores que assistiam aos jogos e anotavam estatísticas importantes como gols, corridas, passes e outros dados relevantes para o esporte em questão. Esse processo era demorado e sujeito a erros humanos, mas era a única maneira de obter dados para análises e previsões.

A evolução da análise de dados nos esportes começou com abordagens simples, como no beisebol, onde estatísticas básicas eram rastreadas desde os primórdios do jogo. A modernização dessa análise começou na segunda metade do século XX, marcada pela publicação de "Percentage Baseball" de Earnshaw Cook e pela Society for American Baseball Research (SABR). Pioneiros como Davey Johnson usaram simulações de beisebol e programas da IBM para melhorar o gerenciamento e a estratégia do jogo. Billy Beane, com sua teoria Moneyball, desempenhou um papel crucial ao adotar uma abordagem baseada em evidências na seleção de jogadores, o que atraiu a atenção de outras equipes da MLB para a análise de dados (Education Dynamics, 2024).

Globalmente, a análise de dados no esporte se tornou um negócio significativo, com o mercado de análise de desempenho esportivo projetado para ultrapassar US$5,2 bilhões até 2024. Exemplos notáveis incluem o Moneyball de Billy Beane e a abordagem inovadora de Matthew Benham no Brentford Football Club, que demonstram como a análise estatística pode transformar equipes esportivas (Pykes, 2022).

A análise de dados se expandiu para outros esportes, adaptando-se às necessidades específicas de cada um. No futebol, por exemplo, os clubes investem em ciência de dados e tecnologia para melhorar o desempenho em campo e as decisões fora dele, monitorando dados como posicionamento em jogo, fadiga durante o treinamento e distâncias percorridas. O basquete também abraçou a análise de dados, com equipes da NBA utilizando câmeras de rastreamento de dados para monitorar cada movimento dos jogadores em quadra, enriquecendo as estatísticas e a análise do desempenho dos jogadores (Education Dynamics, 2024)

Nesse sentido, o objetivo principal deste estudo é construir uma base de dados robusta a partir de estatísticas esportivas oficiais ao vivo e analisar modelos e técnicas de aprendizagem de máquina para realizar predições de um evento na partida, de modo que com a ajuda das métricas de desempenho, essas previsões possam ser confiáveis. Serão desenvolvidos modelos de aprendizagem de máquina, treinados e testados para a previsão de um evento na partida. A busca pelos melhores parâmetros na Rede Neural foi realizada por meio de busca aleatória em um espaço abrangente de parâmetros, enquanto que, com o AutoML, os melhores parâmetros e modelos são automaticamente sugeridos. Este estudo visa a identificar padrões nas estatísticas das partidas com o auxílio dos modelos de aprendizagem, de modo que seja possível fazer uma previsão baseada nos dados históricos das estatísticas das partidas anteriores, prevendo como por exemplo o evento do primeiro gol na partida.

## 2. Trabalhos Relacionados

A exploração de técnicas avançadas de aprendizado de máquina, para a análise e previsão de eventos em partidas de esportes ao vivo, tem adquirido um destaque considerável, refletindo os avanços tecnológicos e a crescente disponibilidade de dados detalhados de jogos. Neste contexto, identificamos três estudos significativos que oferecem *insights* valiosos para o desenvolvimento do presente trabalho, abordando desde a validação de dados em tempo real, até a previsão de resultados utilizando modelos computacionais sofisticados.

O primeiro estudo conduzido por Gong et al (2019), foca na investigação da validade e confiabilidade das estatísticas de partidas de futebol ao vivo. Utilizando variáveis de desempenho coletadas pelo sistema Champdas Master, o estudo emprega análises estatísticas rigorosas para validar a qualidade dos dados capturados em tempo real, estabelecendo uma base sólida para a aplicação de técnicas preditivas em dados esportivos.

Para validar as variáveis de desempenho usadas pelo sistema, um painel de 20 treinadores de futebol profissional participou voluntariamente da validação. Quatro operadores bem treinados, divididos em dois grupos, analisaram de forma independente uma partida da La Liga espanhola. A validação dos indicadores foi avaliada por meio do coeficiente de Aiken, resultando em médias de 0,84 ± 0,03 e 0,85 ± 0,03 para a validação dos indicadores, indicando uma alta confiabilidade intra-operador. Diversas medidas estatísticas, incluindo o coeficiente Kappa, coeficientes de correlação intraclasse (ICC) e erros típicos (TE), foram utilizadas para determinar a confiabilidade das estatísticas coletadas ao vivo, demonstrando sua eficiência tanto intra-operador, quanto entre operadores diferentes. O estudo concluiu que o sistema Champdas Master pode ser usado de maneira válida e confiável para coletar estatísticas de partidas de futebol ao vivo por operadores bem treinados, fornecendo dados confiáveis para treinadores, gerentes, pesquisadores e analistas de desempenho em suas tarefas profissionais e investigações.

Em seguida, em um outro estudo elaborado por Lunelli (2019) é apresentada uma abordagem inovadora para a previsão de resultados em jogos da NBA, explorando uma vasta gama de dados estatísticos e aplicando múltiplas técnicas de aprendizado de máquina, incluindo árvores de decisão e redes neurais. Este estudo destaca a capacidade de modelos computacionais em capturar padrões complexos e oferecer previsões precisas, servindo de referência para a construção de modelos preditivos em diferentes contextos esportivos.

A metodologia adotada é dividida em cinco etapas principais: construção e coleta da base de dados, pré-processamento dos dados, engenharia de atributos, modelagem preditiva e avaliação dos modelos. Dentre os algoritmos testados, destacam-se árvores de decisão, regressão logística, redes neurais, e *ensembles*, visando sempre maximizar a precisão das previsões.

Uma parte significativa do trabalho concentra-se na análise e tratamento dos dados, fundamentais para o sucesso do modelo. Isso inclui a transformação e normalização das variáveis, além da seleção criteriosa das mesmas para inclusão no modelo. A avaliação do modelo é realizada por meio de métricas como a precisão e a aplicação de estratégias de validação cruzada para garantir a robustez e confiabilidade das previsões.

Após a modelagem e avaliação, o estudo avança para a aplicação prática das previsões em um contexto de apostas esportivas, buscando criar estratégias lucrativas baseadas nos resultados do modelo. Esta aplicação prática destaca o potencial do uso de algoritmos de aprendizagem de máquina não apenas como ferramentas analíticas, mas também como suporte para decisões em contextos de alto risco financeiro.

Em suma, o trabalho de Lunelli (2019) representa um avanço significativo na aplicação de técnicas de aprendizagem de máquina para a previsão de resultados esportivos, oferecendo interpretações valiosas para apostadores, treinadores e analistas de desempenho. Através de uma metodologia rigorosa e uma análise detalhada, este trabalho contribui para a literatura existente, ao mesmo tempo em que abre caminho para futuras investigações na interseção entre ciência de dados e esportes.

Leandro Stival (2022) apresenta uma abordagem inovadora para entender melhor como os eventos em partidas de futebol podem ser previstos utilizando dados e técnicas de aprendizado de máquina. Este estudo foca na possibilidade de prever se uma posse de bola resultará em uma jogada perigosa perto da grande área adversária com base nos primeiros segundos da posse de bola. A análise se baseia em dados detalhados de partidas, incluindo a posição dos jogadores e métricas derivadas da teoria de grafos e redes complexas, convertendo estas informações em imagens de forma que seja possível a extração de características por meio de redes neurais convolucionais profundas.

A pesquisa foca na criação de modelos preditivos capazes de antever se uma posse de bola levará a equipe a alcançar a zona de finalização com base nos primeiros segundos da posse de bola. Para isso, foram selecionados e analisados dados de partidas, transformando o posicionamento dos jogadores em séries temporais de grafos. A partir destes, métricas específicas como centralidade, excentricidade, eficiência global e local, vulnerabilidade, coeficiente de clusterização, entropia e *pagerank* foram calculadas e utilizadas para alimentar as redes neurais em um processo de aprendizado profundo.

O estudo emprega a técnica de transferência de aprendizado para refinar um modelo pré-treinado (EfficientNet B0), com o objetivo de extrair características relevantes das imagens de ritmo visual geradas a partir das métricas de grafos. O modelo ajustado é, então, aplicado na classificação das posses de bola, determinando sua potencialidade em resultar em uma jogada ofensiva significativa.

A validação dos modelos empregou métricas de acurácia balanceada, eficiência local, excentricidade e entropia. Além disso, uma análise detalhada da contribuição das características permitiu identificar quais métricas possuem maior influência sobre os resultados preditivos, oferecendo dados para o entendimento das dinâmicas que levam à criação de oportunidades de gol.

Estes trabalhos correlatos e o estudo atual são mostrados na Tabela 1, que demonstram a diversidade de abordagens e técnicas empregadas na análise de dados esportivos, desde a validação da confiabilidade dos dados, até a implementação de modelos preditivos de eventos relacionados a partidas ao vivo. As metodologias e descobertas desses estudos fornecem uma base sólida para a presente pesquisa, inspirando a exploração de novas técnicas e abordagens no campo da previsão esportiva. A integração dessas perspectivas e a aplicação de conhecimentos derivados destas investigações contribuirão significativamente para o avanço do estudo em questão, ampliando o entendimento sobre a aplicação de tecnologias de aprendizado de máquina na previsão de eventos em jogos esportivos ao vivo.

**Tabela 1. Comparação do estudo atual com trabalhos anteriores**

| Título, ano, quem e onde | Objetivo | Variáveis usadas nos dados | Dados | Métricas utilizadas |
|---|---|---|---|---|
| Gong, B., Cui, Y., Gai, Y., Yi, Q., e Gómez (2019) | Investigar a validade das variáveis de partidas de futebol e a confiabilidade do sistema Champdas Master, usado por operadores treinados em partidas de futebol ao vivo. | Desempenho relacionado ao ataque, passe e defesa/goleiro. | Dados coletados ao vivo pelo próprio sistema Champdas Master que combina teclas de atalho de teclado e posicionamento na tela para representar eventos e rótulos | Coeficiente de Correlação Intraclasse (ICC) e índice Kappa de Cohen |
| Lunelli (2019) | Desenvolver modelos de machine learning para prever resultados dos jogos da NBA e criar um sistema de apostas simples com base nos resultados do modelo. | Pontuação, jogadas, e rastreamento dos jogadores e da bola, arremessos de quadra, rebotes, faltas, e lances livres. | Dados coletados principalmente do site Basketball-Reference.com, além de outras fontes estatísticas da NBA. | Precisão (Accuracy), Retorno sobre o Investimento (ROI), Retorno Financeiro Calculado através do ROI |
| Stival, L. (2022) | Investigar se um modelo baseado nos primeiros 5 segundos de posse de bola pode prever se a bola chegará no quarto final do campo e identificar um conjunto reduzido de características mais importantes para essa previsão. | Área de redes complexas, como centralidade e entropia, obtidas de grafos criados a partir dos dados dos jogos. | Dados de 10 jogos completos, incluindo a posição de todos os jogadores a cada 30 frames por segundo. Eventos como faltas, gols e saídas pelas laterais também foram rotulados nos dados. | Acurácia Balanceada, Eficiência Local, Excentricidade, Entropia |
| Gomes, N (2024) | Avaliar modelos preditivos em janelas de tempo de eventos esportivos ao vivo. | Desempenho relacionado ao ataque e defesa, estatística comuns como posse de bola, escanteios, etc. E desempenho histórico das equipes. | Dados coletados da plataforma optaplayerstats. | Precisão, Recall, F1-score, Matriz de confusão |

A Tabela 1 compara o estudo atual com trabalhos anteriores, destacando tanto diferenças quanto semelhanças. Uma diferença notável é o enfoque metodológico: enquanto o trabalho atual emprega uma combinação de redes neurais e AutoML para analisar eventos esportivos em tempo real. Além disso, *TPOT* foi a biblioteca de AutoML utilizada para este trabalho onde foi integrada com técnicas de otimização Bayesiana para melhorar ainda mais sua eficiência e eficácia em espaços de hiperparâmetros contínuos, demonstrando sua capacidade de adaptar-se a limitações computacionais e explorar eficientemente o espaço de pesquisa (Kenny et al., 2023). Os estudos relacionados concentram-se em métodos específicos, como validação de dados em tempo real ou uso de múltiplas técnicas de machine learning para previsões em jogos da NBA e análise preditiva em futebol. Em termos de semelhanças, todos os estudos buscam aprimorar a precisão das previsões em eventos esportivos utilizando dados ao vivo e técnicas avançadas de aprendizado de máquina. Além disso, há uma preocupação comum em validar a confiabilidade dos dados e otimizar parâmetros para melhorar os resultados dos modelos. O presente trabalho se distingue pela sua abordagem integrada que utiliza uma gama mais ampla de dados e técnicas analíticas, visando capturar a complexidade e a dinâmica dos eventos esportivos em tempo real.

