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

LUIZ FERNANDO RODRIGUES FERNANDES

PREDIÇÃO DE RESULTADOS DE UMA PARTIDA DE FUTEBOL UTILIZANDO APRENDIZADO DE MÁQUINA

Ouro Preto, MG
2024

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

LUIZ FERNANDO RODRIGUES FERNANDES

PREDIÇÃO DE RESULTADOS DE UMA PARTIDA DE FUTEBOL UTILIZANDO APRENDIZADO DE MÁQUINA

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

Orientador: Pedro Henrique Lopes Silva

Ouro Preto, MG
2024

SISBIN - SISTEMA DE BIBLIOTECAS E INFORMAÇÃO

F363p Fernandes, Luiz Fernando Rodrigues.
Predição de resultados de uma partida de futebol utilizando aprendizado de máquina. [manuscrito] / Luiz Fernando Rodrigues Fernandes. - 2024.
39 f.: il.: color., tab..

Orientador: Prof. Dr. Pedro Henrique Lopes Silva.
Monografia (Bacharelado). Universidade Federal de Ouro Preto. Instituto de Ciências Exatas e Biológicas. Graduação em Ciência da Computação .

1. Mineração de dados (Computação). 2. Aprendizado do computador. 3. Algoritmo Florestas Aleatórias. 4. Modelos multiníveis (Estatísticas). 5. Futebol. I. Silva, Pedro Henrique Lopes. II. Universidade Federal de Ouro Preto. III. Título.

CDU 004.62

Bibliotecário(a) Responsável: Soraya Fernanda Ferreira e Souza - SIAPE: 1.763.787

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

FOLHA DE APROVAÇÃO

Luiz Fernando Rodrigues Fernandes

Predição de resultados de uma partida de futebol utilizando aprendizado de máquina

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

Aprovada em 14 de Outubro de 2024.

Membros da banca

Pedro Henrique Lopes Silva (Orientador) - Doutor - Universidade Federal de Ouro Preto
Fernando Henrique Oliveira Duarte (Examinador) - Mestre - PPGCC-UFOP
Guilherme Augusto Lopes Silva (Examinador) - Mestre - PPGCC-UFOP

Pedro Henrique Lopes Silva, Orientador do trabalho, aprovou a versão final e autorizou seu depósito na Biblioteca Digital de Trabalhos de Conclusão de Curso da UFOP em 14/10/2024.

Documento assinado eletronicamente por **Pedro Henrique Lopes Silva**, **PROFESSOR DE MAGISTERIO SUPERIOR**, em 16/10/2024, às 09:50, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

A autenticidade deste documento pode ser conferida no site http://sei.ufop.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0 , informando o código verificador **0789786** e o código CRC **9EB201C7**.

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

Este trabalho é uma homenagem especial a minha mãe, Maria Margarete, meu pai, Fernando Abade, e minha irmã, Luiza Cristina. Agradeço profundamente pelo apoio inestimável que eles me proporcionaram ao longo da minha jornada acadêmica, pois sem eles, eu não teria conseguido concluir o meu curso.

Agradecimentos

Desejo expressar minha profunda gratidão, em especial à minha mãe, Maria Margarete, ao meu pai, Fernando Abade, e à minha irmã, Luiza Cristina, que ao longo de toda esta jornada, sempre me ofereceram seu apoio incondicional. Também sou imensamente grato ao meu tio Aroldo e minha tia Rosemere, que não apenas me acolheram, mas me trataram como um de seus próprios filhos. A André Luiz e Filipe Augusto, que se tornaram verdadeiros irmãos para mim. Também agradeço ao meu orientador Professor Pedro, por toda a atenção na correção e detalhes do texto. Por último, mas não menos importante, agradeço à minha namorada, Bárbara Lia, que esteve ao meu lado durante todo o processo de elaboração deste trabalho.

Resumo

O presente trabalho aplica algoritmos de aprendizado de máquina para a predição de resultados de futebol, especialmente no contexto do planejamento de temporada das equipes. A pesquisa propõe a utilização dos algoritmos XGboost, Bayesiano Ingênuo (*Naive Bayes*) e Floresta Aleatória (*Random Forest*), aplicados às bases de dados *Dataset* LaLiga 2021-2022, *Dataset* LaLiga 2022-2023, *Dataset* LaLiga 2023-2024 e LaLiga *Pack Dataset*, construídas através de técnicas *Web Scrapping*. O processo inclui um pré-processamento dos dados para otimização da adequação aos algoritmos, visando obter predições cientificamente embasadas e assertivas. O estudo ressalta a importância dessas abordagens no cenário do futebol, fornecendo valiosos *insights* para o planejamento estratégico das equipes. Alcançando 53% de precisão com o modelo *Naive Bayes*, utilizando as temporadas 2021-2022 e 2022-2023 para treino e a temporada 2023-2024 para teste.

Palavras-chave: Futebol. Aprendizado de máquina. Mineração de dados. XGboost. Bayesiano Ingênuo. Floresta Aleatória. *Web Scrapping*.

Abstract

The present work applies machine learing algorithms for results prediction in football field, especially in the context of team season planning. The research proposes the use of XGBoost, Naive Bayes, and Random Forest algorithms, applied to the datasets LaLiga 2021-2022, LaLiga 2022-2023, LaLiga 2023-2024, and LaLiga Pack Dataset, built using web scraping techniques. The process includes data preprocessing to optimize their suitability for the algorithms, aiming to achieve scientifically grounded and accurate predictions. The study emphasizes the importance of these approaches in football, providing valuable insights for teams' strategic planning. Achieving 53% accuracy with the *Naive Bayes* model, using the 2021-2022 and 2022-2023 seasons for training and the 2023-2024 season for testing.

Keywords: Football. Machine learning. Data mining. XGBoost. Naive Bayesian. Random Forest Web Scrapping.

Lista de Ilustrações

Figura 2.1 – Esportes praticados no Brasil em 2013. 5
Figura 2.2 – Campos científicos cuja integração constitui os fundamentos da mineração de dados. 6
Figura 2.3 – Representação do funcionamento do algoritmo *Random Forest* com um número n de árvores. 10
Figura 2.4 – Representação do funcionamento da validação cruzada 11
Figura 2.5 – Dados utilizados para a predição do resultado derrota em uma partida, utilizando o modelo *Random Forest*. 14
Figura 4.1 – Probabilidade de cada resultado para o time da casa utilizando o modelo *Random Forest* com 5 jogos anteriores. 25
Figura 4.2 – Explicação Local de forma mais detalhada para a predição de derrota no modelo *Random Forest*. 25
Figura 4.3 – Explicação Local de forma mais detalhada para a predição de empate no modelo *Random Forest*. 26
Figura 4.4 – Explicação Local de forma mais detalhada para a predição de vitória no modelo *Random Forest*. 27
Figura 4.5 – Probabilidade de cada resultado para o time da casa utilizando o modelo XGBoost com 5 jogos anteriores. 27
Figura 4.6 – Explicação Local de forma mais detalhada para a predição de derrota no modelo *XGBoost*. 28
Figura 4.7 – Explicação Local de forma mais detalhada para a predição de empate no modelo *XGBoost*. 29
Figura 4.8 – Explicação Local de forma mais detalhada para a predição de vitória no modelo *XGBoost*. 29
Figura 4.9 – Probabilidade de cada resultado para o time da casa utilizando o modelo Naive Bayes com 5 jogos anteriores. 30
Figura 4.10–Explicação Local de forma mais detalhada para a predição de derrota no modelo *Naive Bayes*. 30
Figura 4.11–Explicação Local de forma mais detalhada para a predição de empate no modelo *Naive Bayes*. 31
Figura 4.12–Explicação Local de forma mais detalhada para a predição de vitória no modelo *Naive Bayes*. 32

Lista de Tabelas

Tabela 2.1 – Tabela *Random Forest* representando variáveis nas colunas e eventos indicando se ocorreram ou não nas linhas. 9
Tabela 3.1 – Nomes e tipos dos campos usados na classificação das bases de dados da LaLiga. 18
Tabela 3.2 – Nomes e tipos dos campos adicionados como média nas bases de dados da LaLiga. 19
Tabela 3.3 – Nomes e tipos dos campos adicionados como números de partidas anteriores nas bases de dados da LaLiga. 20
Tabela 3.4 – Nomes e tipos dos campos adicionados como resultados e locais anteriores nas bases de dados da LaLiga. 20
Tabela 3.5 – Campos usados na classificação das bases de dados da LaLiga. 21
Tabela 4.1 – Acurácia em porcentagem (%) dos modelos treinados para as bases de dados com a utilização de 1, 2, 3, 4 e 5 partidas anteriores para cada equipe. Em negrito estão os melhores resultados. 23

Lista de Abreviaturas e Siglas

ANN	Artificial Neural Networks
DECOM	Departamento de Computação
LaLiga	Campeonato Nacional de Liga de Primeira Divisão Espanhola
KDD	Knowledge Discovery in Databases
KNN	K-Nearest Neighbors
MLP	Multi-Layer Perceptron
MMI	Melhor Modelo Individual
NFL	National Football League
SVM	Support Vector Machine
HTML	Hypertext Markup Language

