

UNIVERSIDADE FEDERAL DE UBERLÂNDIA

João Vitor Santos Izidoro

**Predição de Resultados do Campeonato
Brasileiro de Futebol com Aprendizado de
Máquina: Comparação de Modelos e Análise de
Decisões**

Uberlândia, Brasil

2025

UNIVERSIDADE FEDERAL DE UBERLÂNDIA

João Vitor Santos Izidoro

Predição de Resultados do Campeonato Brasileiro de Futebol com Aprendizado de Máquina: Comparação de Modelos e Análise de Decisões

Trabalho de conclusão de curso apresentado à Faculdade de Computação da Universidade Federal de Uberlândia, como parte dos requisitos exigidos para a obtenção título de Bacharel em Sistemas de Informação.

Orientador: Prof. Dr. Rodrigo Sanches Miani

Universidade Federal de Uberlândia – UFU

Faculdade de Computação

Bacharelado em Sistemas de Informação

Uberlândia, Brasil

2025

João Vitor Santos Izidoro

Predição de Resultados do Campeonato Brasileiro de Futebol com Aprendizado de Máquina: Comparação de Modelos e Análise de Decisões

Trabalho de conclusão de curso apresentado à Faculdade de Computação da Universidade Federal de Uberlândia, como parte dos requisitos exigidos para a obtenção título de Bacharel em Sistemas de Informação.

Trabalho aprovado. Uberlândia, Brasil, 14 de Maio de 2025:

Prof. Dr. Rodrigo Sanches Miani
Orientador

Prof. Dr. Bruno Augusto Nassif
Travencolo

Prof. Dr. Renato Aparecido Pimentel
da Silva

Uberlândia, Brasil
2025

Dedico este trabalho aos meus pais, Maria Inês e Jorge, por sempre estarem ao meu lado, por me apoiarem em todos os momentos e por me orientarem ao longo da trajetória. Agradeço, também, por nunca medirem esforços para me proporcionar um ensino de qualidade. Foram, ao longo de todos esses anos, meu alicerce e exemplos de força e perseverança.

Agradecimentos

Agradeço, primeiramente, a Deus. Ele foi a fonte de toda força e sustento que me guiou durante a trajetória, mantendo-me firme em seus propósitos e atendendo-me todas as vezes que a Ele recorri.

Agradeço, à minha família e aos amigos, que celebraram comigo o ingresso na universidade e me ofereceram apoio contínuo ao longo de toda a caminhada. Em especial, agradeço à minha avó Teresinha, que foi uma referência de ser humano e me apoiou incondicionalmente.

Expresso minha gratidão aos amigos da faculdade, com quem compartilhei a jornada acadêmica, tanto nos momentos de alegria quanto nos de dificuldade, sempre repletos de aprendizado.

Agradeço a todos os professores com os quais tive o privilégio de aprender durante a graduação. Sou grato pelo empenho e dedicação de cada um no processo de ensino. Em particular, agradeço ao meu orientador, Professor Doutor Rodrigo Sanches Miani, pela atenção e apoio durante o desenvolvimento deste trabalho, bem como pela oportunidade de pesquisar um tema que tanto me interessa.

Resumo

Este trabalho propôs a construção de modelos preditivos para prever resultados em partidas de futebol, utilizando algoritmos de aprendizado de máquina em dois cenários distintos: com e sem a possibilidade de empates. No cenário com empates, os modelos *Logistic Regression* e XGBoost apresentaram acurácia de 75%, sendo que o XGBoost também se destacou no *F1-score*. O LightGBM obteve o maior número de verdadeiros positivos para vitórias do time visitante, embora com elevado número de falsos positivos. A análise com SHAP possibilitou a identificação de variáveis determinantes, como *odds* e valores de mercado das equipes, que influenciam diretamente nas decisões dos modelos. Após a seleção das variáveis mais relevantes, observou-se leve aprimoramento no desempenho do LightGBM. No cenário sem empates, todas as métricas foram favorecidas: a Regressão Logística manteve-se como a mais precisa, o LightGBM apresentou o melhor *F1-score* e menor desvio padrão, e o XGBoost destacou-se pela identificação equilibrada das vitórias visitantes. Os resultados evidenciam que a formulação do problema impacta significativamente o desempenho dos modelos e que a interpretação dos dados é fundamental para aprimorar a capacidade preditiva das abordagens utilizadas.

Palavras-chave: Aprendizado de máquina, Previsão de resultados, Futebol, Análise de desempenho.

Lista de ilustrações

Figura 1 – Representação do processo de aprendizado sequencial de um modelo LGBM. Fonte: Adaptado de (ZHANG et al., 2021).	15
Figura 2 – Representação da matriz de confusão. Fonte: Adaptado de (ABDEL-RAOUF et al., 2023).	19
Figura 3 – Volumetria - cenário com empates.	27
Figura 4 – Volumetria - cenário sem empates.	28
Figura 5 – Desempenho dos modelos representado pelas matrizes de confusão (cenário com empates).	32
Figura 6 – Comparação das métricas dos modelos por meio de gráfico de barras (considerando empates).	33
Figura 7 – Plotagem do gráfico sumário com os valores <i>SHAP</i> .	34
Figura 8 – Matriz de Confusão modelo LightGBM após novo treinamento e testes com empates.	35
Figura 9 – Desempenho dos modelos representado pelas matrizes de confusão (cenário sem empates).	37
Figura 10 – Comparação das métricas dos modelos por meio de gráfico de barras (não considerando empates).	38
Figura 11 – Plotagem do gráfico sumário com os valores <i>SHAP</i> .	39
Figura 12 – Matriz de Confusão modelo LightGBM após novo treinamento e testes sem empates.	40

Lista de tabelas

Tabela 1	– Atributos do <i>dataset</i> e suas descrições	26
Tabela 2	– Atributo <i>vencedor</i> transformado - conjunto de dados com empate . . .	28
Tabela 3	– Atributo <i>vencedor</i> transformado - conjunto de dados sem empate . . .	28
Tabela 4	– Desempenho dos modelos preditivos utilizando dados de treinamento - cenário com empates	31
Tabela 5	– Desempenho dos modelo preditivo LightGBM utilizando dados de teste - cenário com empates	35
Tabela 6	– Desempenho dos modelos preditivos utilizando dados de treinamento - cenário sem empates	36
Tabela 7	– Desempenho dos modelo preditivo LightGBM utilizando dados de teste - cenário sem empates	40

Lista de abreviaturas e siglas

AM	Aprendizado de Máquina
LR	<i>Logistic Regression</i> (Regressão Logística)
RF	<i>Random Forest</i> (Floresta Aleatória)
XGB	<i>Extreme Gradient Boosting</i> (XGBoost)
NB	<i>Naive Bayes</i> (Bayesiano Ingênuo)
KNN	<i>k-Nearest Neighbors</i> (K-ésimo Vizinho mais Próximo)
SVM	<i>Support Vector Machine</i> (Máquina de Vetores de Suporte)
LGBM	<i>Light Gradient-Boosting Machine</i> (LightGBM)
SHAP	<i>SHapley Additive exPlanations</i> (Explicações sobre aditivos SHapley)
GOSS	<i>Gradient-based One-Side Sampling</i> (Amostragem Unilateral Baseada em Gradiente)
EFB	<i>Exclusive Feature Bundling</i> (Agrupamento de Recursos Exclusivos)
FP	Falso positivo
FN	Falso negativo
VP	Verdadeiro positivo
VN	Verdadeiro negativo

Sumário

1	INTRODUÇÃO	11
1.1	Objetivos	12
1.1.1	Objetivo geral	12
1.1.2	Objetivos específicos	12
1.2	Organização do Trabalho	12
2	FUNDAMENTAÇÃO TEÓRICA	13
2.1	Campeonato de Pontos Corridos	13
2.2	Aprendizado de máquina	13
2.3	Algoritmos de aprendizado de máquina	14
2.3.1	LightGBM (LGBM)	14
2.3.2	XGBoost (XGB)	15
2.3.3	K-Nearest Neighbors (KNN)	16
2.3.4	Regressão Logística (LR)	16
2.3.5	Random Forest (RF)	16
2.3.6	Naive Bayes (NB)	17
2.3.7	Máquina de Vetores de Suporte (SVM)	17
2.4	Interpretabilidade de Modelos de Aprendizado de Máquina	17
2.4.1	SHapley Additive exPlanations (SHAP)	18
2.5	Avaliação de modelos de aprendizado de máquina	18
2.5.1	Matriz de confusão	19
2.5.2	Métricas	20
2.5.2.1	Acurácia	20
2.5.2.2	Precisão	20
2.5.2.3	Recall	20
2.5.2.4	F1-score	20
2.6	Trabalhos Relacionados	20
3	DESENVOLVIMENTO	23
3.1	Fases do Desenvolvimento	23
3.2	Coleta e Fonte dos Dados	24
3.3	Pré-processamento dos Dados	26
3.4	Modelagem e Divisão da Base de Dados	28
3.5	Avaliação e Interpretação	29
4	RESULTADOS	30

4.1	Visão Geral dos Cenários Avaliados	30
4.2	Resultados do Cenário A – Com Empates	31
4.2.1	Treinamento	31
4.2.2	Teste	31
4.2.3	Interpretação com SHAP	33
4.3	Resultados do Cenário B – Sem Empates	35
4.3.1	Treinamento	36
4.3.2	Teste	36
4.3.3	Interpretação com SHAP	38
4.4	Discussão dos resultados	40
5	CONCLUSÃO	43
	REFERÊNCIAS	44

1 Introdução

O futebol é compreendido como mais do que um esporte por seus praticantes e entusiastas. No contexto nacional, constitui-se como a modalidade esportiva de maior popularidade, capaz de provocar intensa comoção e de consolidar um sentimento cultural transmitido ao longo das gerações. Considerando a relevância que a prática futebolística assume no país, observa-se crescente interesse na previsão de resultados de partidas, tanto por parte dos adeptos quanto das organizações que possuem equipes inseridas nas competições, movimentando, assim, um mercado milionário ([VALUE, 2023](#)).

No contexto do futebol, esporte que desperta emoções entre seus adeptos, a previsão do resultado de uma partida transcende a análise puramente lógica e estatística. Para os torcedores, fatores como uma sequência positiva de resultados, o retrospecto favorável ou o jogo contra o rival, são considerados suficientes para indicar o provável vencedor, ainda que, por vezes, ocorra a vitória da equipe considerada tecnicamente inferior. Com o aumento da disponibilidade de informações, torna-se possível acompanhar o ambiente interno dos clubes, evidenciando que aspectos subjetivos, como a motivação individual ou coletiva dos atletas, podem influenciar o desfecho das partidas, desafiando previsões baseadas exclusivamente em dados objetivos.

A previsão do desempenho de uma equipe ao longo da temporada está intrinsecamente relacionada ao seu aproveitamento em cada partida, refletindo diretamente nos investimentos e no planejamento para o ano seguinte. Trata-se, portanto, de um tema de grande relevância, dada sua capacidade de impactar a geração de capital. De acordo com dados publicados por [GE \(2019\)](#), os quatro clubes com maior receita na temporada de 2018 foram: Palmeiras (R\$ 654 milhões), Flamengo (R\$ 536 milhões), São Paulo (R\$ 339 milhões) e Corinthians (R\$ 389 milhões).

Com o aumento da popularidade das previsões no esporte, já existem estudos relacionados à previsão esportiva, como os desenvolvidos por [Bernardes \(2023\)](#) e [DEUS \(2019\)](#), que utilizaram métodos de Aprendizado de Máquina em suas abordagens. O primeiro trata da predição da fase final do campeonato de futebol americano, denominada *playoffs*, com foco na utilização de uma volumetria maior de atributos estatísticos. O segundo, por sua vez, aborda a predição de partidas em campeonatos de futebol, com ênfase na construção de um conjunto de dados balanceados. Além disso, o estudo proposto por [Kleina \(2021\)](#) emprega Redes Neurais Artificiais para prever a classificação final dos times do campeonato, organizando-os em grupos de acordo com prováveis metas ou premiações alcançadas ao término de suas campanhas, apresentando resultados relevantes, como 42,94% de sucesso e 47,46% de erro.

Assim, este estudo propõe a análise comparativa de modelos preditivos de aprendizado de máquina aplicados à previsão de vitórias de equipes visitantes no Campeonato Brasileiro da Série A, utilizando dados históricos compreendidos entre 2018 e 2023. Busca-se, com isso, aproximar-se de um modelo preditivo ideal para a referida tarefa, mediante a identificação dos atributos mais relevantes, a apresentação dos resultados obtidos e a avaliação crítica da eficácia dos modelos analisados.

1.1 Objetivos

1.1.1 Objetivo geral

Este trabalho tem como objetivo geral avaliar, por meio de métricas de desempenho, os algoritmos de predição Regressão Logística, *Random Forest*, *k-Nearest Neighbors*, *Naive Bayes*, Máquina de Vetores de Suporte, XGBoost e LightGBM, aplicados a uma base de dados de resultados de futebol, com o intuito de prever o vencedor das partidas. Ademais, propõe-se analisar o modelo LightGBM utilizando a técnica *SHapley Additive exPlanations*, a fim de interpretar as variáveis mais relevantes para a predição.

1.1.2 Objetivos específicos

- Coletar e organizar uma base com dados estatísticos das partidas da Série A do Campeonato Brasileiro.
- Construir modelos preditivos utilizando diferentes algoritmos de aprendizado de máquina presentes da biblioteca *Scikit-learn* do *Python*.
- Analisar a eficácia dos modelos por meio das métricas Acurácia, Precisão, *Recall* e *F1-Score*.
- Aplicar a técnica SHAP para interpretar as variáveis mais relevantes no modelo LGBM.