# 3. Metodologia

A metodologia é dividida em várias seções essenciais, cada uma projetada para abordar diferentes aspectos da coleta, processamento e análise dos dados. A abordagem começa com a coleta de dados, utilizando a plataforma optaplayerstats onde, por meio de técnicas avançadas de web scraping com Selenium e BeautifulSoup, dados são extraídos em tempo real durante os jogos e organizados em DataFrames para facilitar a manipulação e análise subsequente.

O processo de pré-processamento dos dados é crucial para garantir a qualidade e usabilidade dos mesmos. Esta fase envolve a limpeza de dados, removendo duplicatas e valores nulos, conversão de tempos para formatos decimais, e a padronização dos nomes das ligas, garantindo consistência e precisão na análise. Além disso, é feita a seleção de eventos significativos, como o primeiro gol da partida, e a aplicação de técnicas de balanceamento para manter a equidade entre os dados e minimizar vieses.

A seção subsequente descreve a separação dos dados em conjuntos de treinamento e teste, realizada de forma estratificada para manter uma proporção constante das classes na variável alvo. Segue-se uma análise da importância dos atributos, utilizando o modelo RandomForestClassifier para determinar quais características influenciam mais significativamente o desempenho das equipes. Esta análise ajuda a identificar quais atributos são mais valiosos e quais podem ser descartados.

Após a avaliação, novos atributos são introduzidos e características menos importantes são eliminadas, otimizando o conjunto de dados para a modelagem preditiva. Esta fase é essencial para refinar o modelo, aumentando sua precisão e eficácia. A construção dos modelos de aprendizagem é o último passo metodológico, envolvendo o desenvolvimento e treinamento de modelos avançados, como redes neurais e AutoML. Esta etapa inclui a seleção de hiperparâmetros adequados e a implementação de técnicas de aprendizado profundo para melhorar a classificação e previsão dos resultados.

Cada etapa da metodologia é projetada para construir sobre a anterior, assegurando uma análise compreensiva baseada em dados robustos e bem processados. Essa abordagem metódica não só facilita uma compreensão aprofundada das tendências no futebol, mas também permite fazer previsões precisas e fornecer interpretações valiosas.

## 3.1. Base de Dados

Os dados empregados nesta pesquisa foram construídos e extraídos da *optaplayerstats*[1], que é uma plataforma dedicada a fornecer resultados públicos ao vivo de futebol, das principais competições ao redor do mundo. Está associada à *Stats Perform*, uma empresa conhecida pela aplicação de inteligência artificial no esporte, oferecendo uma gama de serviços de dados e análises esportivas. Os dados empregados são armazenados no *javascript* da página de maneira temporal à medida que o jogo ocorre. Todo o experimento foi feito em *python* para que estas informações fossem coletadas, de maneira sequencial, utilizando técnicas de *web scraping*.

O *script* emprega bibliotecas avançadas como *Selenium*[2] e *BeautifulSoup*[3] para automatizar um navegador, com o objetivo de acessar e extrair informações detalhadas

[1] https://optaplayerstats.statsperform.com
[2] https://selenium-python.readthedocs.io
[3] https://beautiful-soup-4.readthedocs.io/en/latest

de partidas de futebol específicas. Essa automação permite a coleta de dados estruturados, que são organizados e salvos. Após a coleta, as informações são estruturadas em linhas e colunas, proporcionando uma maneira eficiente de manipular e organizar os dados antes de salvá-los. Esse método facilita análises posteriores sobre o desempenho das equipes e tendências nas ligas relacionadas ao comportamento das partidas, como a prevalência de gols no primeiro tempo, incidência de faltas, e a frequência de cartões amarelos ou vermelhos, em momentos específicos do jogo. Adicionalmente, para otimizar a velocidade de coleta de dados, foram utilizadas dez instâncias *Docker* em conjunto com o *Selenium Grid*. Cada instância foi ativada em um intervalo de data que é disponibilizado na plataforma naquele dia, maximizando a eficiência na obtenção dos dados. Este enfoque não só agiliza o processo de coleta, mas também garante uma cobertura ampla e detalhada das partidas analisadas.

Os dados coletados contam com mais de 1.3 milhões de registros com 37 atributos, minuto a minuto de jogo, com cerca de 7891 partidas e 63 ligas de todo o mundo. Na Europa, temos ligas e torneios como a UEFA Champions League, UEFA Europa League, UEFA Europa Conference League, e UEFA Super Cup, junto com as competições nacionais como a Premier League da Inglaterra, La Liga da Espanha, Bundesliga da Alemanha, Série A da Itália, e Ligue 1 da França. Na América do Sul, foram coletadas informações sobre competições como a CONMEBOL Libertadores, CONMEBOL Sudamericana, e a Copa América. Enquanto isso, a CONCACAF Gold Cup e a Major League Soccer (MLS) ressaltam o futebol na América do Norte e Central. No cenário asiático, a AFC Champions League e a AFC Asian Cup destacam-se, enquanto a África tem a CAF Africa Cup of Nations. Na Austrália, temos a A-League Men, e competições globais como a FIFA World Cup e a FIFA Women's World Cup, que unem nações no futebol mundial. Além disso, competições nacionais específicas como a Copa do Brasil, Liga MX do México, e a J1 League do Japão, juntamente com torneios específicos de cada país como a FA Cup, Copa del Rey, e DFB Pokal), foram analisadas. A Tabela 2 apresenta os atributos nessa fase inicial.

**Tabela 2. Descrição dos atributos iniciais.**

| **Atributo** | **Descrição** |
|---|---|
| minute | Minuto atual do jogo |
| homeTeam | Nome do time da casa |
| awayTeam | Nome do time visitante |
| shots_home | Número de chutes do time da casa |
| shots_away | Número de chutes do time visitante |
| league | Liga em que o jogo está sendo disputado |
| corners_home | Escanteios a favor do time da casa |
| corners_away | Escanteios a favor do time visitante |
| shotsOffgoal_home | Chutes para fora do time da casa |
| shotsOffgoal_away | Chutes para fora do time visitante |
| fouls_home | Faltas cometidas pelo time da casa |
| fouls_away | Faltas cometidas pelo time visitante |
| tackles_home | Desarmes realizados pelo time da casa |
| tackles_away | Desarmes realizados pelo time visitante |
| result | Resultado do evento |
| match_id | Identificador único do jogo |
| possessiontime_away | Tempo de posse de bola do time visitante |
| possessiontime_home | Tempo de posse de bola do time da casa |
| shotsOffgoal_home | Chutes para fora do time da casa |
| shotsOffgoal_away | Chutes para fora do time visitante |
| shotsOngoal_home | Chutes no gol do time da casa |
| shotsOngoal_away | Chutes no gol do time visitante |
| yellowcards_home | Cartões amarelos para o time da casa |
| yellowcards_away | Cartões amarelos para o time visitante |
| passes_home | Passes completados pelo time da casa |
| passes_away | Passes completados pelo time visitante |
| fouls_c_home | Faltas cometidas pelo time da casa |
| fouls_c_away | Faltas cometidas pelo time visitante |
| fouls_won_home | Faltas sofridas pelo time da casa |
| fouls_won_away | Faltas sofridas pelo time visitante |
| offsides_home | Impedimentos do time da casa |
| offsides_away | Impedimentos do time visitante |
| tackles_home | Desarmes realizados pelo time da casa |
| tackles_away | Desarmes realizados pelo time visitante |
| result | Resultado do evento (preenchida com zeros inicialmente) |
| match_id | Identificador único do jogo |

## 3.2. Pré-processamento

O processo de pré-processamento dos dados envolveu várias etapas cruciais para limpar e balancear o conjunto de dados, assegurando a qualidade e a utilidade para análises subsequentes. Inicialmente, a base foi carregada e as propriedades básicas dos dados, como o intervalo de datas, foram examinadas. A remoção de registros duplicados e a limpeza de valores nulos foram realizadas para garantir a integridade dos dados. Além disso, foi implementada uma conversão de tempo para transformar os minutos em um formato decimal, facilitando análises futuras. Após essa conversão, os minutos iguais a zero foram removidos, e o conjunto de dados foi mais uma vez filtrado para remover duplicatas e valores nulos.

Para aprimorar a análise, os nomes das ligas foram padronizados através de um mapeamento específico, garantindo uniformidade e facilitando comparações. A padronização dos nomes das ligas é crucial para facilitar a análise dos dados. Diferentes fontes podem referenciar a mesma liga de maneiras variadas, levando a duplicações ou erros na compilação dos resultados. Por exemplo, a Premier League pode ser referida como "English Premier League", "EPL", ou simplesmente "Premier League". A padronização assegura que todas as referências a uma liga sejam unificadas sob um único nome padrão, facilitando comparações precisas e análises agregadas, além de melhorar a qualidade dos resultados gerados a partir dos dados. A seleção de ligas baseou-se em critérios, como o número mínimo de partidas, definido como no mínimo 100 partidas por liga, para garantir uma boa diversidade de dados.

O processo de balanceamento de dados garante que os resultados dos jogos sejam distribuídos de maneira uniforme. Utilizou-se uma técnica de balanceamento para assegurar uma representação proporcional dos diferentes resultados dos jogos. Essa técnica envolveu equilibrar as classes preditivas 0 e 1: o valor 1 indica a ocorrência de um evento chave, como o primeiro gol do jogo, enquanto o valor 0 indica jogos onde esse evento não ocorreu no período analisado. Entretanto, se o evento ocorre durante o jogo, a variável alvo dos registros anteriores é atualizada para 1, refletindo a ocorrência do evento, enquanto os demais registros permanecem com o valor 0. Isso pode levar a um desequilíbrio, pois, dependendo do intervalo de tempo escolhido, pode haver menos registros com a variável alvo igual a 1. Para equilibrar as classes, para cada estatística relacionada à classe 1, buscou-se uma estatística correspondente na classe 0 que ocorresse no mesmo intervalo de tempo do jogo. Essa abordagem minimiza possíveis vieses nos dados e é essencial não apenas para neutralizar desequilíbrios, mas também para garantir a integridade e confiabilidade das análises futuras.

Durante a etapa de pré-processamento dos dados, também foi feita a seleção do evento a ser analisado. O foco recaiu sobre o evento significativo de marcar o primeiro gol durante a partida. Para capturar esse momento, implementou-se um filtro específico: sempre que o primeiro gol é marcado, independente do time, a variável alvo, denominada 'result', recebe o valor 1 em uma janela de tempo específica. Essa abordagem permite uma identificação clara do desempenho e precisa da ocorrência do evento, facilitando análises subsequentes, que buscam compreender os fatores que contribuem para a realização do evento.

Finalmente, foi feita uma divisão da base de dados em janelas de tempo de 1, 5 e 10 minutos. O tamanho dos registros das distribuições são respectivamente 14.644, 71.308 e 137.057, respectivamente. Essa abordagem oferece informações sobre o conjunto de dados processados. Esta etapa destacou a diversidade e a distribuição dos dados entre as diferentes ligas, fornecendo uma compreensão visual da distribuição de resultados após o pré-processamento e balanceamento. Para cada etapa subsequente foi feita a análise para cada uma das janelas de tempo, a fim de adquirir melhores resultados.

### 3.3 Separação para treinamento e teste

Para o modelo de floresta aleatória, optou-se por utilizar este algoritmo específico para calcular a importância de cada atributo, dado que ele é particularmente adequado para este tipo de análise nos dados em questão. A preparação para este modelo concentrou-se na normalização das características numéricas, sem a necessidade de codificação one-hot para variáveis categóricas [Johnson and Khoshgoftaar 2021]. Esse enfoque se deve ao fato de o modelo ser usado exclusivamente para avaliar o impacto de cada atributo no desempenho geral, uma análise que a floresta aleatória facilita.