Sumário

1 Introdução 1
1.1 Justificativa 2
1.2 Objetivos 2
1.3 Organização do Trabalho 3
2 Revisão Bibliográfica 4
2.1 Fundamentação Teórica 4
2.1.1 Esportes 4
2.1.2 Futebol 4
2.1.3 Mineração de Dados 6
2.1.4 Aprendizado de Máquina 7
2.1.4.1 Algortimo *Naive Bayes* ou Bayesiano Ingênuo 8
2.1.4.2 Algoritmo *Random Forest* ou Floresta Aleatória 9
2.1.4.3 Algoritmo *XGBoost* 10
2.1.5 Validação Cruzada 11
2.1.6 *Web Scraping* 12
2.1.7 Intepretabilidade 13
2.1.7.1 LIME 13
2.2 Trabalhos Relacionados 14
3 Desenvolvimento 17
3.1 Base de dados 17
3.2 Pré-Processamento 17
3.2.1 *Datasets* LaLiga 17
3.2.2 Ajustes finais 19
3.3 Treinamento 20
4 Resultados 23
4.1 Interpretabilidade do melhor modelo 24
4.1.1 Interpretabilidade *Random Forest* com 5 jogos anteriores 24
4.1.1.1 Análise dos dados para probabilidades de Derrota 25
4.1.1.2 Análise dos dados para probabilidades de Empate 25
4.1.1.3 Análise dos dados para probabilidades de Vitória 26
4.1.2 Interpretabilidade XGBoost com 5 jogos anteriores 27
4.1.2.1 Análise dos dados para probabilidades de Derrota 27
4.1.2.2 Análise dos dados para probabilidades de Empate 28
4.1.2.3 Análise dos dados para probabilidades de Vitória 28
4.1.3 Interpretabilidade *Naive Bayes* com 5 jogos anteriores 30
4.1.3.1 Análise dos dados para probabilidades de Derrota 30

4.1.3.2 Análise dos dados para probabilidades de Empate 31
4.1.3.3 Análise dos dados para probabilidades de Vitória 31
4.2 Discussão dos resultados 32
5 Considerações Finais 34
5.1 Conclusão 34
5.2 Trabalhos Futuros 35

Referências 36

1 Introdução

O universo esportivo, especificamente o futebol, tem testemunhado uma revolução significativa nas últimas décadas, impulsionada pelo avanço tecnológico e pela crescente disponibilidade de dados (Gazeta Esportiva, 2021). Nesse contexto, a relevância da ciência de dados no futebol transcende a simples análise numérica. Ela revoluciona a perspectiva dos clubes sobre o jogo e as oportunidades de conquistar vantagens competitivas. Através de uma análise estatística aprofundada, é possível discernir padrões de jogo, explorar as vulnerabilidades dos adversários e tomar decisões mais fundamentadas acerca de escalações, substituições e estratégias de jogo (Matias, 2009). A análise do desempenho individual dos jogadores no futebol representa uma das aplicações mais notáveis da ciência de dados. Utilizando métricas como passes precisos, chutes a gol e desarmes, é viável avaliar o rendimento de cada atleta, destacando suas habilidades em áreas específicas do jogo. Essas informações não apenas servem para avaliar jogadores já contratados, mas também desempenham um papel crucial na tomada de decisões relacionadas a contratações futuras (Awari, 2023).

Sendo o futebol um dos esportes mais populares do mundo e atraindo atenção de pessoas variadas, muitos estudos sobre ele foram feitos ao longo do tempo. Estudos como (Tüfekci, 2016), (Carpita, 2015) que foram realizados para predizer resultados de partidas do Campeonato Inglês (*English Premier League*). Há o trabalho de Deus (2019) que tentou prever os resultados de partidas de várias ligas de países da Europa, como, por exemplo, a Liga Belga (*Belgium Jupiler League*), a Liga Francesa (*League* 1), entre outras, usando algoritmos de inteligência artificial e aprendizado de máquina.

Por mais que o objetivo do futebol seja sempre alcançar um resultado positivo, não existe uma formação tática que seja considerada superior a todas as outras. Afinal, uma partida entre duas equipes leva em consideração fatores como as capacidades físicas de cada jogador, o entendimento tático do que foi solicitado pelo treinador e o estado mental de cada atleta. Vale ressaltar que nem sempre os favoritos garantem a vitória em um campeonato ou partida. Um exemplo notável é o *Leicester City*, time inglês que quase sofreu o rebaixamento na temporada 2014-2015, mas conseguiu se sagrar campeão na temporada seguinte, em 2015-2016 (Diniz, 2018).

Um aspecto crucial para aprimorar a precisão das previsões de resultados reside na mineração de dados, que viabiliza a extração de informações em larga escala, particularmente em cenários nos quais abordagens tradicionais seriam impraticáveis (Neves, 2022).

Diversos algoritmos de aprendizado de máquina podem ser empregados na tentativa de prever resultados de partidas de futebol. Neste trabalho, optou-se pela utilização dos algoritmos Floresta Aleatória (*Random Forest*) e XGBoost, ambos versáteis, podendo ser aplicados tanto

em tarefas de classificação quanto de regressão. Contudo, neste estudo, eles serão empregados especificamente para tarefas de classificação (Lima e Amorim, 2020). Além disso, será incluído o algoritmo Bayesiano Ingênuo (*Naive Bayes*), reconhecido por seu enfoque em classificação (Sacramento, 2023).

O objetivo deste trabalho é explorar o campo da predição de resultados no futebol, empregando técnicas avançadas de aprendizado de máquina, *web scraping* e mineração de dados. A classificação foi realizada no contexto de uma competição de futebol espanhola, com o objetivo de prever o resultado dos jogos, indicando se o time da casa ganhou, perdeu ou empatou. O modelo atingiu uma taxa de acerto de 50% nos testes, o que supera significativamente a taxa de acerto esperada por pura aleatoriedade, que seria de 33,3%. Portanto, os resultados demonstram que o modelo conseguiu prever os resultados com uma precisão superior ao acaso.

1.1 Justificativa

Para além de sua posição como um dos esportes mais populares globalmente, o futebol também exerce um impacto considerável no âmbito financeiro. Em 2022, o setor movimentou a expressiva quantia de 286 bilhões de dólares, equivalente ao Produto Interno Bruto (PIB) da Finlândia (Moreira, 2022). A forma como os clubes de futebol estruturam seus planos para cada temporada tem uma influência direta nos resultados obtidos ao longo das partidas. Por essa razão, observa-se uma crescente disponibilidade de bancos de dados contendo informações relevantes sobre partidas e estatísticas dos clubes de futebol.

Ao empregar o aprendizado de máquina e os algoritmos contemporâneos, as equipes esportivas têm a oportunidade de otimizar a alocação de recursos, identificando áreas em que superam ou ficam aquém das expectativas ao longo da temporada. Esse enfoque não apenas fornece *insights* valiosos para as equipes, mas também pode ter repercussões positivas no engajamento dos torcedores. A capacidade de oferecer uma análise detalhada sobre o desempenho da equipe pode servir como fonte de motivação para os fãs, incentivando-os a participar ativamente, comparecer aos estádios, adquirir produtos relacionados, e, consequentemente, fortalecer o apoio à sua equipe favorita. Além disso, tal abordagem pode atrair a atenção de patrocinadores, proporcionando uma visão mais favorável a equipes que buscam se destacar em um cenário esportivo cada vez mais competitivo.

1.2 Objetivos

O propósito principal deste estudo é empregar os métodos de aprendizado de máquina *Random Forest*, *Naive Bayes* e XGBoost, especificamente em dados de partidas anteriores para realizar previsões dos resultados de futuras partidas de futebol, utilizando como fundamentos os desfechos passados e estatísticas esportivas.

Os objetivos específicos são:

- Coletar a base de dados usada neste trabalho.
- Avaliar técnicas de normalização e limpeza de dados.
- Avaliar método de *Random Forest* para predição do resultado de partidas.
- Avaliar método de XGBoost para predição do resultado de partidas.
- Avaliar método de *Naive Bayes* para predição do resultado de partidas.
- Avaliar as melhores métricas para comparar os modelos treinados.

1.3 Organização do Trabalho

O Capítulo 2 (Revisão Bibliográfica), apresenta a fundamentação teórica deste trabalho em conjunto com os trabalhos relacionados. Na sequência o Capítulo 3 (Desenvolvimento) abordará quais os materiais e métodos foram utilizados para dar continuidade a esse estudo. O Capítulo 4 (Resultados) descreve quais resultados foram alcançado após aplicar e testar as teorias propostas. Por último o Capítulo 5 (Considerações Finais) trará as conclusões e trabalhos futuros.

2 Revisão Bibliográfica

Este capítulo abrange a fundamentação teórica do trabalho, proporcionando uma introdução aos esportes, com ênfase no futebol. Além disso, introduz o conceito de mineração de dados e explora a aplicação de aprendizado de máquina. A explicação dos algoritmos específicos utilizados no desenvolvimento do trabalho é apresentada nesta seção. Por fim, são discutidos os trabalhos relacionados que foram analisados como parte do processo de pesquisa para este trabalho.

2.1 Fundamentação Teórica

Esta seção apresenta uma introdução ao esporte, com foco especial no futebol, além de abordar a mineração de dados e o aprendizado de máquina. Também serão discutidos os algoritmos utilizados na tentativa de prever resultados de jogos e eventos esportivos.

2.1.1 Esportes

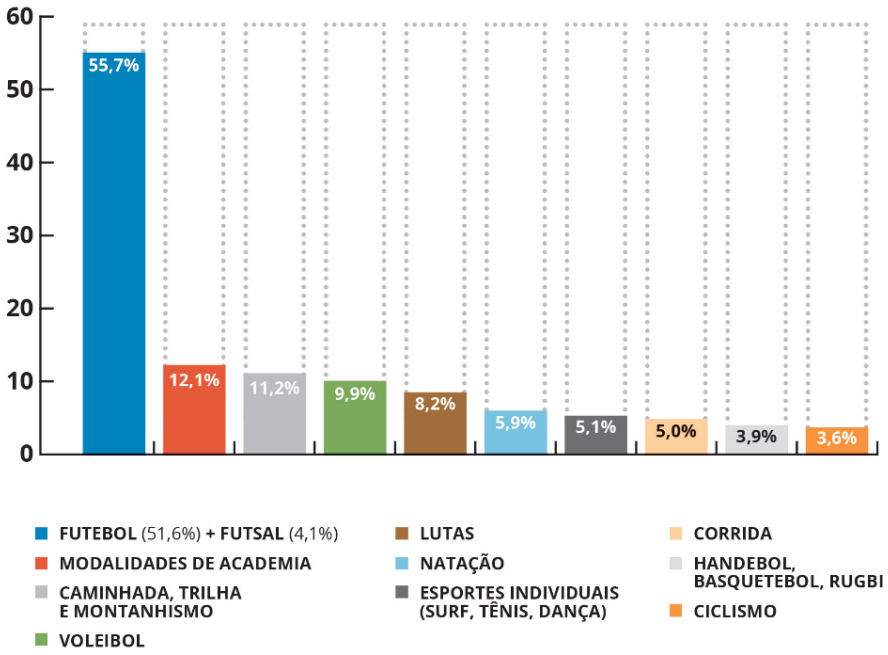
Os esportes são atividades físicas competitivas ou recreativas que envolvem um conjunto de regras e regulamentos, normalmente praticados individualmente ou em equipes, com o objetivo de melhorar a aptidão física, alcançar metas esportivas ou simplesmente se divertir. Eles são uma parte importante da cultura e da sociedade em muitas partes do mundo, envolvendo uma ampla variedade de atividades em termos de intensidade, habilidade e popularidade (Kravchychyn *et al.*, 2012).