1.2 Organização do Trabalho

O trabalho está organizado da seguinte forma: o Capítulo 2 é dedicado à apresentação das definições e dos conceitos fundamentais utilizados, tais como Campeonato de Pontos Corridos, Aprendizado de Máquina, algoritmos e métricas empregadas. O Capítulo 3 aborda o desenvolvimento do trabalho, desde a definição do conjunto de dados até a construção dos modelos preditivos. No Capítulo 4, são discutidos os resultados obtidos pelos modelos e a interpretação das variáveis utilizadas. Por fim, o Capítulo 5 apresenta as conclusões e sugestões para trabalhos futuros.

2 Fundamentação Teórica

O capítulo tem como objetivo apresentar os fundamentos teóricos e os trabalhos relacionados que embasam o desenvolvimento deste estudo. Inicialmente, são discutidos os principais conceitos necessários para a compreensão da temática abordada, o funcionamento dos modelos de Aprendizado de Máquina e os algoritmos aplicados à predição de resultados esportivos, destacando suas características e aplicações no contexto da ciência de dados.

Na sequência, são descritas as métricas de avaliação utilizadas. Por fim, são apresentados trabalhos anteriores que exploraram abordagens semelhantes, servindo como base comparativa e justificativa para a metodologia adotada neste estudo.

2.1 Campeonato de Pontos Corridos

O Campeonato Brasileiro de Futebol passou por diferentes formatos ao longo de sua história. Até 2002, o sistema adotado era o de eliminatórias (mata-mata), caracterizado por controvérsias relacionadas a rebaixamentos e limitações na disputa pelo título.

A partir de 2003, a Confederação Brasileira de Futebol (CBF) implementou o sistema de pontos corridos no Campeonato Brasileiro. Nesse modelo, os times se enfrentam em dois turnos, como mandantes e visitantes, e a classificação é determinada pela soma de pontos obtidos ao longo da competição: a vitória concede três pontos, o empate concede um ponto e a derrota não concede pontos ([CNN Brasil, 2025](#)).

A implantação do sistema de pontos corridos no Campeonato Brasileiro promoveu a padronização da competição em todas as divisões, influenciando sua organização tanto por parte dos clubes quanto da entidade que a administra, proporcionando maior equidade e previsibilidade ([TUESTA et al., 2017](#)).

2.2 Aprendizado de máquina

O Aprendizado de Máquina (AM) está associado ao uso de métodos computacionais que possibilitam o tratamento e o processamento de grandes volumes de dados, cuja manipulação manual demandaria elevado custo de tempo e recursos. Esse processamento pode ser aplicado a diversas áreas do cotidiano, como web, saúde, finanças, marketing, varejo, esportes, entre outras.

De acordo com [Faceli et al. \(2021\)](#), os algoritmos projetados para AM são catego-

rizados em três grupos principais, sendo o primeiro deles o Aprendizado Supervisionado, onde os algoritmos tem como objetivo central a predição de resultados futuros a partir do mapeamento de uma entrada de dados rotulados que serão usados para o treinamento e resultar em uma saída prevista relacionada. Já o Aprendizado Não Supervisionado, visa uma análise descritiva, buscando agrupamentos que podem estar ou não ocultos, identificando assim uma semelhança nos dados de entrada. Por fim, o Aprendizado por Reforço está relacionado a tomada de decisão ótima em tempo real, onde durante o seu aprendizado contínuo, entende e toma as decisões visando a melhor recompensa por ele decidida.

A classificação de modelos de Aprendizado Supervisionado, onde a entrada possui uma saída conhecida, é o modelo ideal a ser utilizado em cenários esportivos, pois para sua utilização contempla um espaço de hipóteses que é fornecido aos modelos no momento de treinamento após o pré processamento de dados. Possuem uma função objetivo que é determinada a otimizar as predições da atributo alvo e a partir das métricas fornecidas é possível realizar uma avaliação do modelo, identificando quais pontos podem ser melhor balanceados para alcançar resultados superiores.

2.3 Algoritmos de aprendizado de máquina

Os algoritmos de aprendizado de máquina desempenham um papel central no campo da computação e da análise de dados, sendo responsáveis pela capacidade dos sistemas de identificar padrões e realizar previsões a partir de conjuntos de dados. Esses algoritmos possibilitam a automatização de tarefas complexas que, tradicionalmente, exigiriam intervenção humana intensiva, promovendo avanços significativos em diversas áreas do conhecimento.

Nesta seção, serão apresentados os modelos de Aprendizado Supervisionado implementados, com ênfase na descrição de seu funcionamento e de sua estruturação interna. O objetivo é oferecer uma compreensão dos princípios que regem esses algoritmos.

2.3.1 LightGBM (LGBM)

De acordo com [Ke et al. \(2017\)](#), o modelo LightGBM (LGBM) tem se tornado popular na implementação de algoritmos de aprendizado de máquina, destacando-se especialmente por sua escalabilidade e eficiência em tarefas de classificação. Desenvolvido para lidar com grandes volumes de dados e altos requisitos computacionais, o LGBM busca otimizar tanto o tempo de processamento quanto o consumo de recursos. Ademais, o modelo é amplamente reconhecido por sua capacidade de manter altos níveis de acurácia, mesmo em cenários que envolvem conjuntos de dados extensos e complexos.

O LightGBM adota uma estratégia de crescimento de árvore orientada por folhas (*leaf-wise*), conforme ilustrado na Figura 1, em contraste com o tradicional crescimento orientado por níveis (*level-wise*). Conforme apresentado por [Ke et al. \(2017\)](#), essa abordagem pode acelerar o processamento em até 20 vezes. Tal desempenho é alcançado mediante a implementação de técnicas como a Amostragem Unilateral Baseada em Gradiente (GOSS) e o Agrupamento de Recursos Exclusivos (EFB). A técnica GOSS elimina instâncias de dados com gradientes pequenos, concentrando-se apenas naquelas com gradientes maiores, o que permite obter uma estimativa precisa do ganho de informação com menor custo computacional. Por sua vez, a técnica EFB realiza o agrupamento de características pouco relevantes, reduzindo o número de campos com valor zero e, consequentemente, otimizando o treinamento do modelo. Com a adoção dessas estratégias, o LGBM consegue reduzir de forma mais eficiente a função de perda em comparação aos métodos tradicionais de crescimento por nível.

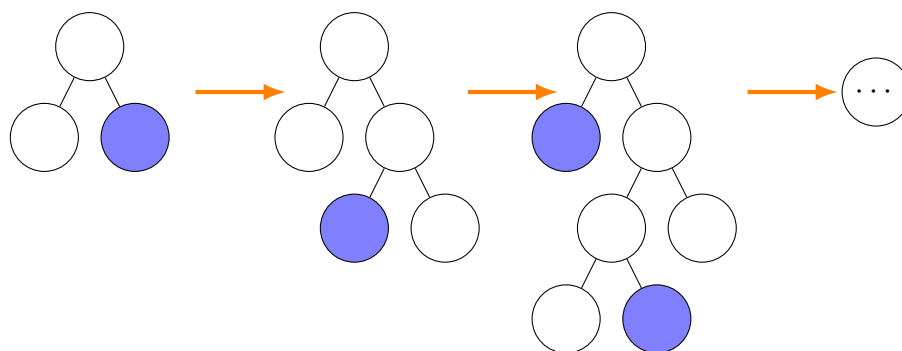


Figura 1 – Representação do processo de aprendizado sequencial de um modelo LGBM.
Fonte: Adaptado de ([ZHANG et al., 2021](#)).

2.3.2 XGBoost (XGB)

Algoritmo de *Boosting* amplamente utilizado em aprendizado de máquina baseado em *Gradient Boosting Decision Tree* (GBDT). o modelo desenvolvido apresentado em ([CHEN; GUESTRIN, 2016](#)) se destaca pela eficiência computacional, evoluindo modelos anteriores e sendo possível a utilização de implementação em paralelismo.

Baseado em modelo de árvore de decisão, o XGBoost tem a proposta de construir modelos preditivos a partir da sequência de árvores de regressão, sequência essa distribuída em etapas sucessivas. Onde a primeira é a realização de uma predição da atributo alvo utilizando uma árvore simples construída com a entrada de dados do modelo, em sequência é realizado o cálculo do resíduo (erro) entre a predição e o valor real. Uma nova árvore é ajustada apenas com esses resíduos, realizando assim uma nova predição de árvores menores.

Com a realização de criação e predição sequencial de novas árvores, no final é possível criar um modelo robusto onde a predição final é obtida pela soma ponderada das

predições de todas as árvores. Um diferencial deste modelo para os outros *Boosting* é que introduz técnicas de regularização.

2.3.3 K-Nearest Neighbors (KNN)

Baseado no princípio de que elementos semelhantes tendem a pertencer à mesma classe, o *K-Nearest Neighbors* (KNN) é considerado um dos métodos mais simples de classificação. Ao receber uma nova amostra, o algoritmo seleciona um número k de amostras previamente rotuladas (vizinhos) para realizar a classificação. A escolha do valor de k influencia diretamente a precisão do modelo, uma vez que valores muito baixos podem resultar em classificações imprecisas, devido à consideração de poucos pontos de diferentes classes. Após essa seleção, é realizado o cálculo da distância euclidiana entre a nova amostra e seus vizinhos mais próximos, determinando, assim, a classe da entrada com base na maioria dos votos (SUYAL; GOYAL, 2022).

O KNN destaca-se por sua facilidade de implementação e adaptação a novos dados, características que o tornam amplamente utilizado em diversas aplicações. Entretanto, para que o modelo atinja uma classificação mais precisa e eficaz, faz-se necessário um processo adequado de normalização e balanceamento dos dados de entrada (TAN; STEINBACH; KUMAR, 2018). Sem essas etapas de pré-processamento, a performance do KNN pode ser significativamente comprometida, especialmente em conjuntos de dados com variáveis de escalas distintas ou desbalanceados entre classes.

2.3.4 Regressão Logística (LR)

A regressão logística é um modelo amplamente utilizado em tarefas de classificação binária, destacando-se por sua simplicidade e elevada capacidade interpretativa. Trata-se de um classificador discriminativo, uma vez que estima diretamente a probabilidade condicional de uma classe com base nos atributos observados (NG; JORDAN, 2002).

Seu funcionamento baseia-se na combinação linear de variáveis independentes, ponderadas por coeficientes ajustados durante o treinamento, sendo esta combinação posteriormente transformada por meio da função sigmoide, o que permite mapear o resultado para um intervalo entre 0 e 1. Essa transformação possibilita interpretar a saída como uma probabilidade associada a determinada classe (HOSMER; LEMESHOW; STURDIVANT, 2013).

2.3.5 Random Forest (RF)

O *Random Forest* é um método de aprendizado de máquina baseado em árvores de decisão, proposto por Breiman (2001), que se destaca pela sua robustez, capacidade de generalização e desempenho em tarefas de classificação e regressão. A técnica consiste

na construção de múltiplas árvores de decisão durante o treinamento, cujo resultado final é obtido por meio da agregação das previsões individuais, como votação majoritária para classificação ou média para regressão.

Além de sua eficiência, o RF apresenta a vantagem de fornecer medidas internas de importância das variáveis, o que facilita a interpretação dos resultados. Essa característica tem sido amplamente explorada em diversas áreas, como evidenciado por [Cutler et al. \(2007\)](#).

2.3.6 Naive Bayes (NB)

Baseado no **Teorema de Bayes**, que consiste em observar uma relação entre as probabilidades que antecedem e sucede o evento. Trata-se de um classificador probabilístico amplamente utilizado em algoritmos de aprendizado supervisionado, cuja simplicidade de cálculo decorre do princípio da independência condicional entre as variáveis preditoras ([MANNING; RAGHAVAN; SCHÜTZE, 2008](#)).

O *Naive Bayes* é particularmente valorizado em cenários que demandam modelos rápidos, interpretáveis e eficientes, especialmente quando se trabalha com conjuntos de dados reduzidos ou em ambientes de alta dimensionalidade ([MANNING; RAGHAVAN; SCHÜTZE, 2008](#)). Sua capacidade de apresentar boas performances mesmo sob condições restritas torna-o uma escolha frequente em aplicações como filtragem de *spam*, análise de sentimentos e classificação de documentos.

2.3.7 Máquina de Vetores de Suporte (SVM)

A Máquina de Vetores de Suporte (SVM) destaca-se por sua capacidade de tratar problemas não linearmente separáveis, transformando o espaço de entrada original em um espaço de dimensões superiores. Dessa forma, torna-se possível realizar a separação linear entre as classes ([SCHÖLKOPF; SMOLA; MÜLLER, 1997](#)).

Após a transformação do espaço de entrada, o algoritmo busca encontrar um hiperplano ótimo que separe as classes com a maior margem possível ([CORTES; VAPNIK, 1995](#)). Esse hiperplano é definido pelos vetores de suporte, que correspondem aos pontos mais próximos às margens das classes e que exercem influência direta na determinação da fronteira de decisão.

2.4 Interpretabilidade de Modelos de Aprendizado de Máquina

A interpretabilidade permite não apenas validar o comportamento do modelo frente a critérios éticos e legais, mas também identificar possíveis vieses, melhorar a confiança dos usuários e facilitar a detecção de falhas. Esta seção discute um dos principais

conceitos e abordagens relacionadas à interpretabilidade, destacando sua importância e os desafios ainda presentes na área.

2.4.1 SHapley Additive exPlanations (SHAP)

A técnica *SHAP* fundamenta-se nos valores de *Shapley*, oriundos da teoria dos jogos cooperativos, para mensurar a contribuição individual de cada variável de entrada na geração de uma previsão. Essa metodologia possibilita uma interpretação aprofundada do comportamento do modelo, permitindo avaliar a influência de cada atributo tanto no panorama geral quanto em casos específicos. Com isso, torna-se mais viável compreender as decisões automatizadas, favorecendo a transparência para especialistas e demais usuários envolvidos no processo.