Por outro lado, a utilização de técnicas de AutoML com a biblioteca TPOT permite a inclusão de diversos modelos, mas exclui modelos baseados em redes neurais. Devido a essa limitação, foi desenvolvida uma rede neural separadamente, com o objetivo de explorar a possibilidade de alcançar resultados superiores.

Antes da divisão dos dados em conjuntos de treino e teste, realizou-se o balanceamento das classes. Este passo crucial foi feito para evitar viés nos modelos devido à desproporção de classes. A divisão subsequente em conjuntos de treinamento e teste foi realizada utilizando a função train_test_split da biblioteca Scikit-Learn, seguindo a proporção padrão de 80% dos dados para treinamento e 20% para teste. Esta divisão foi feita de forma estratificada, garantindo que a proporção das classes na variável alvo fosse mantida igualmente nos conjuntos de treino e teste, assegurando a integridade da análise e a relevância dos modelos testados.

### 3.4 Avaliação da importância dos atributos no desempenho das equipes por liga utilizando modelagem preditiva

Foi feita a implementação de um modelo preditivo para cada liga esportiva em cada janela de tempo, utilizando o algoritmo *RandomForestClassifier* da biblioteca *Scikit-Learn*. Essa abordagem foi feita pelo fato deste modelo ser usado para análise da importância de atributos, pelo seu cálculo de impureza, o cálculo de impurezas é essencial para determinar como os nós nas árvores de decisão são divididos, Wang e Xia (2017) demonstraram como diferentes critérios de divisão de atributos em árvores de decisão podem ser unificados sob a estrutura da entropia de Tsallis, proporcionando uma abordagem eficiente e melhorando o desempenho das árvores de decisão clássicas. A ideia principal é que cada nó de uma árvore dentro da floresta aleatória deve separar as classes dos dados da maneira mais "pura" possível. A medida de impureza ajuda a avaliar quão bem um nó particiona os dados em subgrupos homogêneos com relação à variável de resposta. Existem principalmente dois critérios de impureza utilizados: a entropia e o índice de Gini. Por esse motivo, foi feita análise de cada uma das ligas com objetivo de calcular a importância que cada liga tem em relação às outras.

O índice de Gini mede a frequência com que um elemento aleatoriamente escolhido seria incorretamente classificado se fosse aleatoriamente rotulado de acordo com a distribuição das etiquetas no subconjunto. A pesquisa de Strobl et al. (2007) discutiu a seleção imparcial de divisão para árvores de classificação baseada no Índice de Gini, destacando a necessidade de evitar vieses na seleção de variáveis. Para um dado nó, o Gini é calculado como:

$$G = 1 - \sum_{i-1}^{j} pi^2$$

Aqui, pi representa a proporção de amostras pertencentes à classe i no conjunto de dados. O valor de Gini varia entre 0 (todos os elementos são da mesma classe, pura) e 0.5 (os elementos estão uniformemente distribuídos entre algumas classes, máxima impureza) para o caso binário. Em problemas multiclasse, o máximo pode ser diferente.

A entropia é uma medida de incerteza ou desordem, oriunda da teoria da informação. No contexto de uma árvore de decisão, é usada para calcular a heterogeneidade das amostras dentro de um nó. Louppe (2014) forneceu uma análise profunda dos random forests, explorando os mecanismos, propriedades e limitações do algoritmo, além de discutir a importância das variáveis baseada na redução da impureza média. A entropia para um conjunto é dada pela fórmula:

$$H = -\sum_{i=1}^{J} pi \, log \, pi$$

Onde pi é a proporção de amostras pertencentes à classe i. A entropia será zero quando todas as amostras em um nó pertencerem a uma única classe (máxima pureza) e será máxima quando as amostras estiverem igualmente distribuídas entre as classes.

O algoritmo de RandomForestClassifier utiliza um desses critérios (ou outro especificado pelo usuário) para avaliar a melhor maneira de dividir os nós em cada árvore. Em cada divisão, o modelo seleciona aleatoriamente um subconjunto de recursos e busca a divisão que resulta na maior redução de impureza. A floresta resultante é uma agregação (ensemble) das árvores, onde cada árvore contribui com sua própria previsão. O resultado final do modelo é tipicamente a média (para regressão) ou a moda (para classificação) das previsões de todas as árvores.

### 3.5 Desenvolvimento estratégico dos atributos após avaliação de importância no modelo preditivo de desempenho

Na fase atual, avalia-se a importância dos atributos no desempenho das equipes por ligas, utilizando um modelo preditivo baseado no algoritmo RandomForestClassifier. Esta análise é crucial para identificar quais atributos fornecem as informações mais valiosas para o nosso modelo e quais podem ser descartados ou reavaliados. Observa-se, por exemplo, que atributos como 'goalHome' e 'goalAway' não variam significativamente, sugerindo um impacto limitado sobre as previsões do modelo.

Após essa avaliação, introduzem-se novos atributos ao modelo. O momento ideal para esta adição é após a visualização da importância das características atuais. A adição progressiva de variáveis ao modelo preditivo é uma estratégia chave para entender e otimizar o impacto de cada atributo no desempenho do modelo. Inicialmente, trabalha-se com um conjunto base de variáveis selecionadas por sua relevância teórica e disponibilidade. Este conjunto inicial permite estabelecer uma linha de base do desempenho do modelo. Após esta etapa, a importância de cada atributo é avaliada através de técnicas analíticas, como a análise de importância de atributos. Essa avaliação identifica quais variáveis têm maior influência no resultado preditivo, permitindo uma otimização focada.

Adicionar todas as variáveis de uma só vez poderia não apenas aumentar a complexidade do modelo desnecessariamente, mas também obscurecer o entendimento do impacto individual de cada variável. A introdução gradual de novas variáveis, guiada pela análise de importância, permite ajustar o modelo de maneira mais precisa e entender como cada novo atributo afeta a precisão das previsões. Esse processo iterativo de adição e avaliação assegura que o modelo se mantenha relevante e robusto, adaptando-se às mudanças e refinando continuamente o entendimento dos fatores que influenciam o desempenho da partida. O diagrama abaixo mostra como funciona esta etapa de adição e exclusão gradual dos atributos.

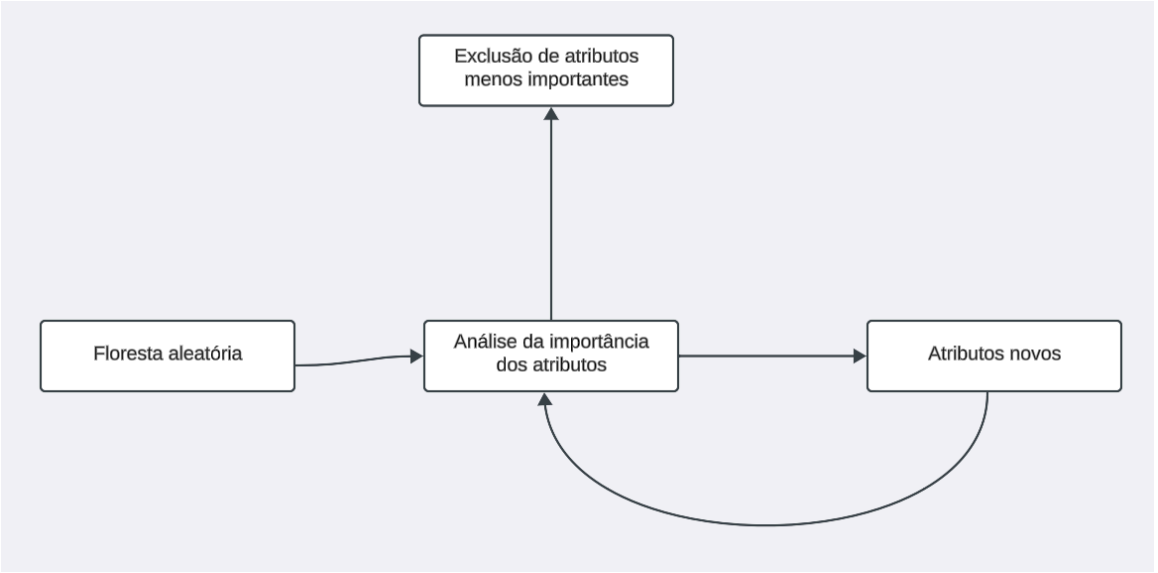


**Figura 1. Diagrama da etapa de análise da importância dos atributos**

A visualização das novas características deve ser apresentada de forma que ilustra claramente a sua contribuição potencial para a precisão do modelo. Ao introduzir novas características, o objetivo é abordar as lacunas deixadas pelas características atuais, potencialmente capturando aspectos mais dinâmicos do desempenho em campo que podem influenciar o resultado das partidas. Ao melhorar o conjunto de características, espera-se não apenas aumentar as métricas de desempenho, mas também fornecer informações importantes das ligas de futebol. Este processo de refinamento contínuo é essencial para manter o modelo relevante e robusto diante das constantes mudanças no mundo dos jogos esportivos. Cada atributo adicionado fornece informação sobre aspectos específicos do jogo.

- **Força de Ataque e Defesa**: Calculadas usando a fórmula de Maher, que considera os gols marcados e sofridos. As médias rolantes dos últimos 10 jogos de cada equipe fornecem uma medida dinâmica do desempenho ofensivo e defensivo, ajustando-se à forma recente das equipes.

- **Taxas de Vitória, Empate e Derrota**: Estas foram determinadas a partir de uma média rolante dos últimos 10 jogos, oferecendo uma perspectiva do histórico recente de resultados das equipes. Isso ajuda a entender a consistência e as tendências de resultados das equipes.

- **Eficiência de Ataque**: Medida pela precisão dos chutes, calculada como a razão entre chutes no gol e o total de chutes. Essa característica indica a qualidade das oportunidades criadas e a eficácia dos atacantes em converter essas chances em gols.

- **Desempenho Defensivo**: Avaliado por métricas como pressão de ataque, agressividade, eficácia defensiva e desarmes ao longo do tempo. Essas métricas fornecem uma compreensão detalhada do comportamento defensivo das equipes, desde a habilidade de bloquear chutes, até a intensidade de pressão e agressividade durante os jogos.

- **Desempenho de Passes**: Indicadores como controle de posse de bola e risco de passe foram calculados. Eles refletem como as equipes gerenciam a posse de bola e a propensão a assumir riscos nas tentativas de passe, revelando estratégias de jogo e

habilidades de controle da partida.

- **Dados sobre Cartões**: O total de cartões (amarelos e vermelhos) foi contabilizado para cada equipe, oferecendo *insights* sobre a disciplina dos jogadores e a tendência de cometer faltas.

A Tabela 3 mostra os atributos adicionados juntos com os atributos iniciais. Cada um desses atributos foi calculado e adicionado ao conjunto de dados para fornecer uma visão mais rica e detalhada do desempenho das equipes, permitindo análises mais precisas.