Conforme ilustrado na Figura 2.1, em 2013, um dos esportes mais praticados no Brasil é o futebol, com uma considerável margem de liderança sobre o segundo colocado em termos de popularidade. Sua popularidade expressiva em relação aos demais esportes faz dele uma escolha relevante para investigações e análises. A vasta participação e interesse da população no futebol motivam a escolha desse esporte como foco deste estudo, proporcionando uma base sólida e representativa para explorar aspectos relacionados à predição de resultados de partidas. Essa decisão é respaldada pela importância cultural e social do futebol no contexto brasileiro, que influencia não apenas a esfera esportiva, mas também outros aspectos da vida cotidiana.

2.1.2 Futebol

O futebol tem raízes antigas e diversificadas em diferentes partes do mundo, com práticas semelhantes documentadas em várias culturas ao longo da história. Um exemplo notável vem da China em 206 a.C., onde há relatos de um manual que regulamentava um jogo estranho parecido com o futebol, já praticado 2500 anos atrás, no império Huang-Ti (Souza, 2013). Essas

Figura 2.1 – Esportes praticados no Brasil em 2013.



FONTE: (Esporte, 2013).

manifestações iniciais, que apresentam semelhanças com o futebol moderno, ilustram como a paixão por esse esporte tem uma história rica e variada, que se espalhou por diferentes continentes ao longo do tempo.

A FIFA, do italiano, *Fédération Internationale de Football Association*, é a entidade máxima do futebol mundial, encarregada de governar e regular o esporte em escala global. A FIFA é composta por associações nacionais de futebol de todo o mundo, cada uma representando seu respectivo país. Ela é liderada por um presidente e um comitê executivo, realizando congressos regulares para tomar decisões importantes no cenário do futebol mundial. A organização desempenha um papel fundamental na promoção e regulamentação do esporte, tornando-se uma instituição influente no cenário esportivo internacional.

Além disso, é importante destacar que o futebol é uma indústria que movimenta bilhões de dólares em transações entre clubes e é uma fonte significativa de receita para casas de apostas esportivas em todo o mundo. A proposta de previsão de resultados não apenas desempenha um papel crucial na preparação das equipes para futuros confrontos, mas também oferece benefícios significativos na avaliação global do desempenho da equipe. Permite que as equipes determinem se estão superando ou ficando aquém das metas estabelecidas para a temporada. Além de seu valor interno, a previsão de resultados emerge como uma ferramenta valiosa na interação com os patrocinadores, proporcionando uma visão clara do potencial da equipe ao longo do ano, incentivando, assim, o apoio financeiro por parte dos patrocinadores.

Não se limitando a seu valor interno, a proposta de previsão de resultados também agrega

um valor adicional aos sócios-torcedores. Ao fornecer *insights* sobre o desempenho previsto da equipe, os sócios-torcedores obtêm uma compreensão mais profunda do potencial do time. Isso fortalece não apenas o vínculo emocional entre os torcedores e a equipe, mas também fornece uma base informada para o engajamento da torcida. Os sócios-torcedores podem sentir-se mais conectados ao acompanhar de perto as metas e expectativas da equipe, contribuindo assim para uma comunidade de torcedores mais envolvida e informada.

2.1.3 Mineração de Dados

O processo chamado de mineração de dados que por vezes também é chamada de descoberta de conhecimento em dados vem sendo utilizado há muito tempo e sua base possui três disciplinas científicas entrelaçadas que já existem há tempos (SAS, 2023). A Figura 2.2 mostra que estatística (*statistics*), inteligência artificial (*artificial intelligence*) e aprendizado de máquina (*machine learning*) são as bases usadas em Mineração de Dados.

Figura 2.2 – Campos científicos cuja integração constitui os fundamentos da mineração de dados.

FONTE: (SAS, 2023).

As técnicas de mineração de dados que fundamentam essas análises podem ser categorizadas em dois propósitos principais: a primeira é descrever o conjunto de dados de interesse, enquanto a segunda visa a previsão de resultados, utilizando algoritmos de *machine learning*. Esses métodos são empregados para extrair os dados mais relevantes em relação aos interesses e objetivos do usuário ou pesquisador em questão (IBM, 2023b).

O processo de mineração de dados abrange diversas etapas, desde a coleta inicial dos dados até a sua visualização, com o objetivo de extrair informações valiosas de vastos conjuntos de dados. Conforme mencionado anteriormente, as técnicas de mineração de dados são aplicadas para criar descrições e previsões relacionadas a um conjunto de dados específico. Os cientistas de dados exploram os dados por meio da identificação de padrões, associações e correlações

significativas. Além disso, eles classificam e agrupam os dados utilizando métodos de classificação e regressão, e identificam valores discrepantes em cenários como a detecção de *spam* (IBM, 2023b).

Os seguintes passos delineiam cada uma das fases a serem seguidas para realizar a mineração de dados.

- Definição dos objetivos.
- Preparação dos dados.
- Mineração de padrões.
- Avaliação dos resultados e aplicação do conhecimento.

Ao compreender o conceito de mineração de dados e as fases a serem seguidas nesse processo, torna-se mais fácil entender um dos elementos cruciais para aplicar a predição de resultados neste trabalho.

2.1.4 Aprendizado de Máquina

Aprendizado de máquina é um ramo da inteligência artificial que busca sintetizar as relações entre dados e informações (Awad; Khanna, 2015). Ele vem sendo usado em muitas áreas nos dias de hoje algumas delas são:

- Cruzamento de dados para detectar problemas de saúde: A tecnologia é capaz de identificar e interpretar padrões em exames de pacientes para perceber doenças que ainda não poderiam ser detectadas pelas atuais ferramentas de diagnóstico (Habehh; Gohel, 2021).
- Economia de energia em empresas: Se conseguir acesso aos dados de como o consumo de energia é realizado dentro da empresa, *machine learning* pode ser utilizado para minimizar esses gastos, indicando formas mais conscientes e inteligentes de uso dos recursos elétricos (Matrenin; Antonenkov; Arestova, 2021).
- Recomendação de produtos extras para clientes: No Walmart, nos Estados Unidos, *machine learning* foi usado para descobrir que quem compra fraldas no supermercado também tende a levar cervejas, e sugeriu colocar os dois no mesmo corredor, conseguindo, assim, vender mais (Santos, 2023).

Há uma variedade de algoritmos para prever o resultado de eventos específicos, como o vencedor de um jogo de futebol, ou classificar se um atleta é promissor, assim como a probabilidade de uma lesão ocorrer (Apostolou; Tjortjis, 2019). Não existe um algoritmo universalmente superior aos demais em todas as situações. Diferentes algoritmos se destacam em diferentes tipos

de problemas. Portanto, é essencial realizar uma análise cuidadosa para identificar qual algoritmo oferece o melhor desempenho em uma situação específica.

Um dos diversos métodos disponíveis são os algoritmos de aprendizagem supervisionada, que estabelecem uma relação entre a saída e a entrada com base em dados rotulados (Fontana, 2018a). Nesse cenário, o usuário alimenta o algoritmo com conjuntos conhecidos de entradas e saídas, geralmente expressos como vetores. Cada saída é associada a um rótulo, que pode ser um valor numérico ou uma classe. O algoritmo, então, adquire a capacidade de prever qual rótulo de saída corresponde a uma entrada específica (Fontana, 2018b).

Dentre os principais algoritmos de aprendizado de máquina para classificação, destacam-se:

- *Naive Bayes*.
- *Random Forest*.
- XGBoost.

2.1.4.1 Algortimo *Naive Bayes* ou Bayesiano Ingênuo

O *Naive Bayes* é um classificador probabilístico simples que aplica o teorema de Bayes com a forte suposição de independência entre as características, de modo que a presença de uma característica individual de uma classe não está relacionada com a presença de outra característica.

Esse algoritmo foi inspirado no Teorema de Bayes, teorema esse que foi desenvolvido por Thomas Bayes (1701-1761) para tentar provar a existência de Deus (Becker, 2019). A Equação (2.1) representa a fórmula do Teorema de Bayes e pode ser explicada da seguinte maneira (Cypriano, 2015):

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)} \tag{2.1}$$

onde:

- $P(B|A)$: é a probabilidade do evento B ocorrer dado que A ocorreu.
- $P(A|B)$: é a probabilidade do evento A ocorrer dado que B ocorreu.
- $P(A)$: é a probabilidade do evento A ocorrer ocorrer.
- $P(B)$: é a probabilidade do evento B ocorrer ocorrer.

O modelo *Naive Bayes* é eficaz e muito atrativo devido à sua simplicidade e robustez (Gomes, 2019). Uma das suas vantagens é que não requer a aplicação de esquemas complexos para estimar parâmetros iterativos em grandes conjuntos de dados, fazendo com que ele seja

muito útil e relativamente fácil de construir e usar. É um algoritmo popular em áreas relacionadas à classificação de texto e filtragem de spam (Awad; Khanna, 2015).