A interpretabilidade em modelos de aprendizado de máquina é um tema amplamente discutido na literatura. [Rudin et al. \(2021\)](#) abordam os principais desafios e fundamentos necessários para garantir que os modelos sejam compreensíveis, especialmente em aplicações críticas. Complementando essa perspectiva, [Murdoch et al. \(2019\)](#) propõem estruturas conceituais que auxiliam na avaliação da qualidade das explicações geradas por diferentes técnicas interpretativas. Por sua vez, [Fryer, Strumbelj e Bosnic \(2021\)](#) analisam as limitações da utilização de valores de Shapley, alertando para cuidados específicos ao aplicar esse método na seleção de atributos. No cenário brasileiro, [Caires \(2022\)](#) apresenta um estudo prático sobre o uso do SHAP em sistemas de análise de crédito e prevenção à fraude, demonstrando sua eficácia na interpretação de decisões realizadas por modelos preditivos.

2.5 Avaliação de modelos de aprendizado de máquina

A avaliação de modelos de aprendizado de máquina é uma etapa fundamental no processo de desenvolvimento de soluções baseadas em dados. Através de métricas apropriadas e procedimentos sistemáticos de validação, é possível mensurar a capacidade do modelo de generalizar para dados não vistos, garantindo a eficácia e a robustez de suas previsões em cenários reais.

A escolha dos métodos de avaliação deve considerar a natureza do problema, as características do conjunto de dados e os objetivos específicos da aplicação. Esta seção discute os principais conceitos, métricas e estratégias utilizados para a avaliação de modelos, ressaltando sua importância para a construção de sistemas confiáveis e eficientes.

2.5.1 Matriz de confusão

A Matriz de Confusão é um método que permite a visualização do desempenho dos modelos, comparando classificações previstas de reais. Tem em sua composição quatro principais grupos que são: Verdadeiros positivos (VP), que significa que o modelo previu corretamente a classe positiva. Verdadeiros Negativos (VN), ocorre quando o modelo previu corretamente uma classe negativa. Falso Positivo (FP) quando previu como positivo, mas era negativo e Falso Negativo (FN), quando o foi previsto que era negativo e o resultado é positivo (POWERS, 2011).

Contribuindo para uma avaliação mais justa e criteriosa do modelo, a matriz de confusão permite que ocorra uma avaliação detalhada principalmente em classes de dados desbalanceados, uma vez que os valores de métricas como acurácia pode induzir a uma conclusão errônea sobre o modelo avaliado. Podendo identificar essas deficiências das outras métricas apresentadas, a matriz pode fornecer uma avaliação precisa (SOKOLOVA; LAPALME, 2009).

Uma interpretação adequada da matriz de confusão pode orientar ajustes nos modelos ou no processo de coleta de dados, resultando em melhorias no desempenho do aprendizado de máquina. A Figura 2 apresenta uma representação de como a matriz organiza e exibe os resultados obtidos.

Classe Real	Negativo	Positivo
	Verdadeiro Neg	Falso Pos
Positivo	Falso Neg	Verdadeiro Pos
		Classe Predita
		Negativo Positivo

Figura 2 – Representação da matriz de confusão. **Fonte:** Adaptado de (ABDELRAOUF et al., 2023).

2.5.2 Métricas

Nesta seção, descrevem-se as métricas adotadas para a avaliação dos modelos de aprendizado de máquina utilizados na predição dos resultados dos jogos. A escolha e a interpretação adequadas dessas métricas são essenciais para garantir uma análise rigorosa do desempenho dos modelos, possibilitando a identificação de seus pontos fortes e limitações.

2.5.2.1 Acurácia

A Acurácia é obtida por meio da razão entre número de previsões corretas (VP + VN) e o total de amostras (VP + VN + FP + FN). Com um valor resultante entre 0 e 1, é um indicativo que mede a qualidade do modelo de forma geral, porém, podendo gerar resultados enganosos em conjuntos desbalanceados ([POWERS, 2011](#)).

2.5.2.2 Precisão

A Precisão é o valor resultante da razão entre a quantidade de verdadeiros positivos (VP) e o somatório de verdadeiros positivos e falso negativos (VP + FN). Esse valor resultante apresenta a eficiência do modelo em não prever falso positivo ([MARCHI; FONSECA; BODÊ, 2023](#)).

2.5.2.3 Recall

O *Recall* é a métrica que apresenta a capacidade do modelo em identificar corretamente todas as instâncias positivas. O valor de *Recall* é obtido da razão entre verdadeiros positivos (VP) e a somatória de verdadeiros positivos e falso negativos (VP + FN) ([MARCHI; FONSECA; BODÊ, 2023](#)).

2.5.2.4 F1-score

O *F1-score* apresenta a média harmônica entre precisão e *recall*, o que proporciona uma avaliação mais equilibrada do modelo. É calculado pela fórmula $2 * ((\text{Precisão} * \text{Recall}) / (\text{Precisão} + \text{Recall}))$ ([YACOUBY; AXMAN, 2020](#)).

2.6 Trabalhos Relacionados

Esta seção apresenta uma revisão concisa dos principais estudos que abordam temas semelhantes aos discutidos neste trabalho, com o objetivo de situar as abordagens adotadas e as contribuições existentes. A motivação para a seleção desses trabalhos relaciona-se, em grande parte, com a busca por identificar lacunas no conhecimento exis-

tente e avaliar como as técnicas e os modelos empregados podem ser aprimorados ou aplicados de forma inovadora em contextos específicos.

[Oliveira \(2024\)](#) conduziu um trabalho voltado à aplicação de algoritmos de aprendizado de máquina na previsão de resultados do Campeonato Brasileiro de Futebol. A dissertação teve como objetivo avaliar e comparar o desempenho de modelos supervisionados, como *Random Forest*, *K-Nearest Neighbors* e *Support Vector Machines*, utilizando variáveis estatísticas das equipes e resultados anteriores das partidas como atributos. O autor organizou um conjunto de dados com informações históricas e extraiu características relevantes, como gols marcados, mando de campo, desempenho recente das equipes e classificação na tabela, adotando um enfoque tradicional na previsão do resultado da partida, com especial atenção ao time mandante.

Além disso, o trabalho concentrou-se nas métricas de desempenho dos algoritmos, como acurácia, *recall*, precisão e *F1-score*, com o objetivo de identificar o modelo mais eficaz na tarefa preditiva. No entanto, a dissertação não explorou mecanismos de interpretabilidade dos modelos, limitando-se à análise quantitativa dos resultados. Também não houve segmentação específica de cenários com baixa probabilidade, como a vitória do time visitante, o que representa uma lacuna abordada neste trabalho.

Nesse contexto, a proposta deste estudo se diferencia ao adotar uma abordagem mais específica e aprofundada. Aqui, o foco está na previsão da vitória do time visitante, uma classe naturalmente mais desbalanceada, exigindo atenção especial na preparação dos dados e na avaliação do modelo. Além da modelagem preditiva, aplicou-se a técnica de interpretação *SHapley Additive exPlanations*, que possibilita compreender a contribuição de cada variável na decisão do modelo. Essa análise interpretativa torna os resultados mais transparentes e confiáveis, permitindo uma leitura mais rica dos fatores que influenciam o desempenho do time visitante, o que representa um avanço em relação ao trabalho anterior.

O trabalho apresentado por [Schneider \(2022\)](#) consistiu em uma análise comparativa entre diferentes algoritmos de classificação aplicados à previsão de resultados de partidas de futebol, utilizando métricas de desempenho, como acurácia e medida F1 (*F1-score*), para a avaliação dos modelos. Embora a proposta compartilhe semelhanças com o presente estudo no uso de técnicas de aprendizado de máquina e na intenção de prever o desfecho das partidas, existem diferenças relevantes entre as abordagens, este trabalho concentra-se especificamente no Campeonato Brasileiro de Futebol, incorporando atributos relacionados ao valor de mercado dos elencos e às cotações de apostas (*odds*), além de empregar técnicas de interpretabilidade, como a SHAP, com o intuito de compreender a influência das variáveis nas decisões tomadas pelos modelos. Dessa forma, o estudo contribui para a literatura ao oferecer não apenas uma comparação de desempenho entre algoritmos, mas também uma análise mais aprofundada dos fatores que impactam

diretamente as predições realizadas.

O objetivo da pesquisa de [Moustakidis et al. \(2023\)](#) consistiu na aplicação de técnicas de aprendizado de máquina combinadas com métodos de interpretabilidade, como a *SHAP*, a fim de identificar os principais fatores que influenciam o desempenho coletivo de equipes de futebol. O estudo destaca-se pelo uso de dados agregados por equipe e pela ênfase na explicabilidade das predições, o que proporciona maior transparência nos resultados gerados pelos modelos. Embora compartilhe a proposta de aplicação de modelos preditivos ao contexto do futebol, o presente trabalho diferencia-se por focar na previsão do resultado final das partidas do Campeonato Brasileiro, adotando uma abordagem direcionada ao desfecho da partida (vitória do mandante, empate ou vitória do visitante) e utilizando variáveis específicas, como as cotações de apostas (*odds*) e o valor de mercado da equipe titular. Ademais, a análise comparativa entre diferentes algoritmos de classificação e a consideração de cenários com e sem empates conferem a este estudo um enfoque mais prático na avaliação da capacidade de generalização dos modelos.

3 Desenvolvimento

O presente capítulo descreve o processo de construção e implementação dos modelos preditivos empregados na previsão dos resultados das partidas da Série A do Campeonato Brasileiro de Futebol. Inicialmente, apresentam-se os procedimentos de coleta, preparação e análise exploratória dos dados, fundamentais para a identificação das variáveis mais relevantes.

Em seguida, abordam-se as estratégias de modelagem adotadas, bem como os métodos utilizados para a otimização dos modelos. Por fim, discute-se o desempenho dos modelos obtidos, com base em métricas específicas de avaliação.

3.1 Fases do Desenvolvimento

O processo de desenvolvimento foi estruturado e conduzido conforme as seguintes etapas:

1. Coleta de dados: Essa etapa envolve a busca e a análise de diferentes conjuntos de dados disponíveis, com o objetivo de identificar aqueles que melhor se alinham aos propósitos da pesquisa. Após a identificação das bases mais relevantes, realiza-se a escolha dos atributos que exercem maior influência no problema estudado. Com base nessa seleção, procede-se à obtenção dos dados necessários, os quais serão organizados e estruturados para compor a base que servirá de suporte às etapas posteriores da investigação.
2. Pré-processamento: Etapa dedicada à aplicação de métodos para o tratamento de dados que possam comprometer a performance preditiva dos modelos ou gerar desequilíbrios no conjunto de dados em relação ao objetivo do estudo.
3. Modelagem consiste na organização do conjunto de dados, com sua divisão em subconjuntos de treino e teste, além da criação, configuração e aplicação dos algoritmos dos modelos selecionados. A predição é realizada com base no conjunto previamente definido e submetido ao pré-processamento.
4. Avaliação e Interpretação dos resultados: Compreende o uso de métricas obtidas a partir do conjunto de teste para mensurar o desempenho dos modelos, bem como a interpretação do comportamento do modelo LightGBM após a criação de uma nova instância otimizada, utilizando as variáveis consideradas mais relevantes para a predição segundo os critérios da própria modelagem.

Com o objetivo de apresentar de forma estruturada as etapas envolvidas neste trabalho. Cada uma dessas etapas será detalhada individualmente nas Seções 3.2, 3.3, 3.4 e 3.5, de modo a aprofundar a discussão sobre suas especificidades, métodos aplicados e contribuições para o desenvolvimento dos modelos preditivos.

3.2 Coleta e Fonte dos Dados

Seguindo abordagem metodológica semelhante à de Oliveira (2024), esta pesquisa emprega a base de dados *Informações e estatísticas - Brasileirão Série A* para avaliação de modelos de aprendizado de máquina. Embora a estrutura geral da análise tenha sido inspirada no estudo mencionado, foram incorporadas modificações e extensões com o objetivo de aprofundar a análise sob a ótica da predição de vitórias dos times visitantes em partidas do Campeonato Brasileiro de Futebol Série A.