**Tabela 3. Descrição dos novos atributos**

| **Atributo** | **Descrição** |
|---|---|
| f_attack_home | Força do ataque do time da casa |
| f_defensive_away | Força defensiva do time visitante |
| f_defensive_home | Força defensiva do time da casa |
| f_attack_away | Força do ataque do time visitante |
| win_rate_home | Taxa de vitórias do time da casa |
| loss_rate_home | Taxa de derrotas do time da casa |
| draw_rate_home | Taxa de empates do time da casa |
| win_rate_away | Taxa de vitórias do time visitante |
| loss_rate_away | Taxa de derrotas do time visitante |
| draw_rate_away | Taxa de empates do time visitante |
| shotAccuracy_home | Precisão de chutes do time da casa |
| shotAccuracy_away | Precisão de chutes do time visitante |
| attackPressureOverTime_home | Pressão de ataque ao longo do tempo do time da casa |
| attackPressureOverTime_away | Pressão de ataque ao longo do tempo do time visitante |
| aggrressionOverTime_home | Agressividade ao longo do tempo do time da casa |
| aggresssionOverTime_away | Agressividade ao longo do tempo do time visitante |
| defensiveEfficacy_home | Eficácia defensiva do time da casa |
| defensiveEfficacy_away | Eficácia defensiva do time visitante |
| tacklesOverTime_home | Desarmes ao longo do tempo do time da casa |
| tacklesOverTime_away | Desarmes ao longo do tempo do time visitante |
| possessionControl | Controle de posse de bola no jogo |
| passRisk_home | Risco nos passes do time da casa |
| passRisk_away | Risco nos passes do time visitante |
| 05ht_home | Mais de um gol no primeiro tempo pelo time da casa |
| 05ft_home | Mais de um gol no segundo tempo pelo time da casa |
| 05_home | Mais de um gol no jogo pelo time da casa |
| 05ht_away | Mais de um gol no primeiro tempo pelo time visitante |
| 05ft_away | Mais de um gol no segundo tempo pelo time visitante |
| 05_away | Mais de um gol no jogo pelo time visitante |
| 15ht_home | Mais de dois gols no primeiro tempo pelo time da casa |

| Atributo | Descrição |
|---|---|
| 15ht_away | Mais de dois gols no primeiro tempo pelo time de fora |
| 15ft_home | Mais de dois gols no segundo tempo pelo time de casa |
| 15ft_away | Mais de dois gols no segundo tempo pelo time de visitante |
| 15_home | Mais de dois gols na partida pelo time da casa |
| 15_away | Mais de dois gols na partida pelo time visitante |
| 25ht_home | Mais de três gols no primeiro tempo pelo time da casa |
| 25ht_away | Mais de três gols no primeiro tempo pelo time visitante |
| 25ft_home | Mais de três gols no segundo tempo pelo time da casa |
| 25ft_away | Mais de três gols no segundo tempo pelo time visitante |
| 25_home | Mais de três gols na partida pelo time da casa |
| 25_away | Mais de três gols na partida pelo time visitante |

Com as novas características adicionadas e remoção das características menos importantes, ou seja, aquelas que obtiveram menos de 0,01 de valor de importância relacionadas umas com as outras, sendo assim, podemos analisar novamente com a mesma abordagem utilizada na etapa anterior, verificar a importância das características em cada liga conforme o diagrama da Figura 1.

### 3.6 Construção dos modelos de aprendizagem

Após o processo de pré-processamento de dados, a divisão da base em conjuntos de treinamento e teste e engenharia de atributos, a fase seguinte compreendeu a construção dos modelos utilizando redes neurais e *AutoML*, para aprimorar a precisão na classificação de dados complexos, aproveitando as capacidades avançadas de aprendizado profundo. Estas tecnologias permitem o modelamento de relações não-lineares e interações complexas entre variáveis, que são comuns em dados esportivos. Redes neurais oferecem uma abordagem flexível e poderosa para capturar padrões sutis nos dados, enquanto o AutoML facilita a seleção do melhor modelo e configuração de hiperparâmetros, otimizando o desempenho sem a necessidade de intervenção manual. Juntos, esses métodos visam construir modelos mais robustos e confiáveis, capazes de prever resultados com maior acurácia. Este processo envolveu a implementação de uma estrutura que utilizou técnicas de aprendizado profundo para aprimorar a precisão na classificação de dados complexos. A metodologia adotada para o desenvolvimento desses modelos é detalhada a seguir, evidenciando as técnicas e ferramentas empregadas para alcançar resultados mais robustos e confiáveis.

#### 3.6.1 Construção do modelo de Redes neurais

Inicialmente, foi definido um espaço de hiperparâmetros abrangente, que inclui variações no número de neurônios por camada, taxas de *dropout* para regularização, tipos de ativação, tamanhos de lote e taxas de aprendizado. Essa diversidade permitiu uma exploração ampla das possibilidades de configuração da rede neural utilizando a técnica de *Random Search*, uma técnica que procura aleatoriamente o melhor parâmetro no espaço de hiperparâmetros, permitindo a identificação da combinação mais eficaz para o problema em questão.

Para a construção do modelo de redes neurais, utilizou-se a biblioteca *Keras*, que facilitou a implementação de redes neurais sequenciais. Para cada janela de tempo foi utilizado parâmetros diferentes, seguindo os parâmetros selecionados durante o processo

de busca aleatória. A otimização dos modelos foi realizada através do algoritmo *Adam*, que se mostrou eficiente na minimização da função de perda de entropia cruzada binária, além de ajustar dinamicamente as taxas de aprendizado durante o treinamento. O processo de treinamento também incluiu a implementação de critérios de parada (*callbacks*), como *EarlyStopping* e *ReduceLROnPlateau*, que contribuíram para a prevenção do superajustamento e a otimização do processo de aprendizado. Os parâmetros mencionados podem ser vistos na Tabela 4.

**Tabela 4. Hiperparâmetro e biblioteca utilizada no modelo de rede neural**

| Nome | Biblioteca/Framework | Hiperparâmetro |
|---|---|---|
| Rede Neural (1 min) | Sequential() da biblioteca keras.models | neurons: 512, 256, dropout_rate: 0.5, activation_type: relu, batch_size: 16, learning_rate: 0.00001 |
| Rede Neural (5 min) | Sequential() da biblioteca keras.models | neurons: 512, 256, 164, dropout_rate: 0.2, activation_type: relu, batch_size: 32, learning_rate: 0.0001 |
| Rede Neural (10 min) | Sequential() da biblioteca keras.models | neurons: 512, 256, 128, dropout_rate: 0.2, activation_type: relu, batch_size: 32, learning_rate: 0.0001 |

A seleção dos melhores hiperparâmetros foi realizada por meio de um procedimento de busca aleatória, onde diferentes combinações foram testadas em um número limitado de iterações. Esse método permitiu a avaliação eficiente de diversas configurações, identificando aquela que proporcionou a melhor acurácia de validação. Com os hiperparâmetros otimizados, procedeu-se à construção e treinamento final do modelo, que foi submetido a um treinamento prolongado até que os critérios de parada fossem atendidos, garantindo assim a obtenção de um modelo bem ajustado e capaz de generalizar para dados não vistos anteriormente, na base de dados de teste.

Os resultados obtidos foram avaliados por meio de métricas como *recall*, AUC (*Area Under The Curve*), ROC (*Receiver Operating Characteristic*), medida F1, tanto no conjunto de treinamento, quanto no de teste. Além disso, foram geradas matrizes de confusão para uma análise detalhada do desempenho do modelo. A curva de aprendizado, representando a evolução da acurácia ao longo das épocas, foi plotada para visualizar a dinâmica do treinamento.

### 3.6.2 Construção do modelo de *AutoML*

Após a construção e avaliação dos modelos de redes neurais, a próxima etapa do estudo envolveu a implementação e o treinamento de um modelo de *AutoML*, utilizando o *TPOT*, uma ferramenta com *pipeline* de otimização baseada em árvore (*Tree-based Pipeline Optimization Tool*).

AutoML visa facilitar e automatizar o desenvolvimento de modelos de aprendizado de máquina. AutoML abrange desde a preparação dos dados até a seleção e

otimização de modelos, passando pela engenharia de recursos. Seu objetivo é encontrar a melhor pipeline de machine learning para um dado problema de forma automática e eficiente (Singh & Joshi, 2022); (He et al., 2019).

O processo de implementação do *TPOT* iniciou com a sua inicialização, definindo parâmetros como o número de gerações para 5 e o tamanho da população para 50, além de ajustar a verbosidade para 2 e fixar o estado aleatório em 42. Essas configurações determinam o escopo da busca pelo *pipeline* ótimo e a profundidade da exploração do espaço de modelos possíveis. Após a inicialização, o *TPOT* foi treinado com o conjunto de dados de treinamento. Esse processo de treinamento não apenas envolveu a seleção do modelo, mas também a otimização de seus hiperparâmetros e a configuração do *pipeline* de pré-processamento de dados, tudo de maneira automática e guiada pelo desempenho nos dados de treinamento. Concluído o treinamento, o *pipeline* otimizado foi exportado e o modelo resultante foi salvo, permitindo sua reutilização em aplicações futuras ou para a realização de novas predições sem a necessidade de executar todo o processo de otimização novamente. A Tabela 5 a seguir mostra os diferentes *pipelines* encontrados para as janelas de tempo.

**Tabela 5. Hiperparâmetro e bibliotecas utilizadas no modelo de AutoML**

| Nome | Biblioteca/Framework | Hiperparâmetro |
|---|---|---|
| AutoML (1 min) | TPOTClassifier() da biblioteca tpot | ExtraTreesClassifier, bootstrap=False, criterion=entropy, max_features=0.85, min_samples_leaf=14, min_samples_split=9, n_estimators=100 |
| AutoML (5 min) | TPOTClassifier() da biblioteca tpot | RandomForestClassifier, bootstrap=False, criterion=entropy, max_features=0.2, min_samples_leaf=1, min_samples_split=2, n_estimators=100 |
| AutoML (10 min) | TPOTClassifier() da biblioteca tpot | ExtraTreesClassifier with MaxAbsScaler, bootstrap=True, criterion=entropy, max_features=0.8, min_samples_leaf=2, min_samples_split=2, n_estimators=100 |

Para avaliar o desempenho do modelo encontrado no *AutoML*, foram utilizadas as predições no conjunto de teste para calcular métricas de desempenho. A probabilidade de cada classe foi obtida e utilizada para construir a curva ROC e a AUC foi calculada, fornecendo uma medida quantitativa da capacidade do modelo de discriminar entre as classes. A visualização da curva ROC, facilitou a interpretação do desempenho do modelo em termos de sensibilidade (taxa de verdadeiro positivo) e especificidade (1 - taxa de falso positivo). Além disso, foi gerada uma matriz de confusão para oferecer

uma visão detalhada do número de previsões corretas e incorretas feitas pelo modelo, categorizadas por classe. Essa visualização foi complementada por um relatório de classificação, que apresentou as métricas de desempenho do modelo mencionadas na seção anterior, permitindo uma análise mais aprofundada nas diferentes categorias de dados.

## 4. Resultados e discussões

Nesta seção será analisada a importância dos diversos atributos e como eles influenciam o desempenho dos modelos, seguido de uma comparação entre os resultados obtidos nas diferentes janelas de tempo, visando entender o impacto temporal de cada modelo. Além disso, será discutido métricas de desempenho como precisão, recall e F1-score, além de apresentar as matrizes de confusão e curvas de aprendizado das diferentes janelas de tempo de cada modelo.

### 4.1. Análise de Importância dos Atributos

Inicialmente, observou-se a variação da importância dos atributos antes e após a etapa de engenharia de atributos para diferentes janelas de tempo (1, 5, e 10 minutos), indicando a relevância da manipulação de dados na melhoria da precisão dos modelos. As imagens ilustrativas da variação da importância dos atributos antes e depois da engenharia, bem como após a remoção de atributos, com importância abaixo de 0,01, oferecem uma visualização clara de como o processo de seleção e modificação de atributos influencia diretamente a capacidade de previsão dos modelos.

A etapa de visualização concentra-se em destacar a variação da importância dos atributos entre as ligas, utilizando um gráfico *boxplot* que apresenta como diferentes atributos impactam o modelo nas janelas de tempo de 1, 5 e 10 minutos como é mostrado na Figura 2, 3 e 4, respectivamente.