2.1.4.2 Algoritmo *Random Forest* ou Floresta Aleatória

O *Random Forest* é um algoritmo de aprendizado supervisionado. Portanto, é necessário escolher uma variável dependente, isto é, a variável que o algoritmo tentará prever. O *Random Forest* cria várias árvores de decisão de forma aleatória e as combina para chegar a um resultado final, revelando a resposta da previsão (Junior, 2021).

No exemplo da Tabela 2.1 a seguir, a variável escolhida para ser prevista foi a variável A.

Tabela 2.1 – Tabela *Random Forest* representando variáveis nas colunas e eventos indicando se ocorreram ou não nas linhas.

Variável A	Variável B	Variável C
SIM	NÃO	SIM
NÃO	SIM	NÃO
SIM	NÃO	SIM

Fonte: Próprio autor.

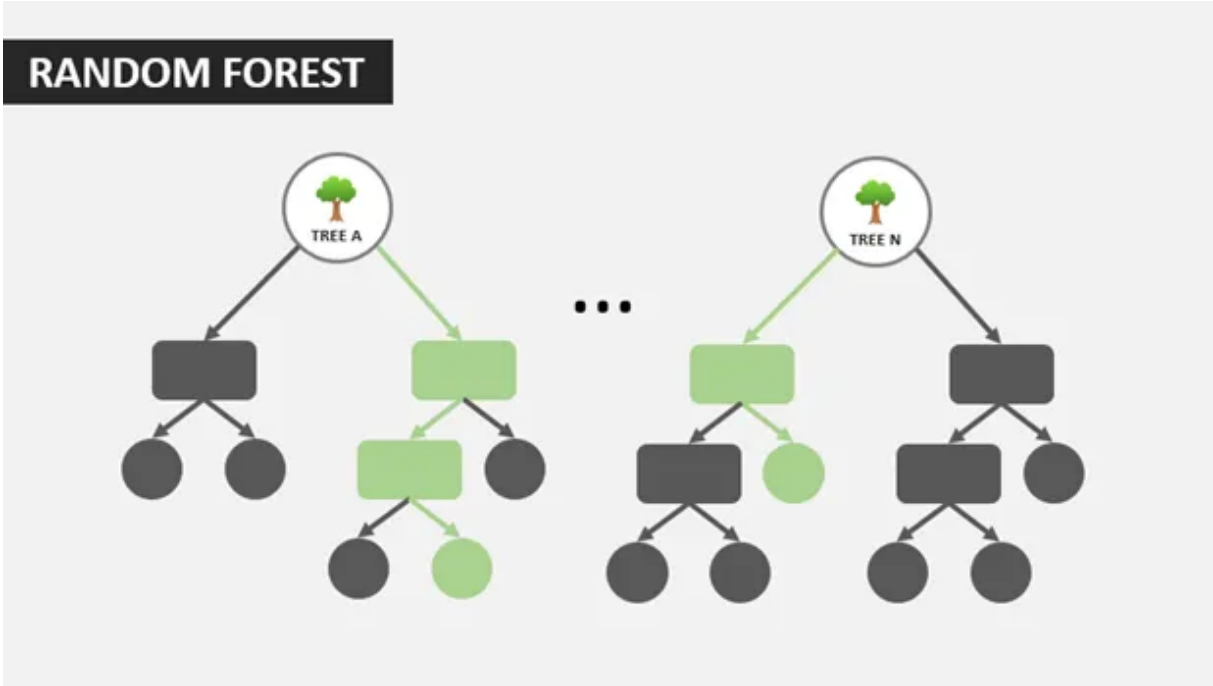
Com base na Tabela 2.1, se a "Variável B" for `NÃO` e a "Variável C" for `SIM` na primeira linha, a decisão tomada em relação à "Variável A" deveria ser `SIM`. Isso se aplica de maneira análoga para a segunda e terceira linha; ou seja, com base nas variáveis B e C, uma resposta deve ser prevista para a variável A. O algoritmo inicia dividindo o conjunto de dados original de forma aleatória em um novo conjunto de treinamento. Em seguida, cria várias árvores de decisão com base nas variáveis independentes, que podem ser "Variável B" ou "Variável C". Cada nó em cada árvore representa uma condição que será analisada para a tomada de decisão. Após a satisfação de cada condição, o algoritmo realizará a previsão da variável A.

O modelo cria várias árvores para evitar o *overfitting*, que ocorre quando o modelo não consegue generalizar e se especializa demais nos dados de treinamento. Isso significa que o modelo memorizou os resultados dos dados em vez de aprender com eles. Quando exposto a novos dados, nesse cenário, o modelo terá taxas de precisão muito baixas em suas previsões (Junior, 2021).

Devido à natureza aleatória do processo de seleção de nós, cada nova árvore será sempre diferente das árvores anteriormente criadas.

Com o modelo devidamente criado, é possível inserir um novo conjunto de dados para que ele faça previsões da variável A. O modelo analisará todas as árvores e coletará as decisões de cada uma delas. No nosso caso, como estamos lidando com um problema de classificação em que a variável A é binária, ou seja, '`SIM`' ou '`NÃO`', o modelo coletará o resultado mais frequentemente observado entre as árvores, que será apresentado como a previsão. No entanto, em problemas de regressão, a escolha será a média dos resultados e ela será apresentada como resultado.

Figura 2.3 – Representação do funcionamento do algoritmo *Random Forest* com um número n de árvores.



FONTE: (Pessanha, nov. 2020).

2.1.4.3 Algoritmo *XGBoost*

Para falar sobre o algoritmo *XGBoost*, primeiro torna-se necessário entender os conceitos de *Boosting* e *Gradient Boosting*, uma vez que esse algoritmo se baseia principalmente nestes dois fatores:

- *Boosting* é uma abordagem em que o avaliador ajusta os critérios de avaliação com base nas informações do avaliador anterior, aumentando assim a eficiência do processo de aprendizagem. Isso resulta em um processo de aprendizado mais dinâmico e adaptativo (Gomes, jun. 2019).
- *Gradient Boosting* é uma técnica de aprendizado de máquina que cria modelos utilizando um conjunto de previsões fracas. É construído um modelo com vários métodos de reforço, permitindo uma otimização com relação a uma perda diferenciável arbitrária. Os ajustes de cada modelo fraco são multiplicados por uma taxa de aprendizado, que determina o impacto de cada árvore no modelo final. Quanto menor o valor da taxa de aprendizado, menor a contribuição da árvore específica (Silva, 2020).

O *Boosting* opera de maneira iterativa, identificando pontos de dados mal classificados e ajustando seus pesos para minimizar o erro de treinamento. O modelo continua otimizando sequencialmente até produzir o preditor mais complexo. O *Gradient Boosting* funciona ao incluir preditores sequencialmente em um conjunto, com cada um deles corrigindo os erros de seus

predecessores. Em vez de alterar os pesos dos pontos de dados, como é feito pelo *Boosting*, o *Gradient Boosting* treina com base nos erros residuais do preditor anterior. O termo *Gradient Boosting* é utilizado porque combina o algoritmo de descida do gradiente e o método de *boosting*. O *XGBoost* é uma implementação do *Gradient Boosting* projetada para velocidade e escala computacional. Ele utiliza vários núcleos na CPU, permitindo que o aprendizado ocorra em paralelo durante o treinamento (IBM, 2023a).

Em resumo, o *XGBoost* é uma ferramenta poderosa para a construção de modelos de *Machine Learning* de alta qualidade. Ele possui recursos avançados, tornando-o uma escolha popular entre cientistas de dados e entusiastas de *Machine Learning*.

2.1.5 Validação Cruzada

A validação cruzada, *ou Cross Validation,* é uma técnica empregada para avaliar modelos de aprendizado de máquina. Essa abordagem consiste em treinar diversos modelos utilizando subconjuntos distintos dos dados de entrada disponíveis e, subsequentemente, avaliá-los nos conjuntos de dados complementares. O objetivo primário da validação cruzada é detectar possíveis problemas de sobreajuste, indicando a capacidade limitada do modelo de generalizar padrões para dados não utilizados durante o treinamento(Amazon, 2023).

Na técnica de validação cruzada, o conjunto de dados é aleatoriamente particionado em "K" grupos. Ao especificarmos um valor para "K", utilizamos esse número para denotar a quantidade de grupos, como exemplificado na validação cruzada com 10 subconjuntos, onde "K" representa o teste em referência (Savietto, 2021).

Figura 2.4 – Representação do funcionamento da validação cruzada

FONTE: (Rosaen, 2016).

- Como exemplificado na Figura 2.4 o conjunto de dados será particionado em 10 grupos.

- Será conduzida uma iteração para cada grupo, em que o grupo selecionado é empregado para fins de teste, enquanto os outros nove grupos são utilizados para o treinamento do modelo.

- Realiza-se o treinamento do modelo utilizando os dados de treinamento, o testa com os dados de teste, registra-se o valor da métrica e posteriormente descarta-se o modelo.

Para obter o resultado final, é comum calcular a média aritmética dos resultados individuais de cada teste. O *k-fold cross validation* é uma abordagem eficaz para avaliar o desempenho do modelo diante de variações nos conjuntos de treinamento. Testes empíricos em aplicações de aprendizado de máquina indicam que valores bastante confiáveis para "K" são 5 e 10, pois esses valores minimizam tanto os erros de teste quanto a variabilidade (Savietto, 2021).

2.1.6 *Web Scraping*

Web Scraping, também conhecido como raspagem de dados, é uma técnica que automatiza a extração de informações de sites da internet (Data Geeks, 2024).

Essa técnica é utilizada para obter dados que normalmente seriam coletados manualmente por um usuário. Ferramentas de *Web Scraping* replicam o comportamento de um usuário navegando em um site, acessando diferentes páginas e extraindo dados específicos de forma eficiente (Data Geeks, 2024).