Os atributos presentes na base de dados utilizada neste estudo são apresentados na Tabela 1, sendo representados por nomes padronizados no formato *snake_case*:

Nº	Nome do atributo	Descrição
1	ano_campeonato	ano corrente do Campeonato Brasileiro de Futebol Série A
2	mes_campeonato	mês corrente da partida
3	data	data da partida
4	rodada	número da rodada da partida
5	time_mandante	nome do time mandante
6	time_visitante	nome do time visitante
7	estadio	nome do estádio da partida
8	PPJ_pre_jogo_mandante	média de pontos por jogo do time mandante antes da partida começar
9	PPJ_pre_jogo_visitante	média de pontos por jogo do time visitante antes da partida começar
10	xG_pre_jogo_mandante	gols esperados da equipe mandante
11	xG_pre_jogo_visitante	gols esperados da equipe visitante
12	GPJ_pre_jogo	média de gols por jogo antes da partida começar
13	AM_porcentagem_pre_jogo	média, antes do jogo começar, da probabilidade de ambas as equipes marcarem
14	A15_porcentagem_pre_jogo	média, antes do jogo começar, da probabilidade entre ambas equipes marcarem mais de 1.5 gols na partida

Nº	Nome do atributo	Descrição
15	A25_porcentagem_pre_jogo	média, antes do jogo começar, da probabilidade entre ambas equipes marcarem mais de 2.5 gols na partida
16	A45_porcentagem_pre_jogo	média, antes do jogo começar, da probabilidade entre ambas equipes marcarem mais de 4.5 gols na partida
17	EPJ_pre_jogo	média do número de escanteios das duas equipes antes do jogo começar
18	odds_mandante_vence	média das odds, no pré-jogo, entre casas de apostas para que o time mandante vença a partida
19	odds_empate	média das odds, no pré-jogo, entre casas de apostas para que acabe com empate a partida
20	odds_visitante_vence	média das odds, no pré-jogo, entre casas de apostas para que o time visitante vença a partida
21	odds_A15	média das odds, no pré-jogo, entre casas de apostas para que haja mais de 1.5 gols no jogo
22	odds_A25	média das odds, no pré-jogo, entre casas de apostas para que haja mais de 2.5 gols no jogo
23	odds_A35	média das odds, no pré-jogo, entre casas de apostas para que haja mais de 3.5 gols no jogo
24	odds_A45	média das odds, no pré-jogo, entre casas de apostas para que haja mais de 4.5 gols no jogo
25	odds_AM_sim	média das odds, no pré-jogo, entre casas de apostas para que ambas as equipes marquem pelo menos um gol cada na partida
26	odds_AM_nao	média das odds, no pré-jogo, entre casas de apostas para que, no máximo, apenas uma equipe marque gol
27	formacao_mandante	formação tática do time mandante que começa a partida

Nº	Nome do atributo	Descrição
28	formacao_visitante	formação tática do time visitante que começa a partida
29	estado_mandante	estado brasileiro de origem do time mandante
30	estado_visitante	estado brasileiro de origem do time visitante
31	colocacao_mandante	colocação na tabela do campeonato do time mandante antes do jogo começar
32	colocacao_visitante	colocação na tabela do campeonato do time visitante antes do jogo começar
33	valor_equipe_titular_mandante	valor financeiro da equipe titular do time mandante para a partida
34	valor_equipe_titular_visitante	valor financeiro da equipe titular do time visitante para a partida
35	idade_media_titular_mandante	média da idade da equipe titular do time mandante para a partida
36	idade_media_titular_visitante	média da idade da equipe titular do time visitante para a partida
37	vencedor	mostra o resultado da partida, que pode ser mandante venceu, ou visitante venceu ou empate.

Tabela 1 – Atributos do *dataset* e suas descrições

Destacam-se, entre os atributos utilizados, as informações extraídas de bases de dados provenientes de sites que disponibilizam os valores de *odds* das casas de apostas. Esses valores já incorporam cálculos de probabilidade sobre possíveis desfechos da partida e, portanto, são relevantes no momento da predição da classe-alvo **vencedor**.

Além disso, alguns atributos mostraram-se especialmente úteis para os modelos, como, por exemplo, o `valor_equipe_titular_mandante` no campeonato corrente. Essa variável teve maior importância na avaliação do modelo LightGBM, conforme será discutido na Seção 4 de resultados.

3.3 Pré-processamento dos Dados

O objetivo foi ajustar a base de dados para o treinamento e a avaliação dos modelos. As alterações focaram na normalização dos dados e seguiram os seguintes passos:

1. Dados nulos: Retirados todos registros que possuem valores nulos. Na coleta de dados foi identificado que dados antes do ano de 2018 possuía informações não preenchidas, por isso a decisão da retirada dos anos anteriores, considerando assim os anos de 2018 a 2023.
2. Idades: os atributos `idade_media_titular_mandante` e `idade_media_titular_visitante` receberam ajustes de ponto flutuante.
3. Valor do time: todos os registros que continham times avaliados abaixo de um milhão de reais, foram removidos.
4. Estádios: Registros que apresentavam o mesmo nome de estádio de formas diferentes foram padronizados.
5. Times: Uma vez que o valor atribuído a coluna de times se trata de um dado categórico, foi utilizado a codificação *one-hot*, que transforma os dados categóricos em binários, sendo assim, possível de serem interpretados.
6. Divisão de *datasets*: Com o objetivo de normalizar o conjunto de dados para o treinamento e teste dos modelos, foram adotados dois cenários para serem abordados. Sendo o primeiro deles, considerando aqueles jogos que tiveram empate ao final das partidas, o outro, para fins de avaliação e comparação, foi desconsiderados os casos onde o resultado final foi empate, sendo assim, para montagem do *datasets* teve apenas partidas onde houve um vencedor. As Figuras 3 e 4 ilustram a distribuição dos resultados após a segmentação dos dois cenários, em que o valor 1 representa a vitória do time visitante e o valor 0 indica a vitória do mandante.

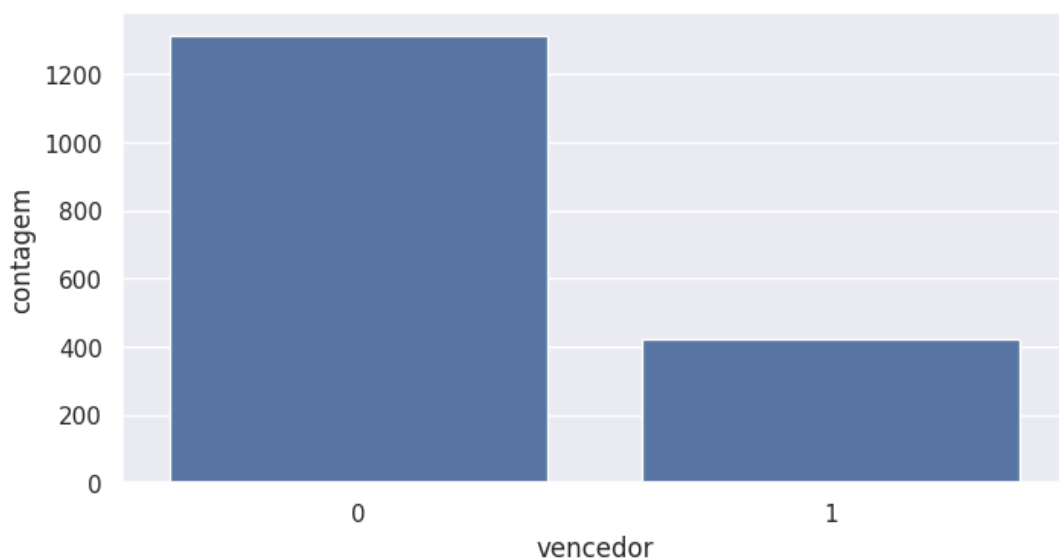


Figura 3 – Volumetria - cenário com empates.

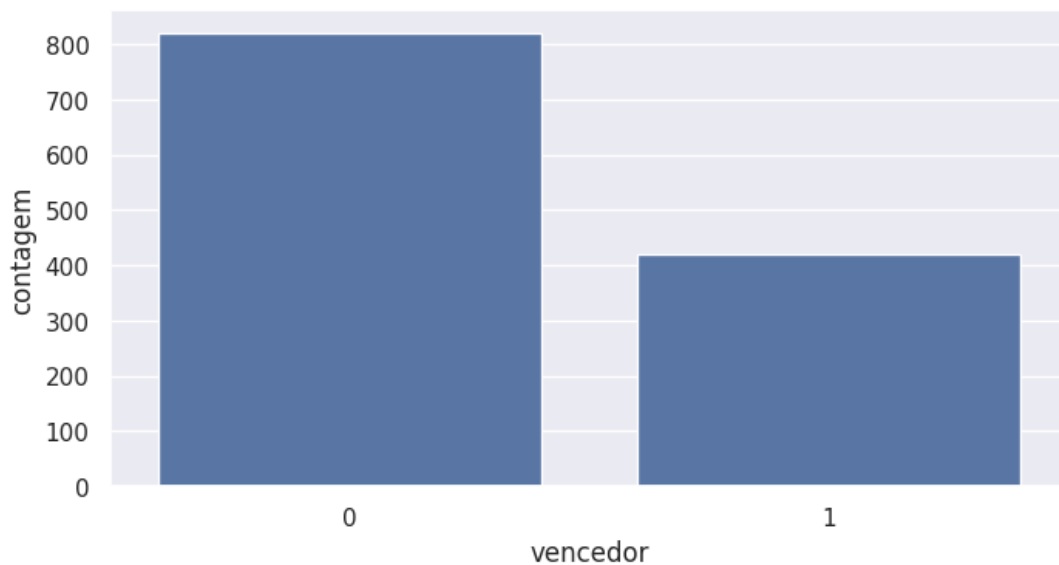


Figura 4 – Volumetria - cenário sem empates.

7. Transformação dos dados: Para fins da utilização dos modelos, para definição de atributo alvo, foi necessário um tratamento de transformação de dados categóricos para dados binários. Dessa forma, com o objetivo é prever a vitória do time visitante, foi adotado para o primeiro cenários (com empates) o valor 1 (um) para o atributo vencedor quando era identificado o valor categórico “fora” e 0 (zero) onde os valores de atributos categóricos em “casa” e “empate”. Já para o segundo cenário (sem empates), valor 1 (um) para o atributo vencedor quando “fora” e 0 (zero) quando atributo vencedor “casa”. As Tabelas 2 e 3 apresentam os resultados das transformações realizadas na atributo alvo.

	Atributo vencedor	Atributo vencedor	Atributo vencedor
Valor antes da transformação	fora	empate	casa
Valor depois da transformação	1	0	0

Tabela 2 – Atributo *vencedor* transformado - conjunto de dados com empate

	Atributo vencedor	Atributo vencedor
Valor antes da transformação	fora	casa
Valor depois da transformação	1	0

Tabela 3 – Atributo *vencedor* transformado - conjunto de dados sem empate

3.4 Modelagem e Divisão da Base de Dados

Esta etapa do desenvolvimento tem como objetivo a remoção do atributo alvo vencedor, a separação da base de dados em conjuntos de treino e teste, além da im-

plementação dos modelos de predição. Com a finalidade de prever um resultado cuja informação não seja previamente conhecida, o atributo alvo é retirado do conjunto de dados, o que permite uma simulação mais próxima de um cenário real. Dessa forma, a partir dos demais atributos disponíveis, o modelo deverá ser capaz de prever o valor do atributo alvo desconhecido.

Empregando uma estratégia amplamente utilizada no aprendizado de máquina, a base foi dividida em proporções de 70% para treinamento e 30% para teste. Foi adotado da seguinte forma, para que com a base de treinamento separada de forma aleatória, os modelos utiliza se de dados conhecidos espelhados da base original para poder realizar as previsões da atributo alvo e com a base de teste, pudesse ser observado através de métricas como validação cruzada, usando *K-fold*, a eficiência de cada modelo ao em cumprir como o objetivo pelo qual foi implementado.

Os algoritmos dos modelos criados para a previsão da atributo alvo são: LightGBM, Regressão Logística, *Random Forest*, *k-Nearest Neighbors*, *Naive Bayes*, Máquina de Vetores de Suporte e XGBoost. Todas as implementações foram realizadas com a biblioteca *scikit-learn* da linguagem *Python*.

Após as primeiras avaliações de resultados dos modelos abordados no Capítulo 4, foi utilizado a biblioteca *SHAP* para identificação e entendimento dos atributos mais relevantes para o modelo LightGBM, dessa forma, a fim de melhorar os resultados, foi criado uma nova base de treinamento e teste, contendo a mesma porcentagem descrita acima, porém contendo apenas os atributos mais relevantes identificados.

3.5 Avaliação e Interpretação

A avaliação, como etapa final do desenvolvimento, é responsável por comparar os modelos após o treinamento e o teste. A análise dos resultados é realizada por meio de Matrizes de Confusão e das métricas Acurácia, Precisão, *Recall* e *F1-score*. A partir da simulação de cenários reais, torna-se possível interpretar a eficiência de cada modelo, cuja discussão será aprofundada no Capítulo 4.

4 Resultados

Este capítulo é dedicado à apresentação e análise dos resultados obtidos a partir da aplicação dos algoritmos de aprendizado de máquina nos dois cenários considerados: com a inclusão dos empates e com a exclusão desses resultados. A distinção entre os cenários visa avaliar o impacto da presença dos empates na capacidade preditiva dos modelos, permitindo uma compreensão mais ampla sobre o comportamento das abordagens implementadas em diferentes condições do problema.

Além disso, são apresentados os resultados referentes à otimização do modelo LightGBM, realizada com base nas interpretações fornecidas pelo método *SHapley Additive exPlanations* (SHAP). Essa etapa permitiu identificar as variáveis com maior influência nas previsões do modelo, contribuindo tanto para o aprimoramento da acurácia quanto para a interpretabilidade do modelo final.

4.1 Visão Geral dos Cenários Avaliados

A presente seção oferece uma visão geral dos cenários definidos para a tarefa de previsão de resultados no Campeonato Brasileiro de Futebol Série A. Para esse fim, foram elaborados dois contextos distintos de modelagem, denominados Cenário A e Cenário B. O Cenário A contempla a classificação tripla dos resultados, abrangendo vitórias, empates e derrotas, o que proporciona uma abordagem mais abrangente, mas com maior complexidade para os algoritmos de classificação.

Por outro lado, o Cenário B adota uma abordagem binária, considerando apenas vitórias e derrotas, o que pode favorecer a performance preditiva ao reduzir a ambiguidade presente na classe de empates. Essa distinção entre os cenários permite avaliar o impacto da granularidade das classes nos desempenhos dos modelos. Além disso, ela oferece subsídios para futuras aplicações em contextos específicos de análise e tomada de decisão.

- **Cenário A:** Previsão considerando vitórias, empates e derrotas.
- **Cenário B:** Previsão considerando apenas vitórias e derrotas.

4.2 Resultados do Cenário A – Com Empates

4.2.1 Treinamento

Aplicou-se a validação cruzada, utilizando-se a divisão em 10 *folds* como estratégia para o treinamento dos modelos, conforme descrito na Seção 3.4. Esse procedimento tem como objetivo garantir maior robustez e capacidade de generalização dos resultados, ao permitir que cada subconjunto da base de dados seja utilizado tanto para treinamento quanto para teste em diferentes iterações. A Tabela 4 apresenta os resultados obtidos para as principais métricas de desempenho, acompanhadas de seus respectivos valores de desvio padrão, o que possibilita a avaliação da consistência dos modelos. Dessa forma, essa representação fornece uma análise preliminar relevante sobre o comportamento dos algoritmos aplicados, oferecendo subsídios para a comparação entre seus desempenhos em diferentes aspectos da predição.