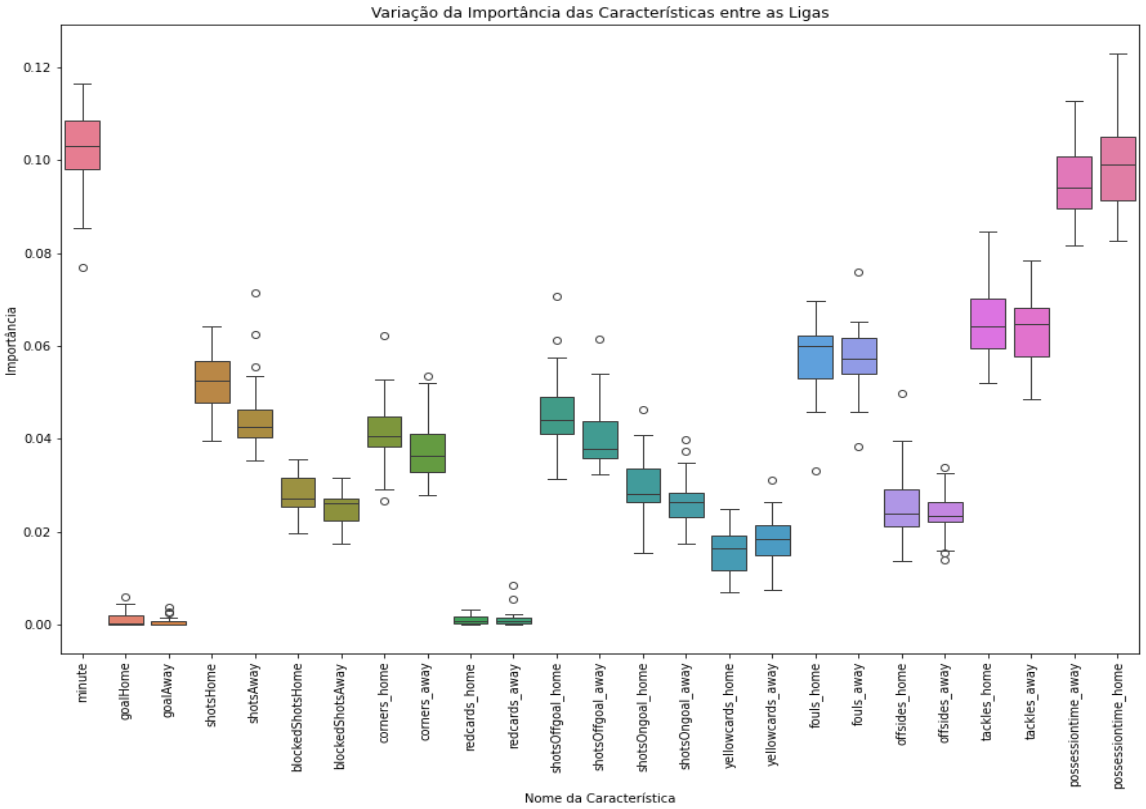


**Figura 2. Gráfico de variação da importância das características em diferentes ligas com janela de tempo de 1 minuto por partida.**

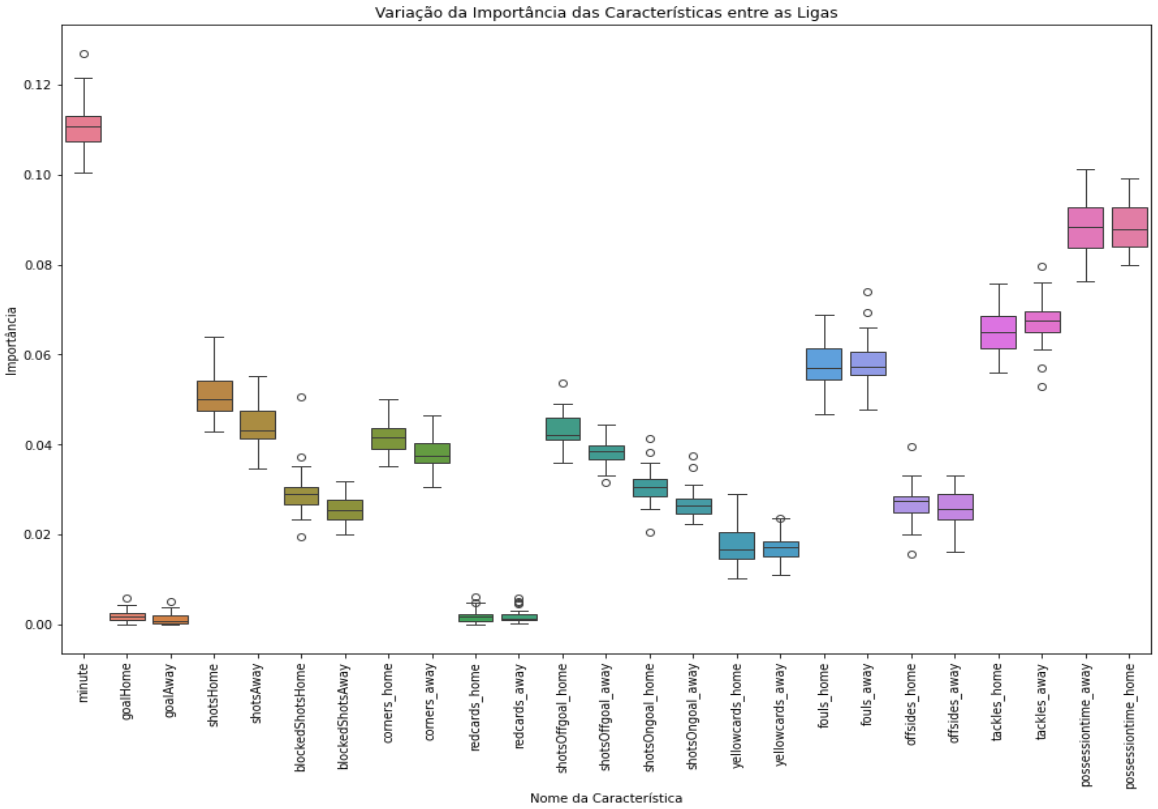


**Figura 3. Gráfico de variação da importância das características em diferentes ligas com janela de tempo de 5 minutos por partida.**

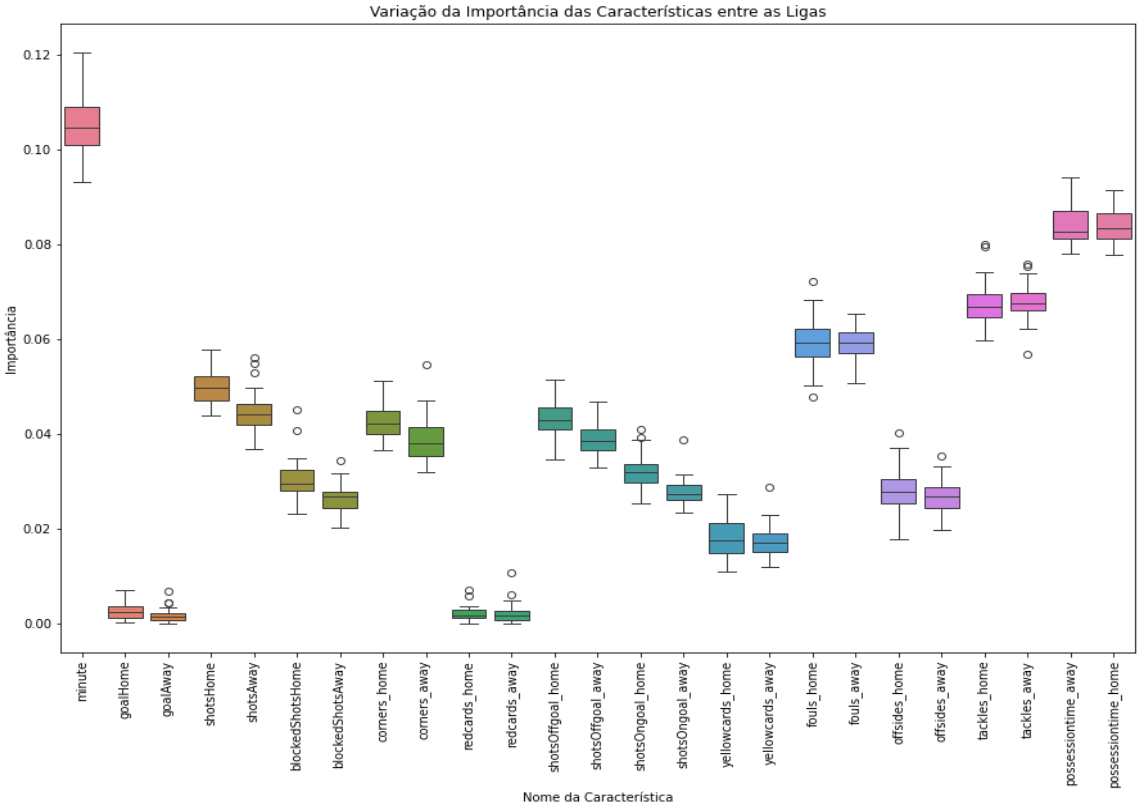


**Figura 4. Gráfico de variação da importância das características em diferentes ligas com janela de tempo de 10 minutos por partida.**

A linha central de cada caixa reflete a mediana da importância atribuída, enquanto os limites superior e inferior da caixa marcam, respectivamente, o terceiro e o primeiro quartil. Pontos que aparecem distantes das caixas representam valores atípicos, denotando uma discrepância significativa em relação às demais medidas de importância. É notável que atributos como ‘goalHome’ e ‘goalAway’ exibem uma variação limitada e uma importância reduzida. Isso pode ser atribuído à natureza desses dados, que se mantêm zerados até o momento de ocorrência do evento correspondente. As características ‘redcards_home’ e ‘redcards_away’ também apresentam baixa variação e importância no modelo, o que pode ser explicado pela maior frequência desses eventos no início das partidas, o que impacta menos a variabilidade ao longo do tempo.

## 4.2. Comparação entre as janelas de tempo

A avaliação da importância dos atributos nas previsões de eventos esportivos em diferentes janelas de tempo (1, 5 e 10 minutos) revela informações sobre a influência de cada atributo nos modelos preditivos. As figuras fornecidas ilustram claramente a variação da importância dessas características, antes e após o processo de engenharia de atributos.

- **Janela de Tempo de 1 Minuto:** A Figura 2 mostra a variação da importância das características em diferentes ligas com uma janela de tempo de 1 minuto, antes da engenharia de atributos. Observa-se que alguns atributos, como 'possesionTime_home' e 'possesionTime_away', possuem uma relevância mais destacada, sugerindo que a posse de bola imediatamente antes de um gol é um indicador significativo. No entanto, após a engenharia de atributos, como visto na Figura 4, a distribuição da importância muda substancialmente, indicando uma otimização do modelo para focar em variáveis mais influentes, com menos variabilidade entre os atributos.

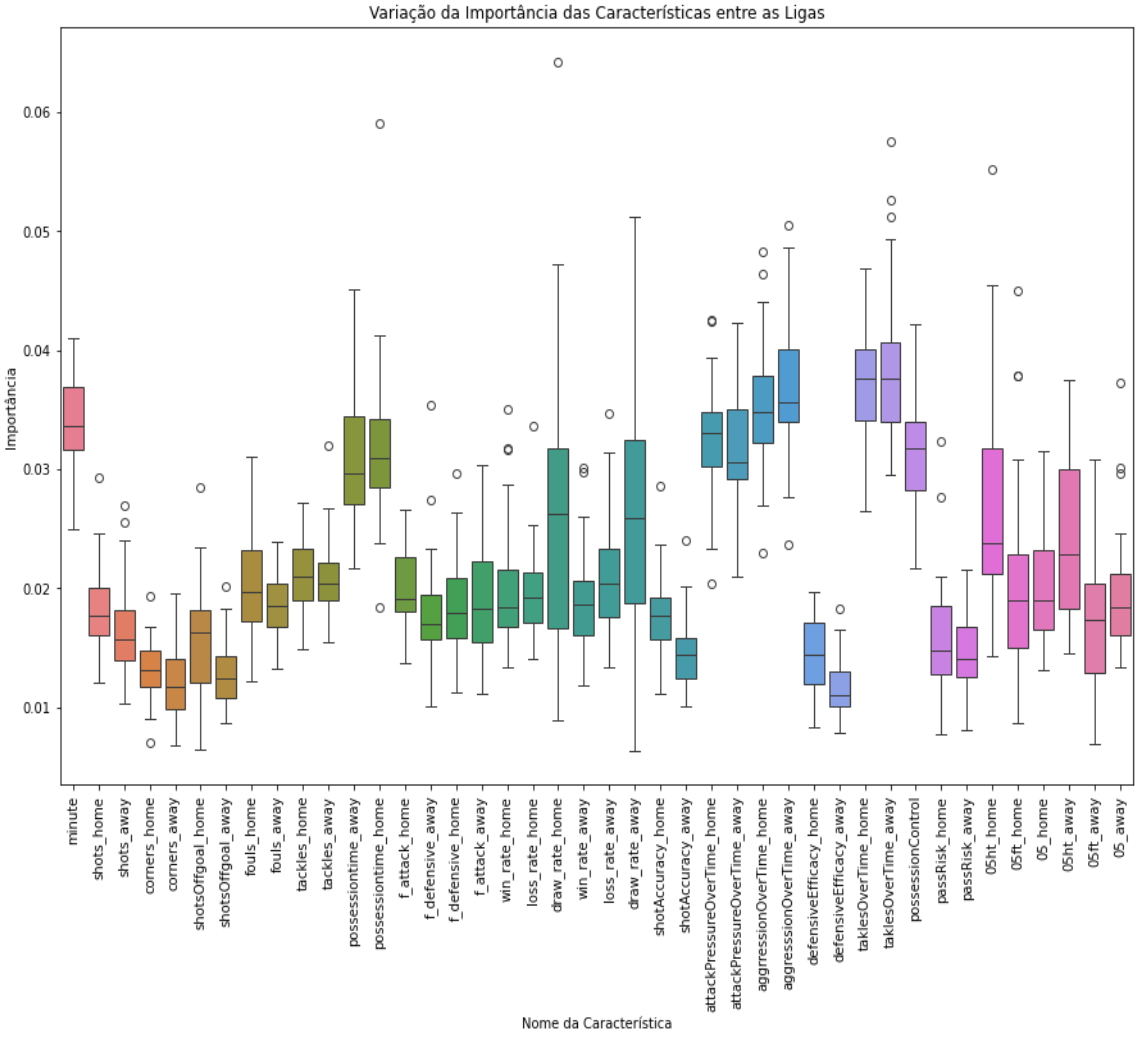


**Figura 5. Gráfico de variação da importância das características em diferentes ligas com janela de tempo de 1 minuto por partida após a engenharia de características.**