O *Web Scraping* é amplamente utilizado em várias indústrias, como o *e-commerce* e o mercado imobiliário. Essa técnica é altamente eficaz para identificar oportunidades de inovação nos negócios e entender as necessidades do público-alvo (Data Geeks, 2024). Os dados coletados por meio de *Web Scraping* podem ser armazenados em bancos de dados ou utilizados para criar visualizações que proporcionem *insights* valiosos.

O *Web Scraping* é uma ferramenta poderosa com grande potencial para otimizar estratégias empresariais e de pesquisa por meio da coleta de dados. No entanto, é fundamental que os profissionais a utilizem de forma ética e responsável, levando em conta tanto as questões legais quanto o impacto sobre terceiros (Data Geeks, 2024).

Com o avanço da era dos dados, o *Web Scraping* tende a se tornar ainda mais essencial, mas também sujeito a regulamentações mais rigorosas (Data Geeks, 2024).

Segundo Data Geeks (2024), algumas técnicas de *web scraping* são:

- Automatização de navegação: Utiliza *bots* ou *scripts* que navegam pela *web* de forma automática.

- Extração de dados: Os dados são extraídos de elementos HTML específicos de uma página da *web*.

- Processamento de dados: Os dados coletados podem ser processados, limpos e formatados para análise ou armazenamento em um banco de dados.

2.1.7 Intepretabilidade

Com o avanço dos modelos de aprendizado de máquina e o aumento de sua complexidade, técnicas de interpretabilidade se tornaram importantes para o entendimento dos resultados obtidos. Ela se tornou fundamental para obter além da compreensão dos resultados, *insights* mais detalhados e compreensíveis sobre os resultados gerados.

A interpretabilidade dos modelos de aprendizado de máquina é especialmente relevante para aqueles utilizados na previsão e análise de problemas com impacto direto na vida real. Ela oferece uma ferramenta essencial para uma melhor compreensão e tomada de decisões(DNC, 2023).

A interpretabilidade é essencial em diversas áreas, como:

Medicina: Na medicina, a interpretabilidade é crítica. Por exemplo, em diagnósticos baseados em imagens médicas, como radiografias ou ressonâncias magnéticas, a interpretabilidade pode ajudar um médico a entender e confiar nas previsões do modelo, melhorando a precisão do diagnóstico e o tratamento do paciente (DIO.ME, 2024).

Finanças: No setor financeiro, a interpretabilidade desempenha um papel crucial na detecção de fraudes, concessão de empréstimos e avaliação de riscos. Modelos interpretáveis permitem que os analistas compreendam e justifiquem as decisões do modelo, aprimorando a transparência e a confiabilidade dos processos de tomada de decisão (DIO.ME, 2024).

A interpretabilidade em modelos de *machine learning* tem se tornado uma área de pesquisa e prática em rápido crescimento, com importantes implicações para a confiança, equidade e transparência em aplicações críticas. À medida que esses modelos se tornam cada vez mais integrados à nossa sociedade e aos processos de tomada de decisão, a demanda por interpretabilidade tende a crescer ainda mais (DIO.ME, 2024).

2.1.7.1 LIME

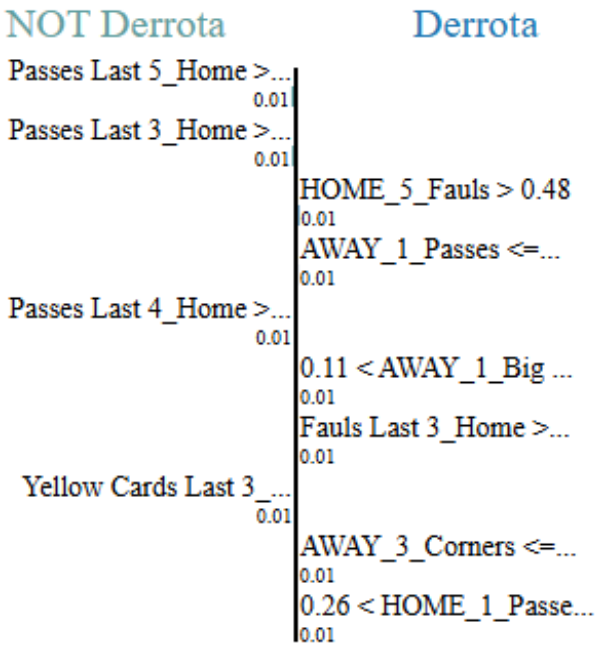
O LIME (*Local Interpretable Model-Agnostic Explanations*) é uma técnica amplamente utilizada na interpretabilidade de modelos. Ela permite explicar de forma mais clara quais dados o modelo está utilizando para fazer as previsões durante a execução dos testes com algoritmos de aprendizado de máquina (DNC, 2023).

Ele funciona gerando novas amostras ao redor de uma amostra real, obtendo as previsões do modelo para essas novas amostras e, a partir disso, gera dados locais em torno da amostra

original. Isso permite uma interpretação mais simples, facilitando ajustes nas *features* utilizadas pelo modelo(DNC, 2023).

Um exemplo do funcionamento do LIME pode ser observado na Figura 2.5, onde são demonstrados os dados utilizados para a predição de derrota ou de um outro resultado, chamado de 'não derrota', que pode ser uma vitória ou um empate. Isso possibilita compreender como o modelo está utilizando as métricas para chegar à sua conclusão, fornecendo uma explicação transparente sobre a tomada de decisão ao se aplicar um modelo de aprendizado de máquina.

Figura 2.5 – Dados utilizados para a predição do resultado derrota em uma partida, utilizando o modelo *Random Forest*.



Fonte: Criada pelo autor.

Ao analisar a Figura 2.5, é possível notar que variáveis como a média de faltas cometidas pelo time da casa nas últimas 5 partidas, além do número de passes e escanteios obtidos pelo time visitante, influenciam o modelo, aumentando a chance de previsão de derrota. Por outro lado, o número de passes realizados pela equipe mandante contribui para aumentar a probabilidade de um resultado diferente da derrota.

2.2 Trabalhos Relacionados

Nesta seção, são apresentados estudos que aplicaram técnicas de aprendizado de máquina e mineração de dados para prever resultados de partidas esportivas, especialmente no futebol.

Nabinger (2018) avaliou quais variáveis eram as mais importantes para determinar o resultado positivo de uma das 64 partidas disputadas na Copa do Mundo de 2018, considerando os dados disponíveis no site da FIFA. O estudo avaliou os métodos *Random Forest*, LASSO, *Boosting* e MARS, concluindo que as variáveis mais importantes para prever os resultados das

partidas da Copa do Mundo de 2018 foram a posse de bola, o número de finalizações e a distância percorrida pela equipe. Além disso, o estudo demonstrou que os algoritmos foram capazes de prever com precisão os resultados das partidas da Copa do Mundo de 2018. Conforme a análise realizada, a precisão dos modelos apresentou variações significativas de acordo com o método empregado. O *Random Forest* obteve o desempenho mais modesto, acertando aproximadamente 43,75% dos resultados. Em contraste, o método LASSO alcançou uma taxa de acerto de 59,375%, enquanto o *Boosting* registrou um aproveitamento de 56,25%. Destaca-se que o modelo mais eficaz foi o MARS, atingindo uma precisão de 60,9375%.

Em (Deus, 2019), adotou-se técnicas de mineração de dados e aprendizado de máquina com o propósito de antecipar resultados de partidas e predizer os times vencedores em campeonatos de futebol. A metodologia utilizada foi baseada no *Knowledge Discovery in Databases* (KDD), que englobou as fases de identificação e coleta de dados, pré-processamento, mineração de dados e extração de conhecimento. Além disso, empregou-se uma variedade de algoritmos de aprendizado de máquina, incluindo Regressão Linear, *Naive Bayes*, Redes *Multi-Layer Perceptron* (MLP) e *Support Vector Machine* (SVM), responsáveis por gerar resultados tanto para tarefas de regressão quanto classificação. A avaliação dos resultados foi conduzida através do Teste T de *Student*, focando na Acurácia e no Coeficiente de Correlação. Os resultados revelaram aproximadamente 60% de precisão para os classificadores, com um coeficiente de correlação em torno de 0,55 para regressão. Destaca-se uma ligeira vantagem das SVMs e MLPs como classificadores, enquanto observou-se um empate técnico geral na regressão.

No estudo apresentado em (Bernardes, 2023), foram empregadas técnicas de Aprendizado de Máquina para prever os resultados dos *playoffs* da NFL (National Football League) um jogo originado nos Estados Unidos, utilizando dados dos últimos 51 anos de jogos. Foram considerados 262 atributos estatísticos distribuídos em três grandes subgrupos: dados estatísticos das partidas (“gs”), dados estatísticos do time da casa (“hts”) e dados estatísticos do time desafiante (“ats”). Esses dados foram extraídos do site *Pro-Football-Reference* [1], que compila estatísticas sobre o futebol americano. Diversos algoritmos de aprendizado de máquina foram modelados, incluindo *Random Forest*, SVM, *Naive Bayes*, *K-Nearest Neighbors* (KNN) e *Artificial Neural Networks* (ANN). Os dados foram divididos em conjuntos de treino e teste (80% treino, 20% teste) usando o método *hold-out*. Foi realizado o *tuning* dos parâmetros internos de cada modelo. As métricas de avaliação incluíram acurácia, acurácia balanceada, precisão, *recall* e *F1-score*. Na segunda fase, foram conduzidas análises temporais considerando a evolução dos times durante a temporada. A acurácia variou conforme o modelo de Aprendizado de Máquina utilizado, sendo que o modelo com ANN obteve o melhor desempenho, alcançando 63,5%, seguido por *Random Forest* e *Naive Bayes* com acurácias de 56,9%. O modelo com (SVM) obteve uma acurácia de 53,4%. A análise temporal revelou ganhos marginais em todas as métricas, indicando potencial significativo nesse tipo de análise.