Algoritmo	Acurácia	Precisão	Recall	F1
LR	0.75 ± 0.03	0.58 ± 0.23	0.12 ± 0.05	0.18 ± 0.06
RF	0.65 ± 0.06	0.46 ± 0.30	0.05 ± 0.04	0.10 ± 0.08
KNN	0.69 ± 0.06	0.30 ± 0.07	0.17 ± 0.07	0.21 ± 0.06
NB	0.75 ± 0.03	0.48 ± 0.13	0.16 ± 0.06	0.24 ± 0.12
SVM	0.72 ± 0.05	0.45 ± 0.36	0.02 ± 0.04	0.09 ± 0.07
XGB	0.75 ± 0.04	0.53 ± 0.20	0.16 ± 0.06	0.23 ± 0.08
LGBM	0.73 ± 0.04	0.46 ± 0.20	0.15 ± 0.07	0.26 ± 0.05

Tabela 4 – Desempenho dos modelos preditivos utilizando dados de treinamento - cenário com empates

Entre os destaques observados nas métricas apresentadas na Tabela 4, tem-se: Regressão Logística e XGBoost com acurácia de 75%; Regressão Logística com precisão de 58%; K-Nearest Neighbors com *recall* de 17%; e XGBoost com *F1-score* de 23%. Embora esse último valor não represente o maior percentual, o modelo apresentou o menor desvio padrão entre os resultados, equivalente a 8%, o que indica maior estabilidade nas previsões.

4.2.2 Teste

Nos testes, utilizou-se a parte da base de dados previamente reservada para essa finalidade, mantendo-se o atributo alvo oculto aos modelos. Com os modelos já treinados, conforme descrito na Seção 4.2.1 a avaliação dos resultados torna-se mais precisa e alinhada com cenários reais de predição.

Para facilitar a visualização dos resultados, foram geradas as matrizes de confusão de cada modelo, bem como reapresentadas as métricas empregadas durante a fase de treinamento. A Figura 5 exibe as matrizes de confusão correspondentes a todos os modelos

analisados, enquanto a Figura 6 apresenta, em formato de gráfico de barras, os valores obtidos para cada métrica avaliada.

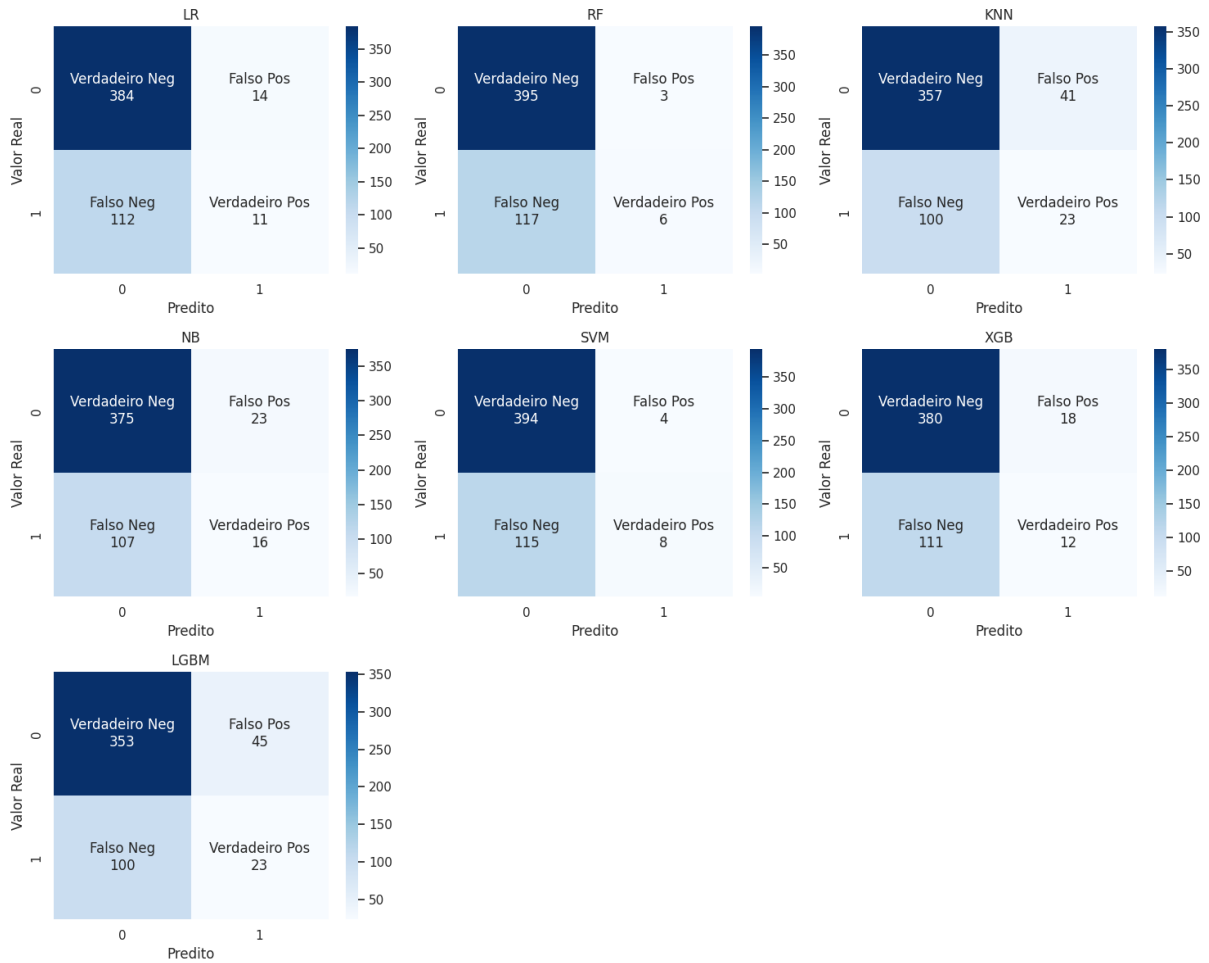


Figura 5 – Desempenho dos modelos representado pelas matrizes de confusão (cenário com empates).

A partir dos resultados obtidos e da análise das matrizes de confusão, representadas na Figura 5, destacam-se os seguintes pontos: o modelo LightGBM apresentou 23 verdadeiros positivos, sendo, portanto, aquele que mais previu corretamente as vitórias do time visitante. No entanto, com 45 casos de falso positivo, foi também o modelo que mais classificou equivocadamente partidas como vitórias do visitante, quando, na realidade, o resultado foi vitória do mandante ou empate. Por sua vez, o modelo de Regressão Logística demonstrou eficiência na previsão de vitórias do mandante ou empates, com 395 acertos. Contudo, apresentou 117 falsos negativos, indicando que, nesses casos, previu incorretamente vitória do mandante ou empate, quando, na verdade, o visitante foi o vencedor.

Na Figura 6, constatarem-se os melhores desempenhos nas métricas de acurácia e precisão para o modelo SVM. Em relação à métrica de *recall*, os modelos LightGBM e *K-Nearest Neighbors* (KNN) apresentaram valores equivalentes. Por fim, a maior pontuação

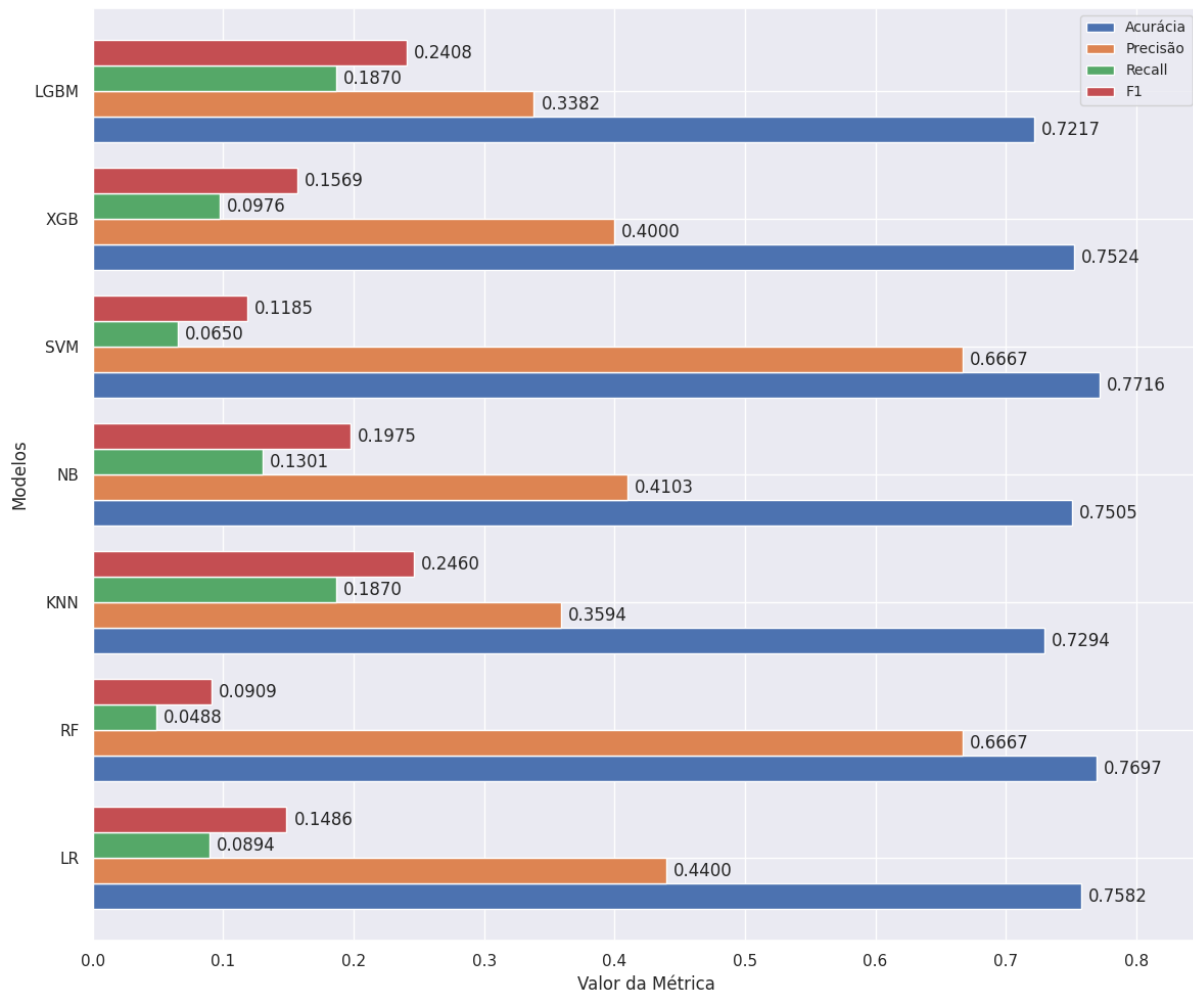


Figura 6 – Comparação das métricas dos modelos por meio de gráfico de barras (considerando empates).

de *F1-score* foi obtida pelo modelo KNN, indicando melhor equilíbrio entre precisão e *recall*.

4.2.3 Interpretação com SHAP

A Figura 7, exibe os valores *SHapley Additive exPlanations*, que quantificam a contribuição de cada variável individual para a predição de um determinado resultado. No gráfico, cada ponto representa uma instância da base de dados, e as cores indicam os valores das variáveis: pontos em vermelho representam valores altos da variável correspondente, enquanto pontos em azul indicam valores baixos. O eixo horizontal mostra o valor SHAP, que representa o impacto da variável na previsão do modelo. Valores mais distantes de zero indicam maior influência na decisão preditiva.

Nesse contexto, as variáveis *odds_mandante_vence*, *valor_equipe_titular_visitante*, *colocacao_mandante* e *valor_equipe_titular_mandante* se destacaram como as mais influentes. Essas variáveis estão diretamente associadas à força relativa das equipes e à ex-

pectativa de vitória, o que reforça sua relevância na lógica de decisão do modelo. Além disso, observou-se uma contribuição significativa de variáveis relacionadas ao desempenho prévio das equipes.

De maneira geral, o modelo demonstrou sensibilidade a informações estratégicas e estatísticas, evidenciando sua capacidade de incorporar variáveis explicativas que contribuem de forma significativa para a previsão dos resultados das partidas.