- **Janela de Tempo de 5 Minutos:** A Figura 3 mostra a variação da importância das características em diferentes ligas com uma janela de tempo de 5 minutos, antes da engenharia de atributos. Comparando com a Figura 6 após a engenharia de atributos, há uma diminuição perceptível na variação e uma distribuição mais uniforme da importância. Isso sugere que essa etapa ajudou a identificar quais características têm uma influência mais consistente no modelo, melhorando a capacidade preditiva ao longo de um período mais extenso antes do evento.

.

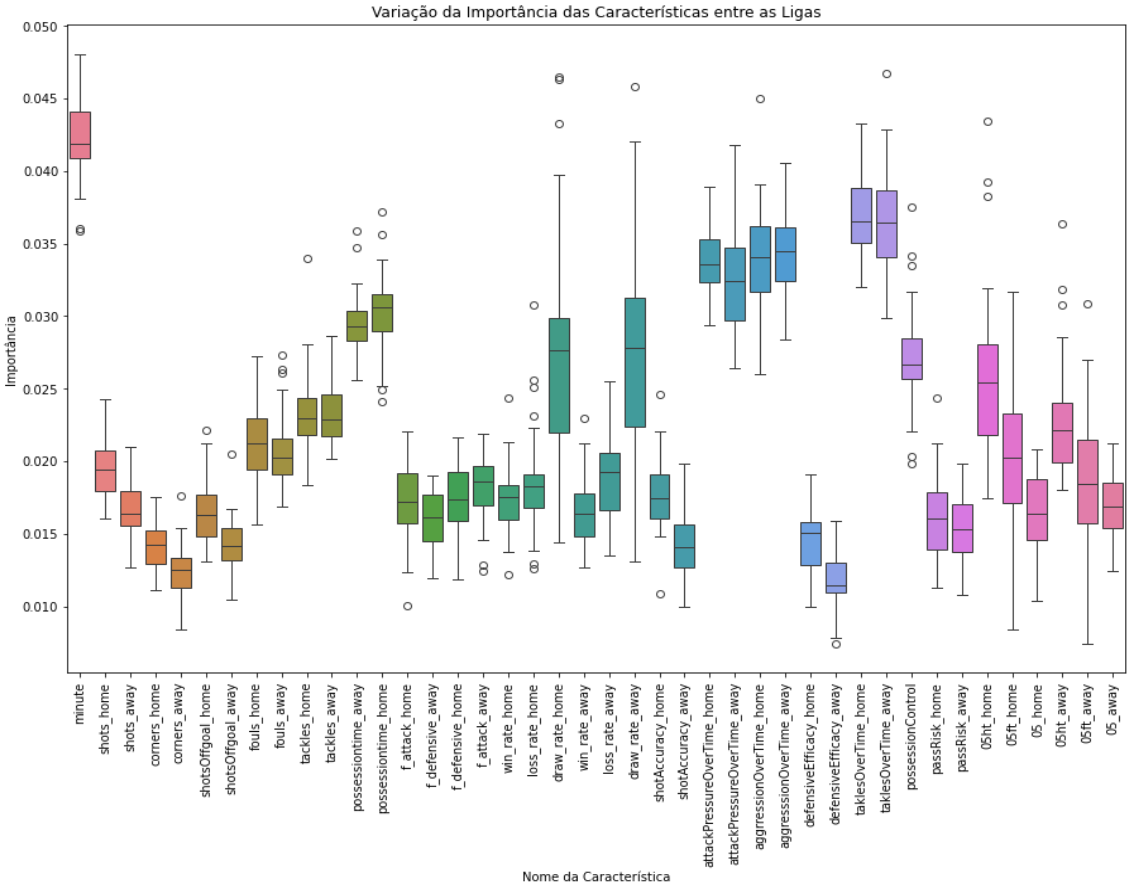


**Figura 6. Gráfico de variação da importância das características em diferentes ligas com janela de tempo de 5 minutos por partida após a engenharia de características.**

- **Janela de Tempo de 10 Minutos:** Por fim, a Figura 4 mostra a variação da importância das características em diferentes ligas com uma janela de tempo de 10 minutos, antes da engenharia de atributos, Comparando com a Figura 7 após a engenharia de atributos, reitera o padrão observado nas janelas de tempo menores: a engenharia de atributos ajuda a equilibrar a importância das variáveis. Nota-se também que características previamente menos valorizadas ganham relevância, o que pode ser atribuído ao maior conjunto de dados disponível para análise, permitindo que o modelo capture padrões complexos que são menos evidentes em períodos mais curtos.

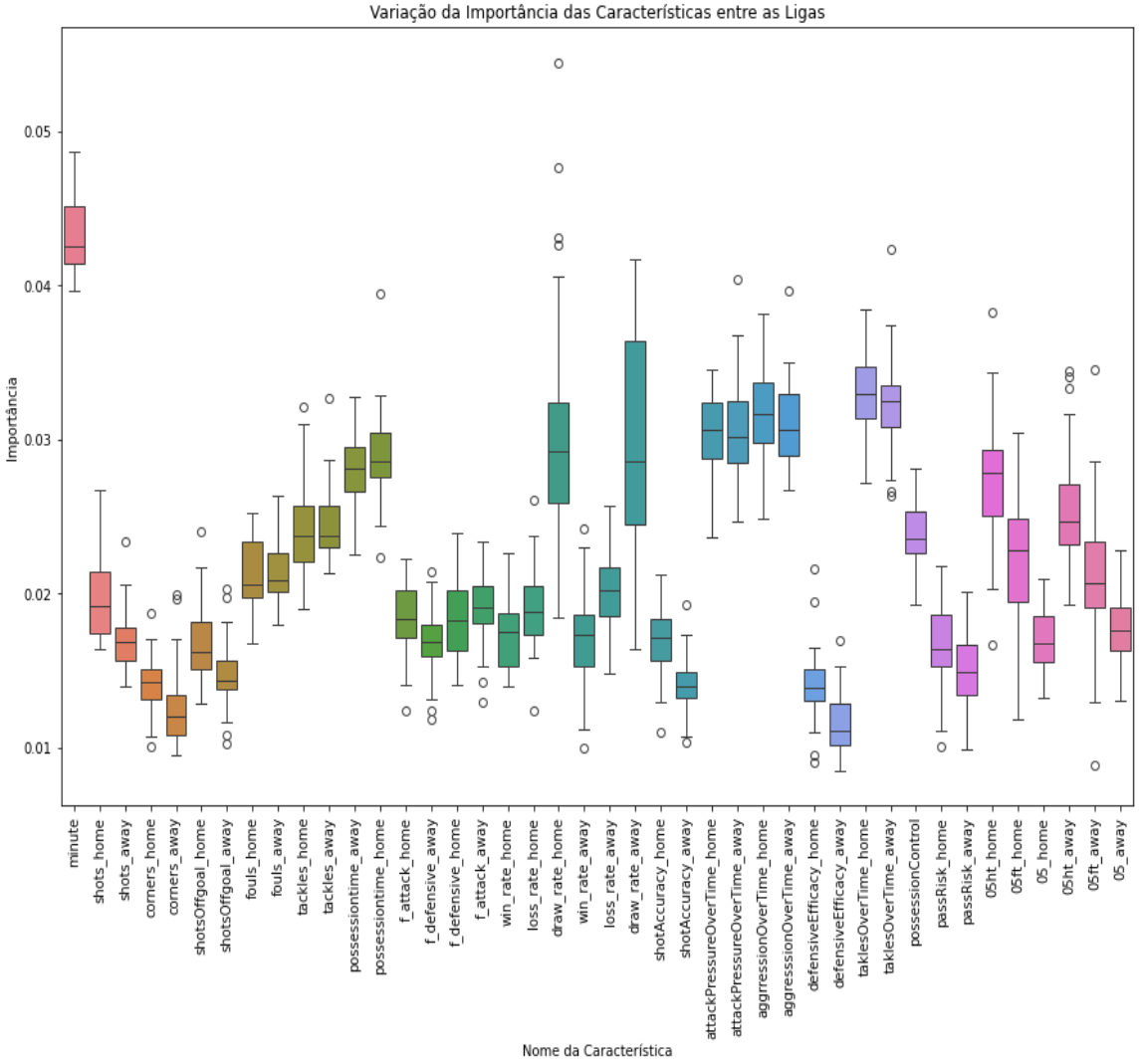


**Figura 7. Gráfico de variação da importância das características em diferentes ligas com janela de tempo de 10 minutos por partida após a engenharia de características.**

A comparação dos gráficos indica claramente que a engenharia de atributos é uma etapa crítica na modelagem preditiva para eventos esportivos ao vivo. Os modelos que utilizam dados de janelas de tempo mais curtas podem ser mais suscetíveis a variações e ruídos, enquanto janelas de tempo mais longas, combinadas com uma seleção sistemática de atributos, podem fornecer previsões precisas e robustas.

### 4.3. Análise dos resultados dos modelos

Nesta seção será realizada uma avaliação detalhada do desempenho dos modelos preditivos, centrando-se nas redes neurais e na AutoML. Esta análise contempla a efetividade dos modelos nas diversas janelas de tempo, oferecendo uma perspectiva abrangente de como cada modelo se comporta nas janelas de tempo. Toda a análise é verificada pelas métricas cruciais como precisão, recall e F1-score, além das visualizações de matrizes de confusão e curvas de aprendizado, que ajudam a ilustrar o sucesso e as limitações dos modelos em prever os eventos esportivos. Esta seção visa não apenas quantificar a performance dos modelos, mas também fornecer insights sobre como diferentes configurações e ajustes de parâmetros impactam a capacidade preditiva, guiando futuras melhorias e aplicações práticas dos modelos.

### 4.3.1. Análise de Resultados da Rede Neural com Janela de Tempo de 1 Minuto

A rede neural configurada com parâmetros otimizados para um contexto de janela de tempo de 1 minuto apresentou uma tendência de aprendizado gradual, conforme evidenciado pela curva de aprendizado. Os melhores parâmetros indicam uma estrutura de duas camadas ocultas com 512 e 256 neurônios, uma taxa de abandono (*dropout*) de 50%, função de ativação ReLU, tamanho de lote de 16 e uma taxa de aprendizado muito baixa de 0,00001. Este conjunto de hiperparâmetros sugere um foco na mitigação do sobreajuste, dada a alta taxa de *dropout* e a baixa taxa de aprendizado.