[1] Disponível em: <https://www.pro-football-reference.com/>

Schmidt (2017) aplicou técnicas de aprendizado de máquina, como *Random Forest*, SVM e ANN, para a previsão de resultados de partidas de futebol. Adicionalmente, procedimentos de importação, limpeza e transformação de dados estatísticos sobre partidas de futebol foram conduzidos, juntamente com a seleção de atributos a serem empregados no programa de previsão e a validação dos resultados obtidos pelo programa em comparação com os resultados reais das partidas. Os dados estatísticos utilizados englobaram informações sobre partidas de equipes do campeonato inglês, abrangendo resultados e estatísticas desde a temporada 1993/1994 até a realização do estudo em 2017. Esses dados foram obtidos do site <www.football-data.co.uk>. Além disso, informações sobre habilidades de clubes e jogadores do campeonato inglês, provenientes do jogo de computador Fifa, foram incorporadas ao estudo a partir do site <futhead.com>. A acurácia dos métodos empregados na previsão de resultados de partidas de futebol variou de acordo com as técnicas utilizadas. Conforme mencionado no estudo, a *Random Forest* alcançou uma taxa de acerto de 70,87%, enquanto as ANNs apresentaram uma acurácia de 68,80%. Adicionalmente, o SVM obteve uma taxa de acerto de 67,88%. Esses resultados evidenciam as diferentes performances das técnicas utilizadas no contexto da previsão de resultados esportivos.

Em Fernandes (2019), diversas técnicas e métodos de aprendizado de máquina foram empregados, incluindo regressão logística, árvores de decisão, ANNs, SVM, *Random Forest*, *Gradient Boosting* e *Naive Bayes*. A análise dos resultados envolveu diversas estatísticas das partidas de futebol, como o número de gols marcados por cada equipe, posse de bola, chutes a gol, faltas cometidas, e cartões amarelos e vermelhos. Esses dados foram extraídos de uma base de dados contendo informações de mais de 25 mil jogos de futebol de onze ligas europeias entre 2008 e 2016. Os métodos adotados apresentaram diferentes níveis de acurácia na predição de resultados de partidas de futebol. Por exemplo, o "Modelo Ingênuo", que classifica sempre o resultado com base no favorito, determinado pelo prêmio pago por aposta da *Bet365*, alcançou uma acurácia de 54.16% e uma métrica F1-score de 0.463. O modelo MMI (Melhor Modelo Individual) registrou uma acurácia de 52.78% nos dados de teste e 52.00% nos dados de avaliação. Comparando com a métrica F1-score do MMI, 0.509, em relação ao Modelo Ingênuo, observa-se uma melhora de 9.99%.

Ao comparar a proposta desta pesquisa com os trabalhos relacionados, é evidente que a maioria deles abordou a previsão de resultados de jogos, principalmente no contexto do futebol. Muitos desses estudos justificaram suas pesquisas com base no mercado de apostas esportivas. Além disso, alguns dos algoritmos empregados nos trabalhos relacionados são semelhantes aos que serão utilizados nesta pesquisa, com o algoritmo *Naive Bayes* sendo um exemplo disso.

3 Desenvolvimento

Neste capítulo, serão delineadas as bases de dados empregadas para o treinamento e teste do modelo, elucidando o processo de pré-processamento aplicado a cada uma delas. Detalharemos os ajustes finais necessários para otimizar a utilidade das bases de dados nos testes subsequentes. Por fim, apresentaremos a abordagem adotada no treinamento, utilizando os algoritmos de aprendizado de máquina mencionados anteriormente. Este capítulo é crucial para compreender a robustez e a eficácia do modelo, pois destaca as escolhas metodológicas e técnicas fundamentais que orientaram o desenvolvimento e aprimoramento do sistema proposto.

3.1 Base de dados

As bases de dados usadas neste trabalho foram coletadas do site <sofascore.com> por meio de *web scraping*. Foram utilizados as temporadas 2021-2022, 2022-2023 e 2023-2024 do campeonato espanhol (LaLiga). Cada temporada foi considerada e treinada de forma independente.

Contudo, um passo natural é utilizar mais dados de treinamento, e para isso, criou-se a base de dados LaLiga *Pack Dataset*, o qual abrange as partidas realizadas pelas equipes durante três temporadas do campeonato espanhol (LaLiga). Trata-se de uma junção das três tabelas mencionadas anteriormente, sendo utilizada com os dados das temporadas 2021-2022 e 2022-2023 para treino, enquanto a temporada 2023-2024 será destinada aos testes. A tabela foi elaborada de forma autoral baseado nas bases já coletadas. Os dados presentes nela incluem os mesmos dados já descritos.

A Tabela 3.1 mostra os campos coletados e os seus respectivos tipos antes das modificações que serão realizadas no pré-processamento.

3.2 Pré-Processamento

Essa etapa tem como objetivo padronizar os dados, tornando-os uniformes para facilitar a aprendizagem efetiva da inteligência artificial. Além disso, visa excluir dados que não contribuirão para o processo e corrigir eventuais erros presentes nas tabelas.

3.2.1 *Datasets* LaLiga

Primeiramente, foi adicionada uma coluna chamada *Match Order*, que atribui um número para indicar a ordem em que as partidas ocorreram. Em seguida, os nomes das equipes foram substituídos por números de 1 a 20, exceto no LaLiga *Pack Dataset*, onde 3 times são rebaixados

Tabela 3.1 – Nomes e tipos dos campos usados na classificação das bases de dados da LaLiga.

Campo	Tipo do Dado	Intervalo dos Dados
Home Team Name	Alfa-numérico	-
Away Team Name	Alfa-numérico	-
Goals Home	Numérico	0-99
Goals Away	Numérico	0-99
Big Chances Home	Numérico	0-99
Big Chances Away	Numérico	0-99
Shots Home	Numérico	0-99
Shots Away	Numérico	0-99
GK Defenses Home	Numérico	0-99
GK Defenses Away	Numérico	0-99
Corners Home	Numérico	0-99
Corners Away	Numérico	0-99
Fouls Home	Numérico	0-99
Fouls Away	Numérico	0-99
Passes Home	Numérico	0-999
Passes Away	Numérico	0-999
Tackles Home	Numérico	0-99
Tackles Away	Numérico	0-99
FreeKicks Home	Numérico	0-99
FreeKicks Away	Numérico	0-99
Yellow Cards Home	Numérico	0-99
Yellow Cards Away	Numérico	0-99
Red Cards Home	Numérico	0-99
Red Cards Away	Numérico	0-99

Fonte: Criado pelo autor.

e 3 promovidos a cada temporada. Dessa forma, para a seção utilizada no treino, os times foram renomeados com números de 1 a 23.

Também foram adicionadas colunas de média móvel, nas quais serão calculadas as médias das estatísticas das últimas 5 partidas. Foram criadas colunas separadas para o time que joga em casa e para o time que joga fora, com um n variando de 2 até 5. Os nomes dessas colunas podem ser observados na Tabela 3.2.

Após isso, foram incluídas colunas com as estatísticas das últimas 5 partidas disputadas pelo time que está jogando em casa e pelo time que está jogando fora, desta vez com n variando de 1 até 5. Os nomes dessas colunas podem ser observados na Tabela 3.3.

Ademais, foram adicionadas quatro colunas extras que indicam se o time venceu a enésima partida anterior e se a partida anterior foi jogada em casa ou fora que podem ser observados na Tabela 3.4.

Além disso, introduziu-se uma nova coluna chamada “resultado” nesta mesma tabela. Essa adição permitirá identificar rapidamente o vencedor da partida. Se o time da casa marcar mais gols, a coluna receberá o valor um; se o time visitante marcar mais gols, a coluna receberá

Tabela 3.2 – Nomes e tipos dos campos adicionados como média nas bases de dados da LaLiga.

Campo	Tipo do Dado	Intervalo dos Dados
Goals Last n_Home	Numérico	0-99
Goals Last n_Away	Numérico	0-99
Big Chances Last n_Home	Numérico	0-99
Big Chances Last n_Away	Numérico	0-99
Shots Last n_Home	Numérico	0-99
Shots Last n_Away	Numérico	0-99
GK Defenses Last n_Home	Numérico	0-99
GK Defenses Last n_Away	Numérico	0-99
Corners Last n_Home	Numérico	0-99
Corners Last n_Away	Numérico	0-99
Fouls Last n_Home	Numérico	0-99
Fouls Last n_Away	Numérico	0-99
Passes Last n_Home	Numérico	0-999
Passes Last n_Away	Numérico	0-999
Tackles Last n_Home	Numérico	0-99
Tackles Last n_Away	Numérico	0-99
FreeKicks Last n_Home	Numérico	0-99
FreeKicks Last n_Away	Numérico	0-99
Yellow Cards Last n_Home	Numérico	0-99
Yellow Cards Last n_Away	Numérico	0-99
Red Cards Last n_Home	Numérico	0-99
Red Cards Last n_Away	Numérico	0-99

Fonte: Criado pelo autor.

o valor dois. No caso de ambas as equipes terem marcado o mesmo número de gols, a coluna receberá o valor zero. Esta recém-introduzida coluna será utilizada exclusivamente como um identificador para a classe do problema e não será empregada no conjunto de treinamento.

Ao final, o conjunto de dados resultante é o apresentado na Tabela 3.5. Além disso, o conjunto final de dados conta com as informações de aproximadamente 380 jogos por temporada.

3.2.2 Ajustes finais

Como as partidas já estão ordenadas pela coluna *Match Order*, normaliza-se os dados de todas as colunas e remove-se as linhas com valores *NaN* para que elas não interfiram no processo de treino e teste dos modelos.

Com os preparativos concluídos, os algoritmos mencionados anteriormente são utilizados para realizar a previsão dos resultados futuros das partidas em cada um dos conjuntos de dados, considerando suas versões já atualizadas com as exclusões e adições de colunas.