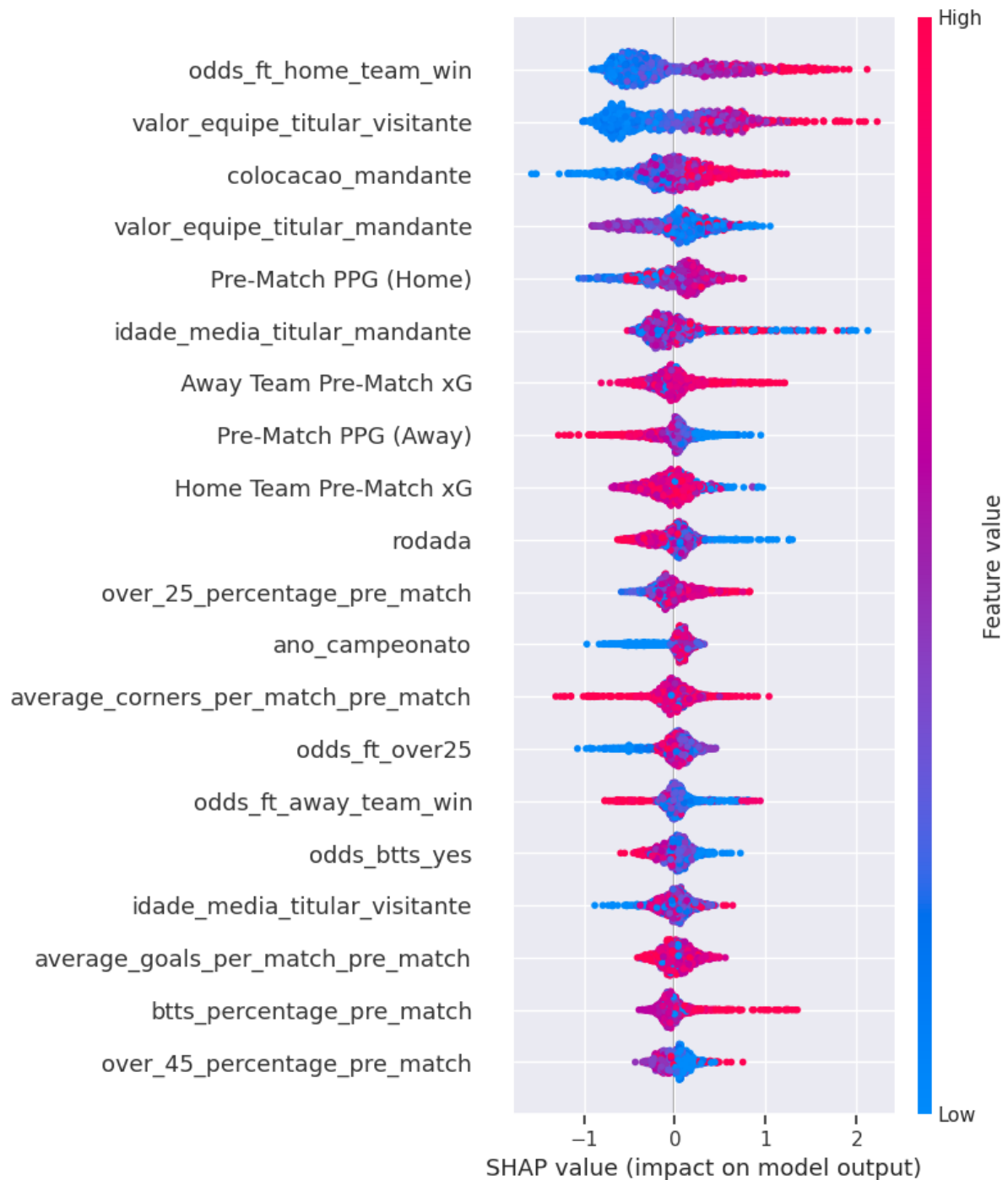


Figura 7 – Plotagem do gráfico sumário com os valores *SHAP*.

Com base na análise anterior, foi criado um novo modelo utilizando apenas as

variáveis mais relevantes do LightGBM. A Tabela 5 apresenta as métricas de desempenho e a Figura 8, a matriz de confusão resultante.

Algoritmo	Acurácia	Precisão	Recall	F1
LGBM	0.727447	0.358209	0.195122	0.252632

Tabela 5 – Desempenho dos modelo preditivo LightGBM utilizando dados de teste - cenário com empates

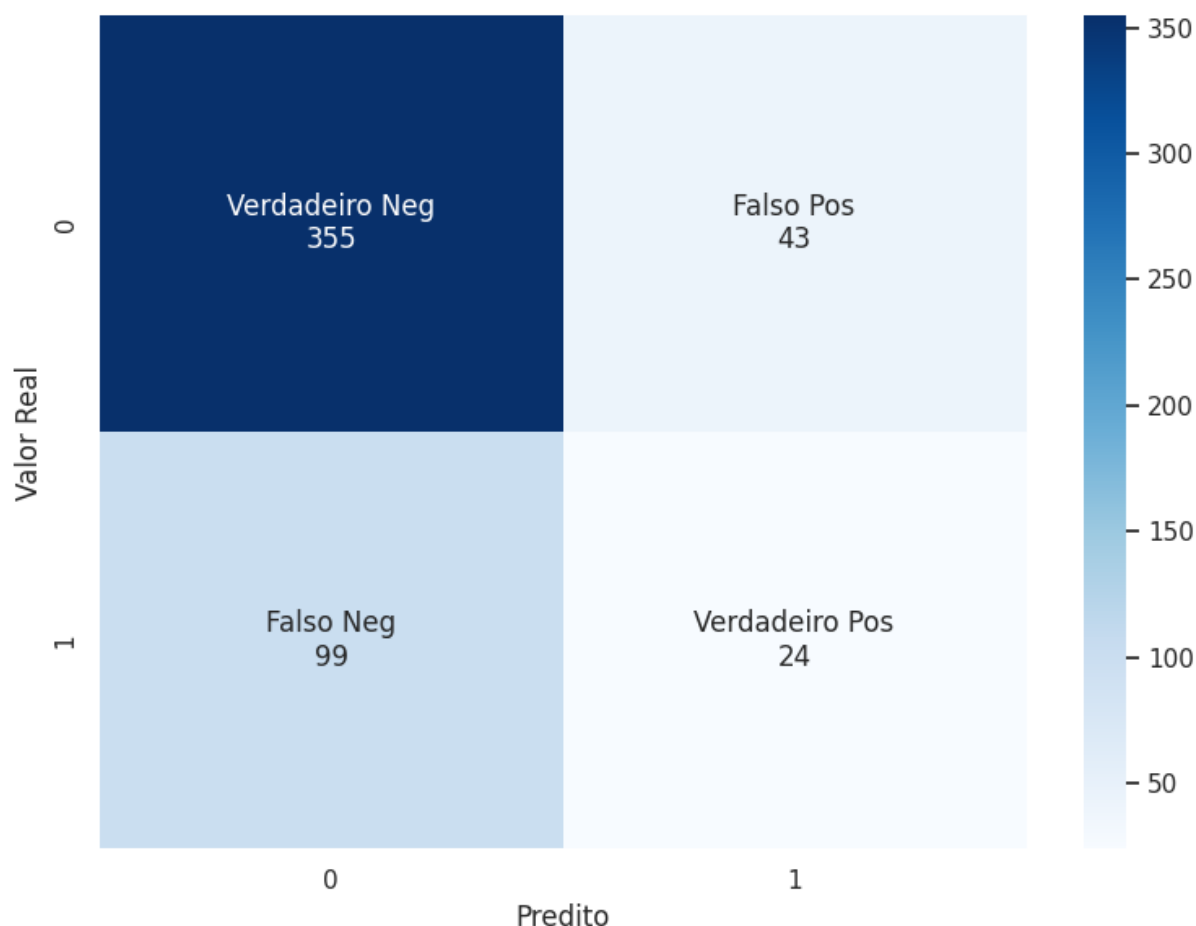


Figura 8 – Matriz de Confusão modelo LightGBM após novo treinamento e testes com empates.

A substituição pelas variáveis mais relevantes no modelo LightGBM resultou em uma leve melhoria no desempenho, com um pequeno aumento na precisão, bem como nos valores de verdadeiros positivos e verdadeiros negativos, embora sem variações estatisticamente significativas.

4.3 Resultados do Cenário B – Sem Empates

Nesta seção, são apresentados os resultados do cenário em que os empates foram desconsiderados da base de dados. O objetivo é observar a possibilidade de melhora no desempenho dos modelos ao se considerar um resultado binário para cada partida.

4.3.1 Treinamento

Adotou-se a mesma estratégia de validação cruzada do cenário com empates, conforme descrito na Seção 4.2.1, para o treinamento dos modelos. Na Tabela 6, apresentam-se os primeiros resultados obtidos, utilizando-se as mesmas métricas e o mesmo modelo de representação.

Algoritmo	Acurácia	Precisão	Recall	F1
LR	0.69 ± 0.06	0.65 ± 0.09	0.36 ± 0.10	0.45 ± 0.09
RF	0.67 ± 0.04	0.58 ± 0.19	0.31 ± 0.07	0.43 ± 0.09
KNN	0.65 ± 0.02	0.51 ± 0.07	0.39 ± 0.05	0.46 ± 0.08
NB	0.68 ± 0.07	0.62 ± 0.09	0.29 ± 0.05	0.40 ± 0.11
SVM	0.68 ± 0.05	0.61 ± 0.15	0.27 ± 0.08	0.38 ± 0.07
XGB	0.68 ± 0.03	0.59 ± 0.11	0.41 ± 0.12	0.46 ± 0.09
LGBM	0.65 ± 0.04	0.54 ± 0.08	0.40 ± 0.05	0.47 ± 0.07

Tabela 6 – Desempenho dos modelos preditivos utilizando dados de treinamento - cenário sem empates

Entre os modelos avaliados, a Regressão Logística apresentou a maior acurácia, com valor médio de 69% e desvio padrão de 6%, além de ter alcançado também o melhor desempenho em termos de precisão, com 65%. Em relação ao *recall*, o destaque foi o modelo XGBoost, com 41%, seguido de perto pelo LightGBM, com 40%.

O maior F1-score foi obtido pelo modelo LightGBM, com 47%, que também se destacou por apresentar o menor desvio padrão nessa métrica, de 7%, demonstrando maior consistência em seus resultados. O modelo KNN apresentou resultados medianos, mas foi o que obteve o menor desvio na acurácia, de apenas 2%, o que indica maior estabilidade nesse aspecto específico.

4.3.2 Teste

Empregando as mesmas normas adotadas no cenário com empates, foi desenvolvido o experimento correspondente ao cenário discutido nesta seção. Observa-se, pelas Figuras 9 e 10, que houve uma melhora no desempenho dos modelos, evidenciada tanto pelas métricas quanto pelas matrizes de confusão, uma vez que a exclusão dos empates elimina uma fonte de ambiguidade que poderia comprometer as previsões.

Ao analisar as matrizes de confusão na Figura 9, verifica-se que o modelo XGBoost apresentou um desempenho mais equilibrado, destacando-se na identificação de vitórias do time visitante, com 50 verdadeiros positivos e 55 falsos negativos. O que demonstra uma boa capacidade de identificação desse tipo de resultado, com um número relativamente reduzido de erros. Por outro lado, o modelo KNN foi o que mais cometeu equívocos ao prever vitórias do visitante, registrando 74 falsos positivos, ou seja, previu vitórias do visitante quando, na realidade, houve vitórias do mandante. Já o modelo *Naïve Bayes*

apresentou 237 verdadeiros negativos, indicando um bom desempenho na identificação de vitórias do time mandante. No entanto, também cometeu 74 falsos negativos, evidenciando limitações na detecção de vitórias do visitante, cujo desempenho preditivo foi mais discreto, com apenas 31 acertos.

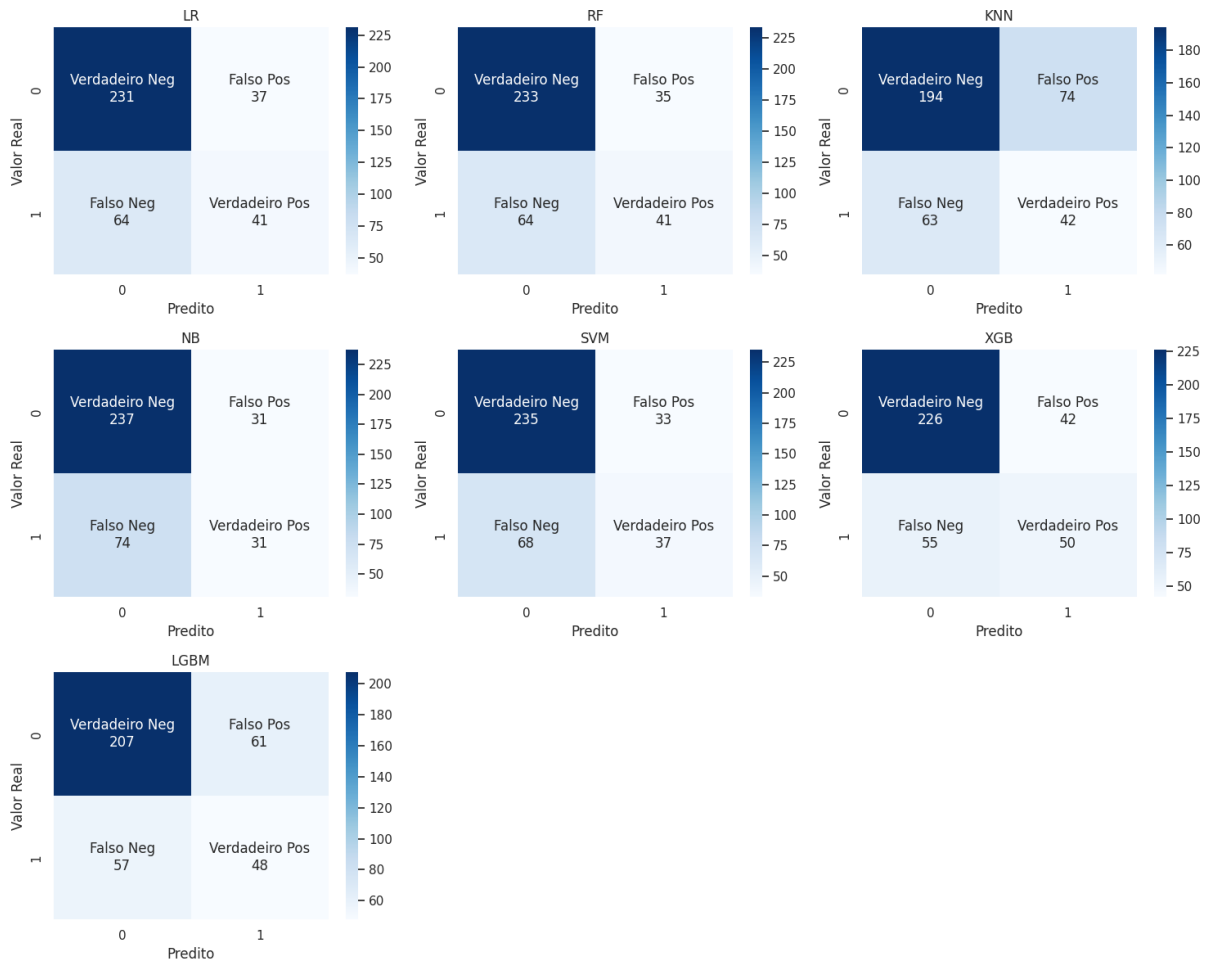


Figura 9 – Desempenho dos modelos representado pelas matrizes de confusão (cenário sem empates).