Durante o treinamento, a acurácia do modelo alcançou cerca de 58% no conjunto de treinamento e aproximadamente 60% no de validação como é mostrado entre as curvas de treino e validação na Figura 8. O *recall* de 70,4% no treino e 70,2% no teste sugere uma tendência do modelo em capturar a maioria dos eventos positivos, enquanto a precisão em torno de 63% implica que o modelo possui uma proporção razoável de acertos e erros entre as previsões.

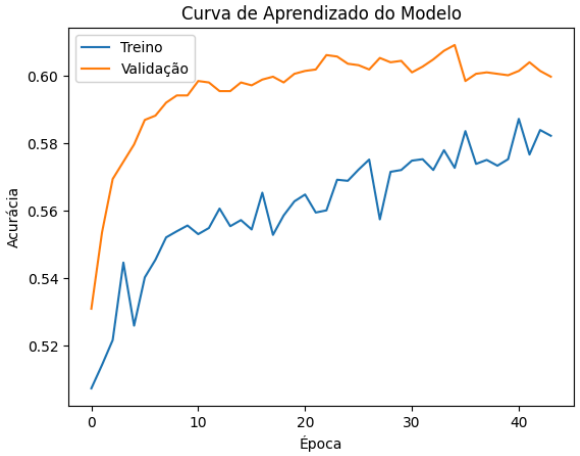


**Figura 8. Curva de aprendizado da Rede Neural com janela de tempo de 1 minuto.**

A Matriz de Confusão na Figura 9 reforça essa interpretação, com o modelo prevendo corretamente a classe positiva (Classe 1) em muitas ocasiões, mas ainda confundindo uma quantidade considerável de eventos negativos (Classe 0) como positivos, o que se reflete em um número substancial de falsos positivos (FP). Esta é uma indicação de que o modelo pode estar inclinado para prever eventos positivos com mais frequência do que o ideal.

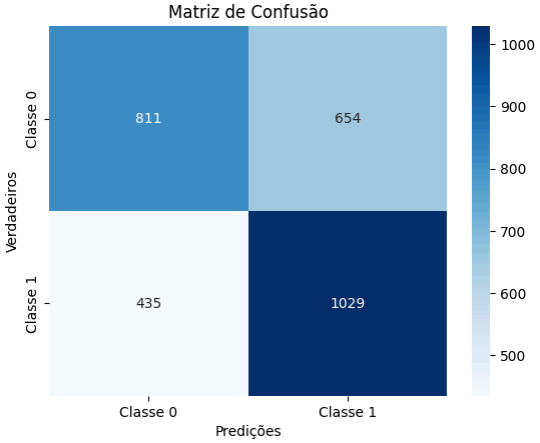


**Figura 9. Matriz de confusão da Rede Neural com janela de tempo de 1 minuto.**

As métricas do Relatório de Classificação (mostrados na Tabela 6)

complementam a análise, mostrando que a Classe 1 (eventos positivos, como gols) tem um *recall* mais alto em comparação com a Classe 0. Isto é consistente com a maior quantidade de verdadeiros positivos (VP) em comparação com os verdadeiros negativos (VN) na Matriz de Confusão. A acurácia global do modelo, bem como a média ponderada e macro de precisão, *recall* e pontuação F1, ficou em torno de 63%, o que indica uma performance moderada.

**Tabela 6. Resultado das métricas do Relatório de Classificação**

| Modelo / Métrica | Precisão Classe 0 | Precisão Classe 1 | Recall Classe 0 | Recall Classe 1 | F1-Score Classe 0 | F1-Score Classe 1 |
|---|---|---|---|---|---|---|
| RN - 1 Minuto | 65% | 61% | 55% | 70% | 60% | 65% |
| AutoML - 1 Minuto | 68% | 64% | 60% | 71% | 64% | 68% |
| RN - 5 Minutos | 85% | 78% | 75% | 87% | 80% | 82% |
| AutoML - 5 Minutos | 96% | 90% | 89% | 97% | 93% | 93% |
| RN - 10 Minutos | 93% | 85% | 84% | 93% | 88% | 89% |
| AutoML - 10 Minutos | 98% | 92% | 91% | 98% | 94% | 95% |

### 4.3.2 Análise dos Resultados do AutoML com Janela de Tempo de 1 Minuto

O AutoML, aplicando o ExtraTreesClassifier, parece ter se saído ligeiramente melhor que o modelo de rede neural anterior para a mesma janela de tempo de 1 minuto, conforme evidenciado pelo escore de validação cruzada interno constante de aproximadamente 0,6484. O melhor *pipeline* identificado não utilizou o método *bootstrap* e optou por um critério de entropia, o que é típico para otimizar a ganho de informação em problemas de classificação.

A Matriz de Confusão mostrada na Figura 10 revela que o modelo teve um desempenho relativamente equilibrado entre as classes, com uma quantidade razoável de verdadeiros positivos e verdadeiros negativos. Entretanto, o número de falsos positivos permaneceu significativo, indicando que, apesar da melhoria em relação ao modelo de rede neural, ainda existem desafios na distinção precisa entre os eventos de cada classe.

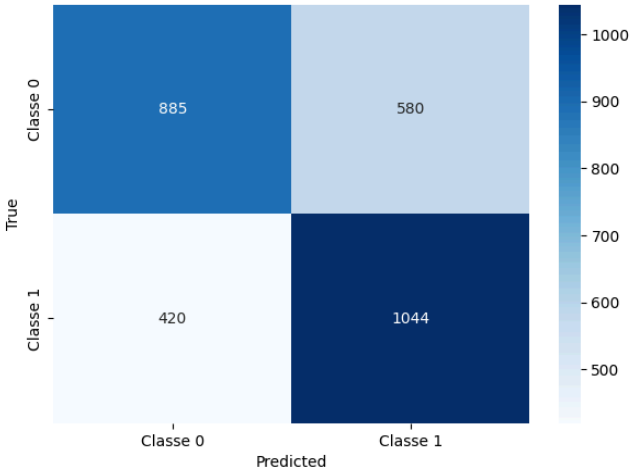


**Figura 10. Matriz de confusão do AutoML com janela de tempo de 1 minuto.**

As métricas do Relatório de Classificação (Tabela 4) mostram que a precisão e o *recall* para a Classe 1 estão alinhados com os da Classe 0, indicando uma boa capacidade do modelo em identificar corretamente as instâncias de cada classe. A acurácia geral do modelo de AutoML atingiu 0,66, ligeiramente superior ao modelo de rede neural, o que sugere que o modelo AutoML pode ser mais adequado para este cenário específico.

### 4.3.3 Análise de Resultados da Rede Neural com Janela de Tempo de 5 Minutos

Os resultados obtidos pela rede neural configurada com três camadas ocultas (512, 256, 164 neurônios) e com uma taxa de *dropout* de 0,2 sugerem uma capacidade robusta do modelo em lidar com superajustamento, dada a menor taxa de *dropout* comparada à janela de 1 minuto. A função de ativação *ReLU*, o tamanho do lote de 32 e uma taxa de aprendizado um pouco maior de 0,0001 também indicam um modelo que busca um equilíbrio entre a capacidade de adaptação e a prevenção de ajuste excessivo aos dados de treino.

Observando a Curva de Aprendizado da Figura 11, vemos que as acurácias de treino e validação convergem de forma mais próxima, implicando um modelo bem calibrado que generaliza bem para dados não vistos, além disso, o modelo poderia continuar aprendendo mas foi parado antecipadamente por alcancar a tolerância de alguns parâmetros do modelo do *EarlyStopping*.

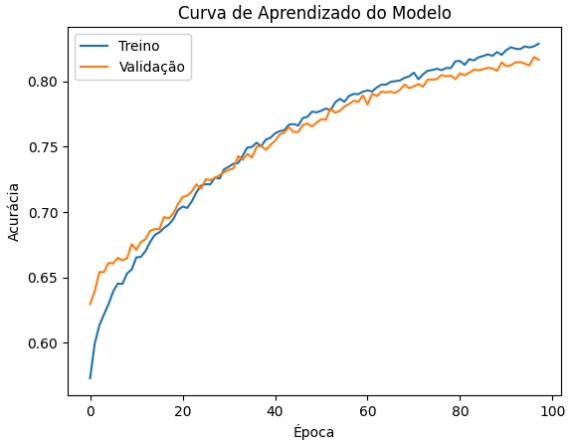


**Figura 11. Curva de aprendizado da Rede Neural com janela de tempo de 5 minutos**

Os valores de precisão, *recall* e pontuação F1 no Relatório de Classificação (mostrados na Tabela 4) são consideravelmente altos, com a Classe 1 (eventos positivos) tendo um *recall* de 0,87, o que indica que o modelo é eficaz em identificar corretamente a maioria dos eventos positivos. A precisão um pouco mais baixa para a Classe 1 (0,78), em comparação com a Classe 0 (0,85), sugere que o modelo pode estar predizendo mais falsos positivos, quando se trata de prever eventos.

A Matriz de Confusão da Figura 12 apoia essa análise, com um número significativo de verdadeiros positivos e verdadeiros negativos, mas também com a presença de falsos positivos e falsos negativos, apontando para a necessidade de possivelmente reajustar o limiar de decisão do modelo para melhorar a precisão sem comprometer a métrica de o *recall*.

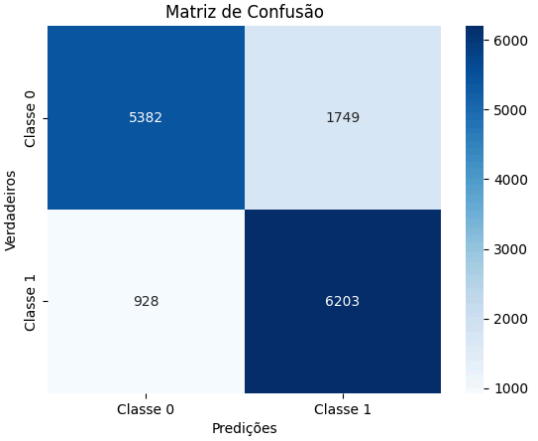


**Figura 12. Matriz de confusão da Rede Neural com janela de tempo de 5 minutos.**

A diferença entre a perda de treino e teste é relativamente pequena e sugere que o modelo não está apenas memorizando os dados de treino. Além disso, as métricas de AUC (Área sob a Curva) estão acima de 0,88, o que demonstra a boa capacidade do modelo de discriminar entre as classes.

### 4.3.4 Análise dos Resultados do AutoML com Janela de Tempo de 5 Minutos

O AutoML, com o *pipeline* RandomForestClassifier, mostrou resultados melhores na janela de tempo de 5 minutos, como evidenciado pelo aumento consistente do escore de validação cruzada interna ao longo das gerações, culminando em um valor de aproximadamente 0,91.

As métricas de desempenho apresentadas no Relatório de Classificação são particularmente altas, com precisão e *recall* para a Classe 1 (eventos positivos) atingindo 0,90 e 0,97, respectivamente. Estes valores indicam que o modelo é bom em identificar corretamente os eventos positivos (gols), e que a maior parte das previsões para esta classe é de fato correta. Da mesma forma, a Classe 0 (eventos negativos) tem uma precisão de 0,96 e um *recall* de 0,89, o que mostra que o modelo também é eficaz em identificar corretamente quando um evento não aconteceu.

A Matriz de Confusão (Figura 13) reforça a eficácia do modelo, com uma grande maioria de verdadeiros positivos e verdadeiros negativos, e relativamente poucos falsos positivos e falsos negativos. Isso é particularmente notável dada a dificuldade inerente à previsão de eventos esportivos ao vivo, onde as circunstâncias do jogo podem mudar rapidamente.