Tabela 3.3 – Nomes e tipos dos campos adicionados como números de partidas anteriores nas bases de dados da LaLiga.

Campo	Tipo do Dado	Intervalo dos Dados
HOME_n_Goals	Numérico	0-99
AWAY_n_Goals	Numérico	0-99
HOME_n_Big Chances	Numérico	0-99
AWAY_n_Big Chances	Numérico	0-99
HOME_n_Shots	Numérico	0-99
AWAY_n_Shots	Numérico	0-99
HOME_n_GK Defenses	Numérico	0-99
AWAY_n_GK Defenses	Numérico	0-99
HOME_n_Corners	Numérico	0-99
AWAY_n _Corners	Numérico	0-99
HOME_n_Fouls	Numérico	0-99
AWAY_n_Fouls	Numérico	0-99
HOME_n_Passes	Numérico	0-999
AWAY_n _Passes	Numérico	0-999
HOME_n_Tackles	Numérico	0-99
AWAY_n_Tackles	Numérico	0-99
HOME_n_FreeKicks	Numérico	0-99
AWAY_n_Free Kicks	Numérico	0-99
HOME_n_Yellow Cards	Numérico	0-99
AWAY_n_Yellow Cards	Numérico	0-99
HOME_n_Red Cards	Numérico	0-99
AWAY_n_Red Cards	Numérico	0-99

Fonte: Criado pelo autor.

Tabela 3.4 – Nomes e tipos dos campos adicionados como resultados e locais anteriores nas bases de dados da LaLiga.

Campo	Tipo do Dado	Intervalo dos Dados
HOME_n_Resultado	Numérico	0-1
AWAY_n_Resultado	Numérico	0-1
HOME_n_Location	Numérico	0-1
AWAY_n_Location	Numérico	0-1

Fonte: Criado pelo autor.

3.3 Treinamento

Os algoritmos que vão ser utilizados para o treinamento foram citados anteriormente no Capítulo 2, sendo eles *Random Forest*, XGBoost e o *Naive Bayes*. Eles serão utilizados em cada uma das duas bases de dados separadamente.

O número de árvores utilizadas para o treinamento com os algoritmos *Random Forest* e *XGBoost* foi de 100 árvores, já a profundidade para o *Random Forest* foi de 100, enquanto que para o *XGBoost* foi de 6. O *Naive Bayes* utilizou uma distribuição Gaussiana.

A métrica adotada para avaliar a eficácia do modelo será a acurácia de classificação,

Tabela 3.5 – Campos usados na classificação das bases de dados da LaLiga.

Campo	Tipo do Dado	Intervalo dos Dados
Goals Last n_Home	Numérico	0-99
Goals Last n_Away	Numérico	0-99
Big Chances Last n_Home	Numérico	0-99
Big Chances Last n_Away	Numérico	0-99
Shots Last n_Home	Numérico	0-99
Shots Last n_Away	Numérico	0-99
GK Defenses Last n_Home	Numérico	0-99
GK Defenses Last n_Away	Numérico	0-99
Corners Last n_Home	Numérico	0-99
Corners Last n_Away	Numérico	0-99
Fouls Last n_Home	Numérico	0-99
Fouls Last n_Away	Numérico	0-99
Passes Last n_Home	Numérico	0-999
Passes Last n_Away	Numérico	0-999
Tackles Last n_Home	Numérico	0-99
Tackles Last n_Away	Numérico	0-99
FreeKicks Last n_Home	Numérico	0-99
FreeKicks Last n_Away	Numérico	0-99
Yellow Cards Last n_Home	Numérico	0-99
Yellow Cards Last n_Away	Numérico	0-99
Red Cards Last n_Home	Numérico	0-99
Red Cards Last n_Away	Numérico	0-99
HOME_n_Goals	Numérico	0-99
AWAY_n_Goals	Numérico	0-99
HOME_n_Big Chances	Numérico	0-99
AWAY_n_Big Chances	Numérico	0-99
HOME_n_Shots	Numérico	0-99
AWAY_n_Shots	Numérico	0-99
HOME_n_GK Defenses	Numérico	0-99
AWAY_n_GK Defenses	Numérico	0-99
HOME_n_Corners	Numérico	0-99
AWAY_n_Corners	Numérico	0-99
HOME_n_Fouls	Numérico	0-99
AWAY_n_Fouls	Numérico	0-99
HOME_n_Passes	Numérico	0-999
AWAY_n_Passes	Numérico	0-999
HOME_n_Tackles	Numérico	0-99
AWAY_n_Tackles	Numérico	0-99
HOME_n_FreeKicks	Numérico	0-99
AWAY_n_FreeKicks	Numérico	0-99
HOME_n_Yellow Cards	Numérico	0-99
AWAY_n_Yellow Cards	Numérico	0-99
HOME_n_Red Cards	Numérico	0-99
AWAY_n_Red Cards	Numérico	0-99
HOME_n_Resultado	Numérico	0-1
AWAY_n_Resultado	Numérico	0-1
HOME_n_Location	Numérico	0-1
AWAY_n_Location	Numérico	0-1
Resultado	Numérico	0-2

Fonte: Criado pelo autor.

definida como a proporção entre o número de previsões corretas e o total de previsões. Essa métrica servirá como critério para determinar se o desempenho do modelo é satisfatório ou não. Ela pode ser definida como:

$$ACURACIA = \frac{PREDICOES \quad CORRETAS}{TOTAL \quad DE \quad PREDICOES.} \tag{3.1}$$

Serão utilizadas diferentes abordagens de treinamento e teste, nas quais serão consideradas os dados das últimas 5, 4, 3, 2 e 1 partidas disputadas por cada uma das equipes.

No processo de treinamento, será utilizada a validação cruzada, onde os conjuntos de dados serão divididos em 10 grupos, usando um grupo para teste e os outros nove para treinamento. O modelo será treinado com os dados de treinamento e avaliado com os dados de teste. A métrica utilizada, neste caso a acurácia, será registrada, e em seguida o modelo será descartado. Essa abordagem oferece uma avaliação mais precisa do desempenho dos modelos frente a diferentes particionamentos dos dados, garantindo uma análise mais confiável e abrangente.

4 Resultados

Neste capítulo, são apresentados os resultados obtidos por meio da execução de cada um dos algoritmos em ambos os conjuntos de dados. Além disso, as linguagens de programação e os pacotes utilizados para conduzir os testes são apresentadas, fazendo comparações com os resultados obtidos em outros estudos mencionados nas referências teóricas.

O ambiente de teste adotado foi o Google Colab, e a linguagem selecionada para a realização dos testes foi o Python. Os pacotes empregados incluíram o Pandas, Scikit-Learn e XGBoost.

Os resultados da acurácia (média e desvio padrão) dos algoritmos aplicados aos problemas se encontram na Tabela 4.1.

Tabela 4.1 – Acurácia em porcentagem (%) dos modelos treinados para as bases de dados com a utilização de 1, 2, 3, 4 e 5 partidas anteriores para cada equipe. Em negrito estão os melhores resultados.

Algoritmos	LaLiga 21-22	LaLiga 22-23	LaLiga 23-24	LaLiga DF
Base de dados utilizando 1 jogo				
Naive Bayes	0.42 ± 0.07	0.46 ± 0.06	0.46 ± 0.07	$\mathbf{0.51 \pm 0.00}$
Random Forest	$\mathbf{0.47 \pm 0.06}$	$\mathbf{0.50 \pm 0.05}$	0.46 ± 0.05	0.50 ± 0.02
XGBoost	0.46 ± 0.08	0.49 ± 0.06	$\mathbf{0.46 \pm 0.04}$	0.46 ± 0.00
Base de dados utilizando 2 jogos				
Naive Bayes	0.41 ± 0.06	0.46 ± 0.03	$\mathbf{0.47 \pm 0.07}$	$\mathbf{0.49 \pm 0.00}$
Random Forest	$\mathbf{0.44 \pm 0.05}$	$\mathbf{0.50 \pm 0.04}$	0.45 ± 0.03	0.48 ± 0.01
XGBoost	0.42 ± 0.10	0.42 ± 0.06	0.42 ± 0.06	0.47 ± 0.00
Base de dados utilizando 3 jogos				
Naive Bayes	0.45 ± 0.12	0.45 ± 0.04	0.45 ± 0.06	$\mathbf{0.53 \pm 0.00}$
Random Forest	$\mathbf{0.47 \pm 0.06}$	$\mathbf{0.49 \pm 0.07}$	$\mathbf{0.47 \pm 0.08}$	0.51 ± 0.01
XGBoost	0.41 ± 0.06	0.45 ± 0.08	0.41 ± 0.07	0.47 ± 0.00
Base de dados utilizando 4 jogos				
Naive Bayes	$\mathbf{0.47 \pm 0.09}$	0.42 ± 0.06	0.45 ± 0.06	$\mathbf{0.52 \pm 0.00}$
Random Forest	0.45 ± 0.07	$\mathbf{0.49 \pm 0.04}$	$\mathbf{0.47 \pm 0.04}$	0.52 ± 0.01
XGBoost	0.43 ± 0.08	0.47 ± 0.10	0.45 ± 0.08	0.48 ± 0.00
Base de dados utilizando 5 jogos				
Naive Bayes	0.48 ± 0.12	0.47 ± 0.08	0.45 ± 0.05	$\mathbf{0.53 \pm 0.00}$
Random Forest	0.48 ± 0.07	$\mathbf{0.51 \pm 0.11}$	$\mathbf{0.48 \pm 0.06}$	0.52 ± 0.01
XGBoost	$\mathbf{0.50 \pm 0.09}$	0.50 ± 0.07	0.45 ± 0.07	0.52 ± 0.00

Fonte: Criada pelo autor.