O modelo LightGBM também se destacou, sendo o segundo com maior número de verdadeiros positivos (48 casos). No entanto, apresentou desempenho inferior na previsão de vitórias do time mandante, o que se refletiu em uma precisão de apenas 44%.

Com base nos resultados apresentados na Figura 10, é possível destacar as seguintes conclusões: o modelo XGBoost obteve o melhor desempenho geral, com acurácia de 73%, além dos melhores índices de precisão (54%), superando modelos com desempenho semelhante, como a Regressão Logística (53,9%). O XGBoost também apresentou os maiores valores de *recall* (48%) e F1-score (50%).

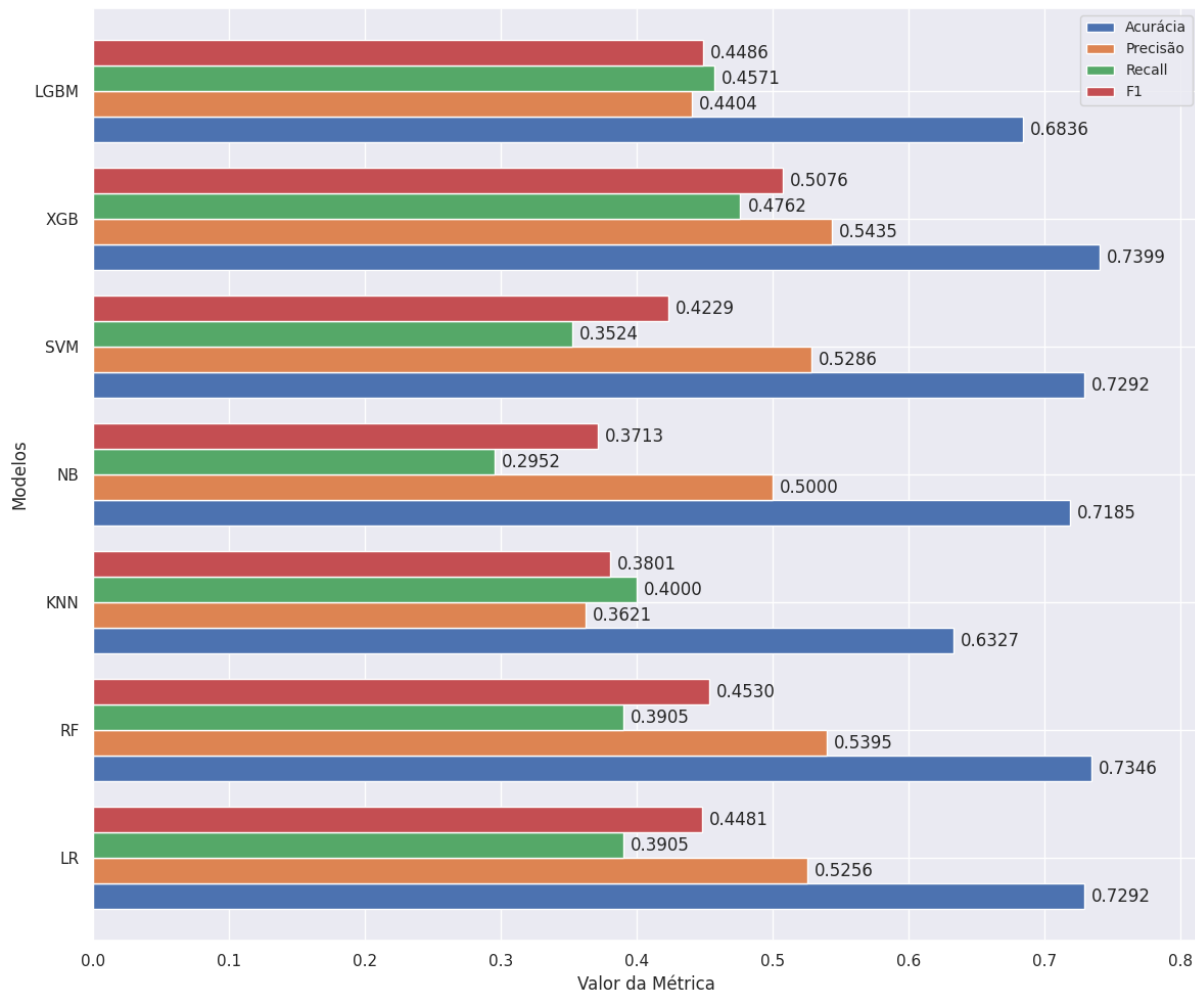


Figura 10 – Comparação das métricas dos modelos por meio de gráfico de barras (não considerando empates).

4.3.3 Interpretação com SHAP

A Figura 11 apresenta o gráfico de interpretação dos valores *SHAP* aplicado ao modelo LightGBM como realizado no cenário com empates, permitindo analisar quais variáveis que são mais importantes nas predições do modelo e de que forma isso ocorre.

As três variáveis com maior impacto foram: `odds_ft_home_team_win` (mantendo-se em primeiro lugar), `valor_equipe_titular_visitante` e `valor_equipe_titular_mandante`. Além dessa troca de variáveis mais importantes, outra que aparece em melhor colocação é `odds_ft_away_team_win`. Isso mostra que tanto as *odds* das casas de apostas quanto o valor de mercado das equipes titulares são fatores determinantes para o modelo prever o resultado da partida, se direcionado de forma mais objetiva a prever corretamente.

Outras variáveis importantes incluem a colocação do time mandante, a idade média dos titulares e métricas pré-jogo, como a média de gols. Esses dados indicam que o desempenho recente e o perfil dos jogadores também influenciam significativamente as decisões do modelo, reiterando o que foi observado na Seção 4.2.3.

Embora apresentem menor impacto, variáveis como a formação tática da equipe visitante (por exemplo, 4-2-3-1) e a rodada do campeonato também se mostram relevantes em contextos específicos. De forma geral, o gráfico contribui para a compreensão do processo decisório do modelo, confirmando que ele considera uma combinação de fatores relacionados à qualidade das equipes, estatísticas históricas e probabilidades indicadas pelo mercado.

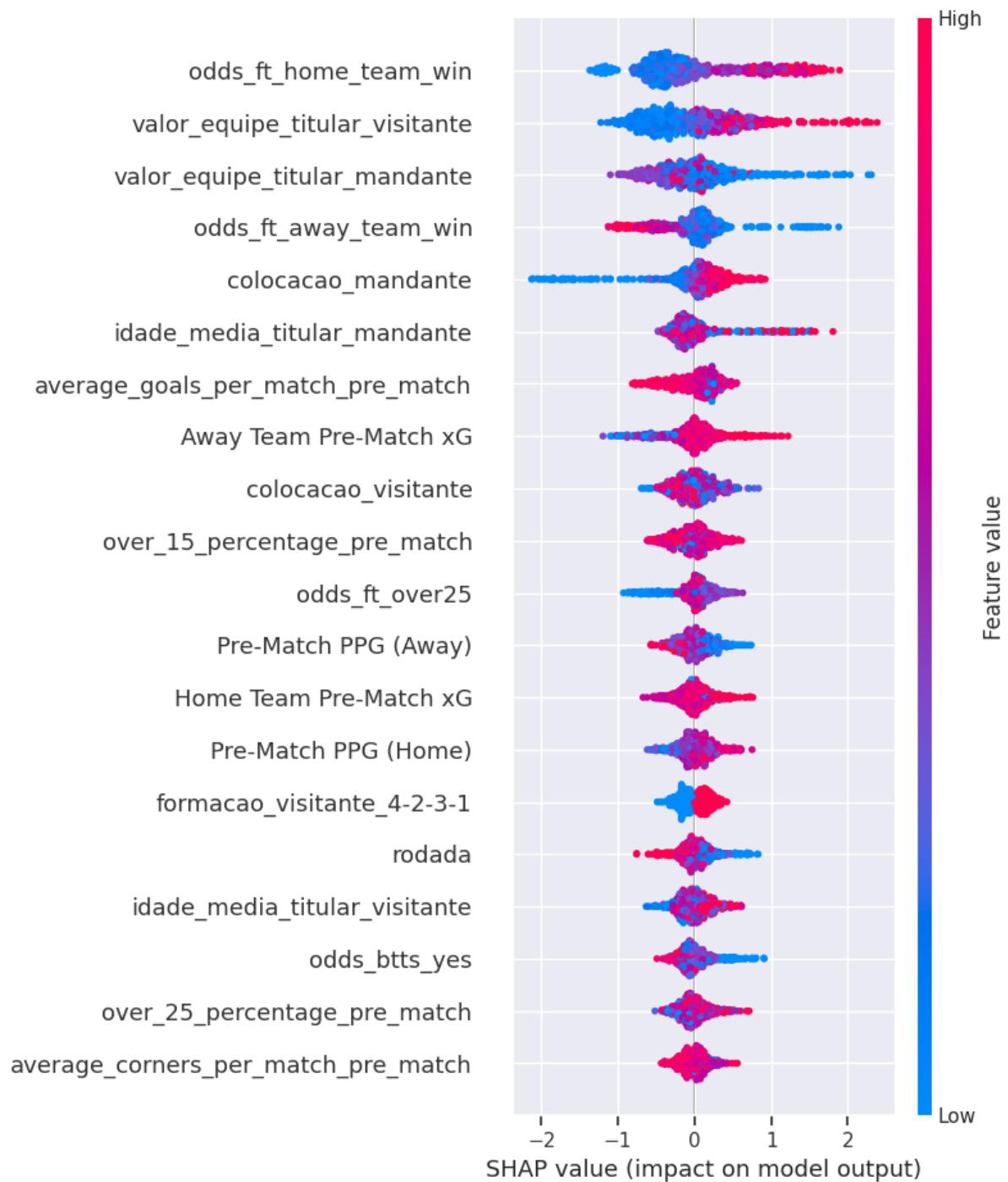


Figura 11 – Plotagem do gráfico sumário com os valores *SHAP*.

Assim como nos casos em que os empates foram considerados, para o cenário

descrito foi realizada uma nova modelagem, utilizando as variáveis mais relevantes, seguida de novo processo de treinamento e testes. Os resultados obtidos estão apresentados na Tabela 7 e na Figura 12.

Algoritmo	Acurácia	Precisão	Recall	F1
LGBM	0.702413	0.472727	0.495238	0.483721

Tabela 7 – Desempenho dos modelo preditivo LightGBM utilizando dados de teste - cenário sem empates



Figura 12 – Matriz de Confusão modelo LightGBM após novo treinamento e testes sem empates.

Conforme verificado, os ajustes realizados para a otimização do modelo resultaram em uma melhoria na eficácia da previsão dos resultados, com aumento nos casos de verdadeiros positivos e verdadeiros negativos, além da consequente redução nas previsões incorretas.

4.4 Discussão dos resultados

Em ambos os cenários do trabalho, a avaliação dos modelos de classificação revelou diferenças relevantes no desempenho das técnicas aplicadas. As métricas de acurácia,

precisão, *recall* e *F1-score*, acompanhadas pelas matrizes de confusão, permitem uma compreensão mais aprofundada de como cada algoritmo se comportou diante do desafio de prever o resultado das partidas.

No Cenário A, onde o modelo foi treinado com todas as variáveis e contemplou resultados de partidas com empates, os resultados obtidos estiveram dentro do esperado, mas apresentaram valores mais baixos em relação às classes de verdadeiros positivos previstos. Isso ocorreu devido aos modelos apresentarem valores baixos de precisão e *recall*, o que se reflete na grande quantidade de previsões de verdadeiros negativos. Em outras palavras, os modelos demonstraram um melhor desempenho na previsão de vitórias do time mandante, em vez de prever a vitória do time visitante, que era o objetivo e a classe alvo.

Com base nos resultados, destaca-se o **K-Nearest Neighbors (KNN)**, que obteve, dentre os modelos avaliados, a maior quantidade de verdadeiros positivos, contabilizando 23 registros, além de 357 verdadeiros negativos. No entanto, apresentou a segunda menor precisão entre os modelos, com apenas 36%. Isso indica que o KNN tem maior propensão a falsos positivos.

O modelo LightGBM, que foi selecionado para identificar as variáveis importantes e criar um novo modelo para as rodadas de treinamento e testes, apresentou uma avaliação semelhante à do modelo KNN. Após a criação de um modelo otimizado, ele não atendeu às expectativas de melhoria na previsão da atributo alvo, com poucas alterações na volumetria apresentada na matriz de confusão.

No Cenário B, os resultados evidenciam uma melhoria significativa no desempenho dos modelos, especialmente em relação à precisão das previsões. Entre os modelos analisados, a **Regressão Logística (LR)** destacou-se com uma acurácia média de 0.69 ± 0.06 , além da maior precisão entre os algoritmos testados (0.65 ± 0.09), o que demonstra sua capacidade de acertar ao prever vitórias do time visitante. No entanto, o *recall* relativamente baixo (0.36 ± 0.10) revela uma limitação na detecção de todos os casos reais de vitória do visitante. A matriz de confusão desse modelo reforça essa análise: enquanto apresentou 231 verdadeiros negativos, teve 64 falsos negativos, ou seja, deixou de identificar corretamente algumas vitórias dos visitantes.