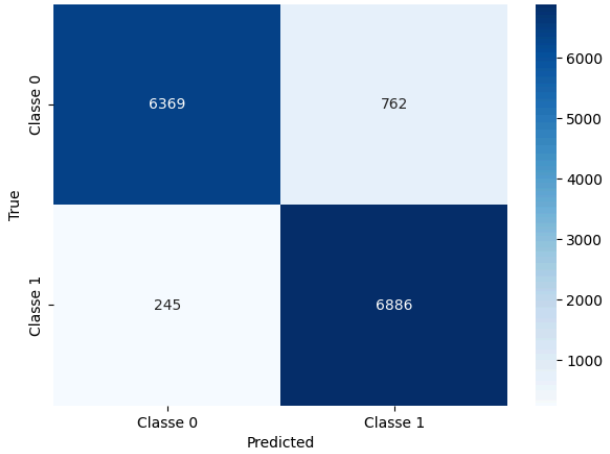


**Figura 13. Matriz de confusão do AutoML com janela de tempo de 5 minutos.**

A acurácia geral do modelo, bem como a média ponderada e macro de todas as métricas, ficou em 0,93, sugere que o modelo é bem equilibrado e oferece uma previsão confiável para ambos os tipos de eventos. O modelo parece estar bem ajustado aos dados, evitando tanto o sobreajustamento, quanto o superajustamento.

### 4.3.5 Análise de Resultados da Rede Neural com Janela de Tempo de 10 Minutos

Os resultados da rede neural para a janela de tempo de 10 minutos apresentam uma alta acurácia, tanto no conjunto de treino, quanto no de validação, como demonstrado pela curva de aprendizado na Figura 14, indica que a rede neural progrediu de forma consistente ao longo das épocas de treinamento. As curvas de treino e validação mostram que o modelo se generaliza bem, com pouco sobreajuste.

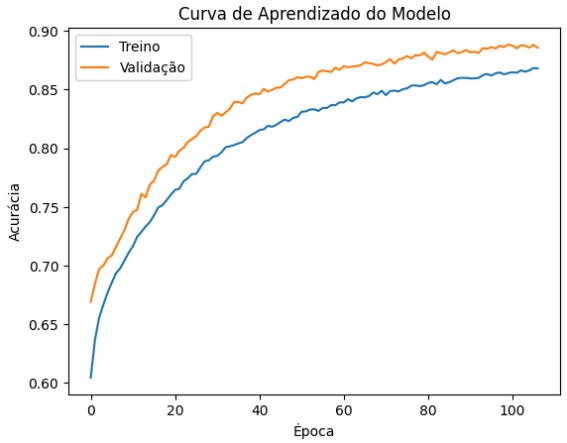


**Figura 14. Curva de aprendizado da Rede Neural com janela de tempo de 10 minutos.**

O resultado das métricas do Relatório de Classificação (Tabela 4) revela uma precisão e um *recall* elevados para ambas as classes, com a Classe 1 tendo uma pontuação particularmente alta de *recall* (0,93), o que indica que o modelo é eficaz em capturar eventos positivos. A precisão da Classe 0 de 0,93 indica que o modelo tem uma capacidade alta de prever corretamente a ausência de eventos.

A Matriz de Confusão mostrada na Figura 15 reforça esses achados, exibindo uma alta taxa de verdadeiros positivos e verdadeiros negativos. A quantidade de falsos negativos é relativamente baixa, o que é promissor para a aplicação do modelo em situações reais, onde a identificação correta de não-eventos é crucial.

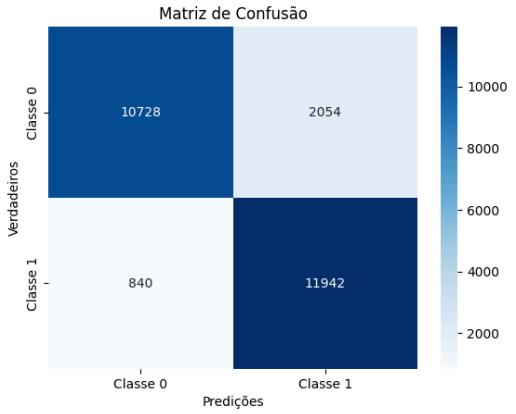


**Figura 15. Matriz de confusão da Rede Neural com janela de tempo de 10 minutos**

As métricas de perda, acurácia, *recall*, precisão e AUC nos conjuntos de treino e teste são todas elevadas, evidenciando a eficiência do modelo. O AUC, tanto no treino quanto no teste, está próximo ou acima de 0,94, indicando uma alta capacidade do modelo em diferenciar entre as classes.

### 4.3.6 Análise dos Resultados do AutoML com Janela de Tempo de 10 Minutos

A aplicação do AutoML para a janela de tempo de 10 minutos resultou em uma performance alta, como é evidenciado pelos escores de validação cruzada interna que melhoraram consistentemente ao longo das gerações. A configuração final do melhor *pipeline* utilizou um ExtraTreesClassifier com o pré-processamento via MaxAbsScaler, *bootstrap* ativado, e um conjunto de hiperparâmetros focados em maximizar a entropia, com um número relativamente alto de estimadores.

O resultado das métricas do Relatório de Classificação (Tabela 4) mostra uma precisão, *recall* e medida f1 altos para ambas as classes, com um desempenho melhor para a Classe 1 (eventos positivos). Estes resultados são indicativos de um modelo altamente capaz de prever corretamente tanto a presença, quanto a ausência de eventos.

A acurácia geral do modelo, juntamente com a média macro e ponderada das métricas, é de 0,95, o que sugere que o modelo AutoML é capaz de fazer previsões muito confiáveis.

A Matriz de Confusão evidenciada na Figura 16 complementa esses achados, mostrando um número muito alto de verdadeiros positivos e verdadeiros negativos, com apenas uma pequena proporção de falsos positivos e falsos negativos. Este padrão na Matriz de Confusão reafirma a habilidade do modelo em classificar corretamente as instâncias da maioria das vezes.

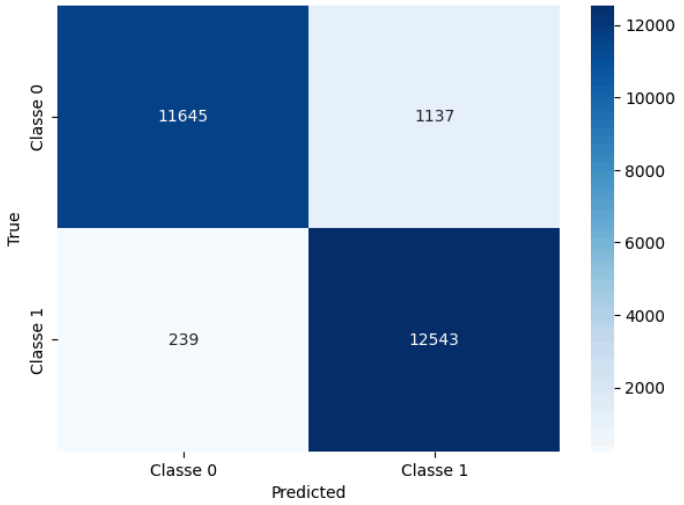


**Figura 16. Matriz de confusão do AutoML com janela de tempo de 10 minutos.**

## 5. Conclusão

É inegável que a seleção de características e a subsequente engenharia de atributos tem um impacto positivo sobre a importância das variáveis nos modelos. Ao refinar os dados de entrada, os modelos tornam-se mais adeptos a discernir os atributos relevantes dos irrelevantes, o que é vital na previsão de eventos tão dinâmicos quanto os esportes ao vivo.

Em suma, os resultados enfatizam a complexidade da modelagem preditiva no contexto esportivo e a necessidade de uma engenharia de dados sistemática para extrair o verdadeiro valor das estatísticas de partidas ao vivo. A análise das características em diferentes janelas de tempo demonstra a interconexão entre o volume de dados, a relevância das características e a precisão do modelo, estabelecendo uma base sólida para a previsão de eventos esportivos em tempo real.

Os modelos de redes neurais demonstraram capacidade substancial de aprendizado e adaptação, com melhorias significativas nas métricas de desempenho à medida que a quantidade de dados disponíveis aumentava com janelas de tempo maiores. A precisão, o *recall* e a medida F1 melhoraram progressivamente, refletindo a importância de uma janela de tempo adequada na captura das nuances e padrões dos dados esportivos. A inclusão de múltiplas camadas e ajustes finos nos hiperparâmetros foram cruciais para alcançar um equilíbrio entre a sensibilidade e a especificidade das previsões.

Os algoritmos de AutoML, especialmente o ExtraTreesClassifier, emergiram como ferramentas poderosas, oferecendo modelos precisos, que são capazes de generalizar bem a partir dos dados de treino. O AutoML destacou-se pela sua capacidade de automatizar a seleção e otimização do modelo, o que resultou em pontuações de validação cruzada altas e métricas de classificação robustas. A consistência dos escores de validação cruzada e a precisão das previsões ressaltam o potencial do AutoML em ambientes de decisão rápida e análise preditiva.

Através da análise comparativa, foi evidente que, enquanto ambos os métodos são válidos e oferecem *insights* significativos, o AutoML mostrou-se particularmente promissor, alcançando uma precisão notável e um equilíbrio entre o *recall* e a especificidade nas previsões. Este estudo reforça a viabilidade do uso de técnicas avançadas de aprendizado de máquina na previsão de eventos esportivos ao vivo, abrindo caminho para futuras pesquisas e aplicações práticas que podem se beneficiar da

automação e eficiência destes modelos preditivos.

Em última análise, este trabalho não apenas fornece uma contribuição metodológica ao campo da análise preditiva em esportes, mas também demonstra a aplicabilidade prática dessas técnicas avançadas. As descobertas aqui apresentadas podem servir como base para o desenvolvimento de sistemas de previsão em tempo real para uma variedade de aplicações, desde aprimorar a experiência de visualização de torcedores, até oferecer *insights* estratégicos para equipes e organizações esportivas.

**Referências**

Johnson, J. M. and Khoshgoftaar, T. M. (2021). Encoding techniques for high-cardinality features and ensemble learners. In 2021 IEEE 22nd international conference on information reuse and integration for data science (IRI), pages 355–361. IEEE.

Gong, B., Cui, Y., Gai, Y., Yi, Q., e Gómez, M.-Á. (2019). The Validity and Reliability of Live Football Match Statistics From Champdas Master Match Analysis System.

Lunelli, L. M., e Castelli, M. (2019). Previsão de Resultado de Jogos da NBA com Algoritmos de Machine Learning.

Stival, L., e Dias, U. M. (2022). Análise preditiva de fatores que influenciam na entrada na zona de finalização em partidas de futebol.

Pykes, K. (2022). Sports Analytics: How Different Sports Use Data Analytics. DataCamp. Disponível em: https://www.datacamp.com/blog/sports-analytics-how-different-sports-use-data-analysis

Education Dynamics (2024). Sports Analytics: Revolutionizing Decision-Making in Sports Management. Carson-Newman University. Disponível em: https://www.cn.edu/cps-blog/sports-analytics-revolutionizing-decision-making-in-sports-management

Liang, J.; Meyerson, E.; Hodjat, B. et al. Evolutionary neural AutoML for deep learning. Proceedings of the Genetic and Evolutionary Computation Conference. DOI: 10.1145/3321707.3321721. 18 fev. 2019.

Wang, H.; Zhou, Z. Unifying attribute splitting criteria of decision trees by Tsallis entropy. Machine Learning and Knowledge Extraction, v. 1, n. 1, p. 1-15, 2021. DOI: 10.3390/make1030018.

Strobl, C.; Boulesteix, A.-L.; Zeileis, A.; Hothorn, T. Unbiased split selection for classification trees based on the Gini Index. Statistics and Computing, v. 17, n. 3, p. 219-236, 2007. DOI: 10.1007/s11222-006-8026-2.

Louppe, G. Understanding Random Forests: From Theory to Practice. arXiv: Machine Learning, 28 jul. 2014. Disponível em: https://consensus.app/papers/understanding-random-forests-from-theory-practice-louppe/49aa5a3434db5c61b4bb10ecf2b74ba1/.

SINGH, V. K.; JOSHI, K. Automated Machine Learning (AutoML): an overview of opportunities for application and research. Journal of Information Technology Case and Application Research, v. 24, p. 75-85, 2022. Disponível em: https://consensus.app/papers/automated-machine-learning-automl-overview-singh/c0

748d4efa945bd4b17468eb2e5c65ba/. Acesso em: 17 abr. 2024.

HE, X.; ZHAO, K.; CHU, X. AutoML: A Survey of the State-of-the-Art. ArXiv, 2019. Disponível em: https://consensus.app/papers/automl-survey-stateoftheart-he/77d1e99f45ce5d0783d66451c7d6525f/. Acesso em: 17 abr. 2024.