O modelo **XGBoost** também apresentou bom desempenho geral, com *F1-score* de 0.46 ± 0.09 e o maior número de verdadeiros positivos (50 casos), o que indica uma boa performance na previsão de vitórias do visitante. Além disso, seu desvio padrão das métricas foi relativamente baixo, evidenciando estabilidade nas execuções.

O **K-Nearest Neighbors (KNN)** alcançou um dos maiores valores de *recall* (0.39 ± 0.05), o que pode indicar uma maior sensibilidade na detecção de vitórias do visitante, embora isso tenha sido acompanhado por uma menor precisão (0.51 ± 0.07), ou seja,

uma maior propensão a falsos positivos. Esse comportamento é confirmado pela matriz de confusão, que revelou 42 acertos de vitórias do visitante e 74 previsões equivocadas nesse sentido.

Por outro lado, o **LightGBM** foi o modelo que mais previu vitórias do visitante, com 48 acertos, mas também foi o que cometeu mais erros nessa previsão, com 61 falsos positivos. Esse comportamento sugere uma tendência do modelo em superestimar vitórias do visitante, o que pode ser problemático em cenários nos quais esse resultado é menos frequente. No entanto, essa tendência foi parcialmente corrigida após a utilização do *SHAP* e a definição das variáveis relevantes para o modelo, o que resultou na redução dos falsos positivos para 58 casos. Embora essa melhoria não tenha sido expressiva, ela aponta para um direcionamento positivo na otimização da previsão.

De forma geral, nenhum modelo apresentou desempenho superior em todas as métricas analisadas. Portanto, a escolha do modelo ideal depende do objetivo da análise: caso o foco seja acertar quando o visitante vence (maior *recall*), modelos como o KNN ou o XGBoost podem ser mais indicados.

5 Conclusão

O futebol, o esporte mais praticado no país e um dos maiores do mundo, com o avanço da capacidade de armazenamento e processamento de dados junto à popularidade de casas de apostas, as informações sobre o assunto se tornam mais disponíveis e robustas para trabalhos relacionados a campeonatos e previsão de resultados. O trabalho desenvolvido, nichado no Campeonato Brasileiro de Futebol da Série A, na modalidade de Pontos Corridos, torna-se atrativo a previsão da vitória do time visitante pois cada ponto somado fora dos seus domínios, ao final do campeonato e a soma dos pontos, pode diferenciar um time que vai ser consagrado campeão da edição ou um que se livra do rebaixamento garantindo a sua permanência na divisão de elite.

O treinamento e teste foram realizados em dois cenários, um conjunto de dados com empates, e conjunto de dados sem empates. Com uma apresentação de resultados pouco eficaz no primeiro cenário, o segundo apresentou uma melhora significativa, onde as métricas e resultados de matrizes de confusão superaram a dos conjuntos com empates por simplificar a problemática para os modelos.

Relacionado a identificação da importância das variáveis para o modelo LightGBM, para ambos os cenários é possível identificar uma melhora nos resultados principalmente quando visto pela óptica do cenário sem empates. Mesmo com pouco avanço após a tentativa de otimização do modelo com os resultados do SHAP, é um indicativo de como melhorar os resultados do modelo.

Por se tratar de um problema de previsão, no qual o atributo alvo assume um valor binário, o modelo XGBoost demonstrou-se eficaz. Entre os modelos avaliados, foi o que apresentou o desempenho mais equilibrado, com métricas satisfatórias e boa adequação ao conjunto de dados utilizado. Esse resultado indica sua capacidade de capturar padrões relevantes associados às vitórias fora de casa, mesmo diante da complexidade inerente ao contexto esportivo.

Dessa forma, a utilização de modelos preditivos em partidas de futebol revela-se promissora. Há um amplo campo para melhorias, especialmente por meio da escolha de algoritmos mais alinhados à natureza do problema e do desenvolvimento de bases de dados mais representativas e balanceadas.

Para trabalhos futuros, sugere-se a interpretação dos modelos com técnicas de explicabilidade. Aplicando **SHAP**, **LIME** ou outros métodos para explicar as decisões dos modelos e entender melhor quais fatores mais influenciam o resultado das partidas. Além do aprimoramento do balanceamento de classes, que poderá possibilitar a identificar corretamente resultados menos frequentes, como vitórias do visitante.

Referências

- ABDELRAOUF, H.; AAAA, S.; MUTHANNA, A.; IBRAHIM, M. **Confusion Matrix-Based Performance Evaluation Metrics**. 2023. <https://www.researchgate.net/figure/Confusion-matrix-and-evaluation-metrics_fig5_370920203>. Accessed: 2025-04-30. Citado 2 vezes nas páginas 6 e 19.
- BERNARDES, D. R. **Predizendo os Vencedores dos Playoffs: Um Estudo de Caso com Aprendizado de Máquina em Partidas de Futebol Americano**. 2023. <https://sol.sbc.org.br/index.php/sbbd_estendido/article/view/25609/25427>. Citado na página 11.
- BREIMAN, L. Random forests. **Machine Learning**, Springer, v. 45, n. 1, p. 5–32, 2001. Citado na página 16.
- CAIRES, T. d. C. **Técnicas de Interpretabilidade para Aprendizado de Máquina: Um Estudo Abordando Avaliação de Crédito e Detecção de Fraude**. Dissertação (Dissertação de Mestrado) — Universidade de São Paulo, São Paulo, SP, Brasil, 2022. Citado na página 18.
- CHEN, T.; GUESTRIN, C. XGBoost: A Scalable Tree Boosting System. In: **Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining**. New York, NY, USA: Association for Computing Machinery, 2016. (KDD '16), p. 785–794. ISBN 978-1-4503-4232-2. Disponível em: <<https://arxiv.org/abs/1603.02754>>. Citado na página 15.
- CNN Brasil. **CNN Brasil: Notícias ao vivo do Brasil e do mundo**. 2025. <<https://www.cnnbrasil.com.br/>>. [Online; accessed 13-Janeiro-2025]. Citado na página 13.
- CORTES, C.; VAPNIK, V. Support-vector networks. **Machine Learning**, Springer, v. 20, n. 3, p. 273–297, 1995. Citado na página 17.
- CUTLER, D. R.; JR, T. C. E.; BEARD, K. H.; CUTLER, A.; HESS, K. T.; GIBSON, J.; LAWLER, J. J. Random forests for classification in ecology. **Ecology**, Wiley Online Library, v. 88, n. 11, p. 2783–2792, 2007. Citado na página 17.
- DEUS, G. A. D. **UTILIZAÇÃO DE APRENDIZADO DE MÁQUINA PARA PREVISÃO DE RESULTADOS DE JOGOS DE FUTEBOL**. Dissertação (Trabalho de Conclusão de Curso) — Universidade Tecnológica Federal do Paraná (UTFPR), Paraná, Brasil, 2019. Disponível em: <<https://repositorio.utfpr.edu.br/jspui/handle/1/12482>>. Citado na página 11.
- FACELI, K.; LORENA, A. C.; GAMA, J.; ALMEIDA, T. A. d.; CARVALHO, A. C. d. **Inteligência Artificial: Uma Abordagem de Aprendizado de Máquina**. [S.l.]: LTC, 2021. Citado na página 13.
- FRYER, D.; STRUMBELJ, E.; BOSNIC, Z. Shapley values for feature selection: The good, the bad, and the axioms. **arXiv preprint arXiv:2102.10936**, 2021. Citado na página 18.

- GE. **Palmeiras e Flamengo concentram quase um quarto das receitas dos maiores clubes brasileiros**. 2019. <<https://ge.globo.com/sp/futebol/noticia/palmeiras-e-flamengo-concentram-quase-um-quarto-das-receitas-dos-maiores-clubes-brasileiros.ghtml>>. [Online; accessed 30-Abril-2025]. Citado na página 11.
- HOSMER, D. W.; LEMESHOW, S.; STURDIVANT, R. X. **Applied Logistic Regression**. 3. ed. [S.l.]: John Wiley & Sons, 2013. Citado na página 16.
- KE, G.; MENG, Q.; FINLEY, T.; ... LightGBM: A highly efficient gradient boosting decision tree. In: CURRAN ASSOCIATES, INC. **Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS)**. Long Beach, CA, USA, 2017. p. 3146–3154. Citado 2 vezes nas páginas 14 e 15.
- KLEINA, M. **ARTIFICIAL INTELLIGENCE TECHNIQUES APPLIED TO PREDICT TEAMS POSITION OF THE BRAZILIAN FOOTBALL CHAMPIONSHIP**. 2021. <<https://periodicos.uem.br/ojs/index.php/RevEducFis/article/view/50871/751375152401>>. Citado na página 11.
- MANNING, C. D.; RAGHAVAN, P.; SCHÜTZE, H. **Introduction to Information Retrieval**. Cambridge: Cambridge University Press, 2008. Citado na página 17.
- MARCHI, A. J.; FONSECA, M. Z.; BODÊ, J. Machine learning: aplicabilidade em monitoramento de redes. In: **FatecSeg-Congresso de Segurança da Informação**. [S.l.: s.n.], 2023. Citado na página 20.
- MOUSTAKIDIS, S.; PLAKIAS, S.; KOKKOTIS, C.; TSATALAS, T.; TSAOPOULOS, D. Predicting football team performance with explainable ai: Leveraging shap to identify key team-level performance metrics. **Future Internet**, MDPI, v. 15, n. 5, p. 174, 2023. Disponível em: <<https://www.mdpi.com/1999-5903/15/5/174>>. Citado na página 22.
- MURDOCH, W. J.; SINGH, C. K.; KUMBIER, K.; ABBASI-ASL, R.; YU, B. Interpretable machine learning: definitions, methods, and applications. **arXiv preprint arXiv:1901.04592**, 2019. Citado na página 18.
- NG, A. Y.; JORDAN, M. I. On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. In: **Advances in neural information processing systems**. [S.l.: s.n.], 2002. v. 14, p. 841–848. Citado na página 16.
- OLIVEIRA, J. V. d. **Análise comparativa de algoritmos de aprendizado de máquina aplicados ao Campeonato Brasileiro de Futebol**. Dissertação (Trabalho de Conclusão de Curso) — Universidade Federal de Uberlândia (UFU), Minas Gerais, Brasil, 2024. Disponível em: <<https://repositorio.ufu.br/handle/123456789/5142/browse?type=author&order=ASC&rpp=100&value=Oliveira%2C+Jo%C3%A3o+Victor+de>>. Citado 2 vezes nas páginas 21 e 24.
- 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. Citado 2 vezes nas páginas 19 e 20.
- RUDIN, C.; CHEN, C.; CHEN, H.; HUANG, Q.; SEMENOVA, L.; ZHONG, Q. Interpretable machine learning: Fundamental principles and 10 grand challenges. **arXiv preprint arXiv:2103.11251**, 2021. Citado na página 18.

- SCHNEIDER, C. F. **Machine Learning Aplicado na Previsão de Resultados de Partidas de Futebol: Um Estudo de Caso para Comparação de Diferentes Classificadores**. Dissertação (Trabalho de Conclusão de Curso) — Universidade Federal de Santa Catarina (UFSC), Santa Catarina, Brasil, 2022. Disponível em: acesso em 11 abr. 2025. Disponível em: <<https://lume.ufrgs.br/handle/10183/179461>>. Citado na página 21.
- SCHÖLKOPF, B.; SMOLA, A.; MÜLLER, K.-R. Kernel principal component analysis. In: SPRINGER. **International Conference on Artificial Neural Networks**. [S.l.], 1997. p. 583–588. Citado na página 17.
- SOKOLOVA, M.; LAPALME, G. A systematic analysis of performance measures for classification tasks. **Information Processing & Management**, Elsevier, v. 45, n. 4, p. 427–437, 2009. Citado na página 19.
- SUYAL, M.; GOYAL, P. A review on analysis of k-nearest neighbor classification machine learning algorithms based on supervised learning. **Review Article**, Seventh Sense Research Group®, v. 70, n. 7, p. 43–48, 2022. Department of CA & IT, SGRR University, Uttarakhand, India. Disponível em: <<https://www.sevensense.org>>. Citado na página 16.
- TAN, P.-N.; STEINBACH, M.; KUMAR, V. **Introduction to Data Mining**. 2. ed. [S.l.]: Pearson, 2018. Citado na página 16.
- TUESTA, E. F.; DELGADO, K. V.; BARBIERI, A. S.; ALVES, C. B.; CARVALHO, G. L.; SABINO, A. U.; ANDRE, L. A. M. L.; ZANOTI, L. C.; TOYOTA, P. A. Análise comparativa da competitividade do campeonato brasileiro de futebol mediante redes complexas. In: **Anais do Brazilian Workshop on Social Network Analysis and Mining (BraSNAM)**. SBC - Sociedade Brasileira de Computação, 2017. Disponível em: <<https://sol.sbc.org.br/index.php/brasnam/article/view/3253>>. Acesso em: 29 abr. 2025. Disponível em: <<https://sol.sbc.org.br/index.php/brasnam/article/view/3253>>. Citado na página 13.
- VALUE, S. **Valuation TOP 30 clubes do Brasil 2023 – 4ª edição**. 2023. <<https://www.sportsvalue.com.br/case-studies/valuation-top-30-clubes-do-brasil-2023-4a-edicao/>>. [Online; accessed 30-Abril-2025]. Citado na página 11.
- YACOUBY, R.; AXMAN, D. Evaluation metrics for multiclass classification: an overview and case study. **arXiv preprint arXiv:2008.05756**, 2020. Citado na página 20.
- ZHANG, Y.; YANG, Z.; LIU, X. et al. Application of the machine learning lightgbm model to the prediction of estuarine water levels. **Journal of Marine Science and Engineering**, MDPI, v. 9, n. 5, p. 496, 2021. Disponível em: <<https://www.mdpi.com/2077-1312/9/5/496>>. Citado 2 vezes nas páginas 6 e 15.