Conforme pode ser visto na Tabela 4.1, os melhores resultados foram alcançados utilizando a base de dados composta pela junção das temporadas LaLiga 21-22 e LaLiga 22-23 para treino, enquanto a LaLiga 23-24 foi usada exclusivamente para teste. Independentemente do número de partidas anteriores consideradas, os três algoritmos obtiveram seus melhores desem-

penhos com a junção das bases. O algoritmo que obteve o melhor resultado foi o *Naive Bayes*, que chegou a uma acurácia de 53%, tanto utilizando 5 jogos anteriores no modelo quanto usando apenas 3 jogos anteriores. No trabalho proposto por Bernardes (2023), os métodos *Random Forest* e *Naive Bayes* alcançaram ambos uma acurácia média de 56,9%, utilizando uma divisão de 80% para treinos e 20% para testes.

Olhando apenas para a base de dados LaLiga 21-22, presente na Tabela 4.1, os melhores resultados foram obtidos ao se utilizar 5 partidas anteriores nos treinamentos, e o melhor algoritmo foi o *XGBoost*, com uma acurácia média de 50%.

Já analisando a base de dados LaLiga 22-23, o melhor resultado também foi obtido ao se utilizar 5 partidas anteriores nos treinamentos, e o melhor algoritmo foi o *Random Forest*, com uma taxa de acurácia de 51%. Este resultado está presente na Tabela 4.1.

Observando os resultados obtidos com a base de dados LaLiga 23-24, os melhores resultados também foram alcançados ao se analisar um número maior de partidas anteriores. No entanto, esse foi o pior resultado entre todas as bases de dados utilizadas, com a maior acurácia chegando a apenas 45% ao utilizar o algoritmo *Random Forest*. Este resultado também está presente na Tabela 4.1.

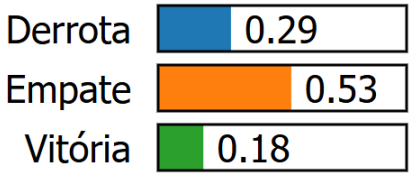
4.1 Interpretabilidade do melhor modelo

Apesar dos resultados obtidos (53% de acerto ao prever os jogos), os resultados ainda não estão claros do porquê o modelo possivelmente errou. Uma das formas de analisar esse erro, é aplicando técnicas de interpretabilidade para apropriadamente entender os resultados e, para isso, será avaliada a interpretabilidade dos três algoritmos propostos, no melhor cenário dos resultados, ou seja, com uma janela de cinco jogos anteriores e usando as temporadas 21-22 e 22-23 para treinamento e 23-24 para teste.

4.1.1 Interpretabilidade *Random Forest* com 5 jogos anteriores

Na Figura 4.1 é possível visualizar a tendência das predições para cada um dos resultados possíveis para o time da casa. Neste exemplo, a maior probabilidade é de que o jogo termine em empate. Nos jogos utilizados nos treinamento, a proporção de resultados foi a seguinte: 27% de derrotas, 29% de empates, e 47% de vitórias, ou seja, o algoritmo tende a resultados diferentes da distribuição observada nos dados dos quais foi treinado.

Figura 4.1 – Probabilidade de cada resultado para o time da casa utilizando o modelo *Random Forest* com 5 jogos anteriores.

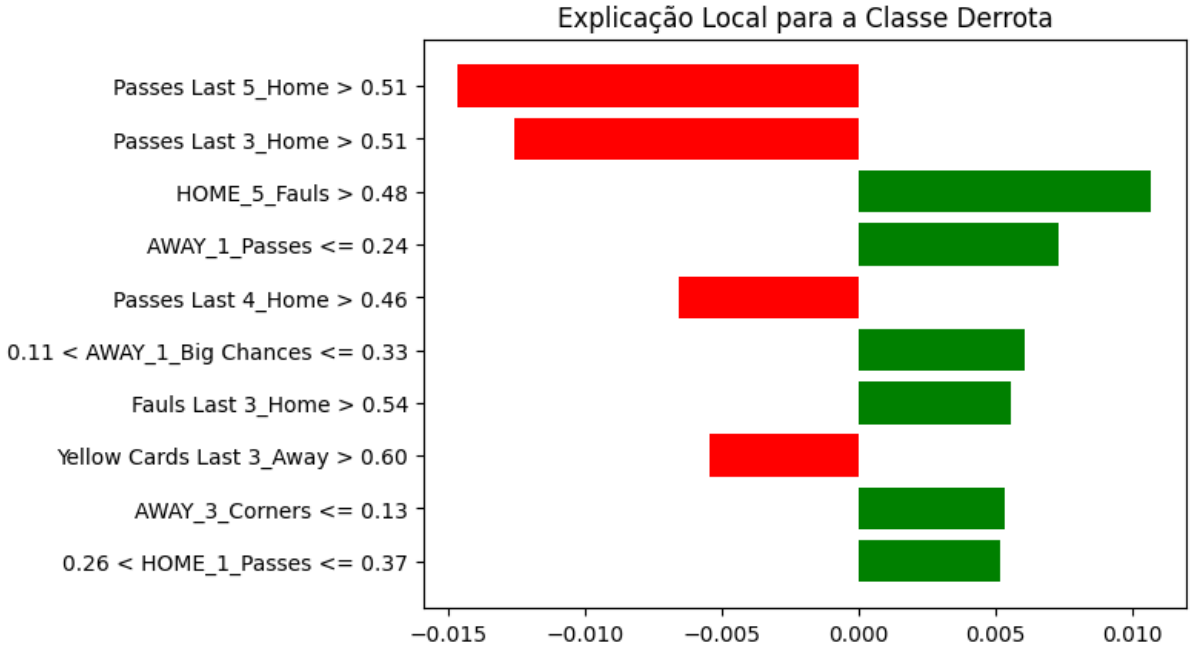


Fonte: Criada pelo autor.

4.1.1.1 Análise dos dados para probabilidades de Derrota

A explicação dos principais componentes utilizados na classificação de uma derrota pode ser vista na Figura 4.2. Pode-se observar que passes e disciplina em campo (faltas e cartões) são fatores determinantes na previsão de derrota, com maior impacto nos jogos recentes e na gestão de oportunidades do adversário. Isso mostra que o time precisa focar na redução de faltas e cartões amarelos, especialmente em jogos fora de casa, enquanto um controle mais rigoroso dos passes e das chances do adversário pode ajudar a diminuir a probabilidade de derrota, o que já é esperado.

Figura 4.2 – Explicação Local de forma mais detalhada para a predição de derrota no modelo *Random Forest*.



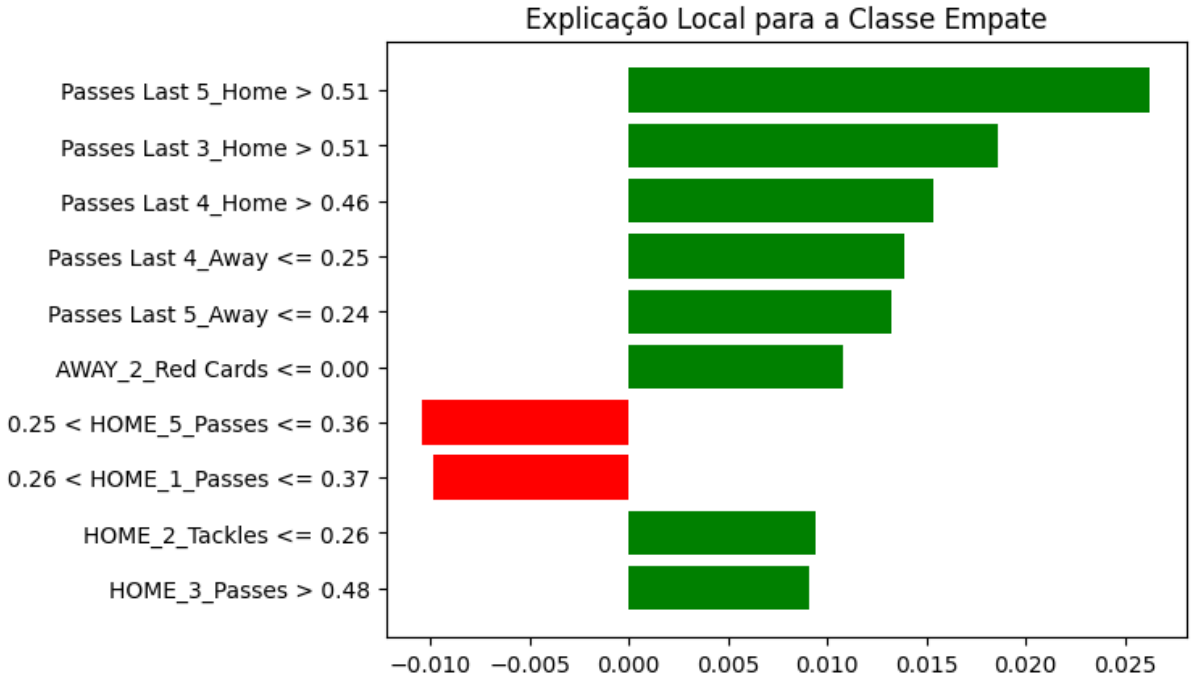
Fonte: Criada pelo autor.

4.1.1.2 Análise dos dados para probabilidades de Empate

A Figura 4.3 apresenta a explicação local gerada pelo LIME para a previsão de um empate no modelo *Random Forest*. O gráfico sugere que o controle do jogo, especialmente em partidas

em casa, por meio de passes, é um dos fatores que mais contribui para um empate. À medida que o time mantém maior posse de bola e consegue distribuir os passes de forma eficaz, aumenta a probabilidade de que o jogo permaneça equilibrado. Além disso, a ausência de cartões vermelhos para o adversário indica que o jogo tende a ficar mais balanceado, o que favorece o empate. Por outro lado, quando o time tem dificuldade em realizar passes suficientes, a chance de empate diminui, provavelmente porque o jogo passa a pender para um dos lados.

Figura 4.3 – Explicação Local de forma mais detalhada para a predição de empate no modelo *Random Forest*.

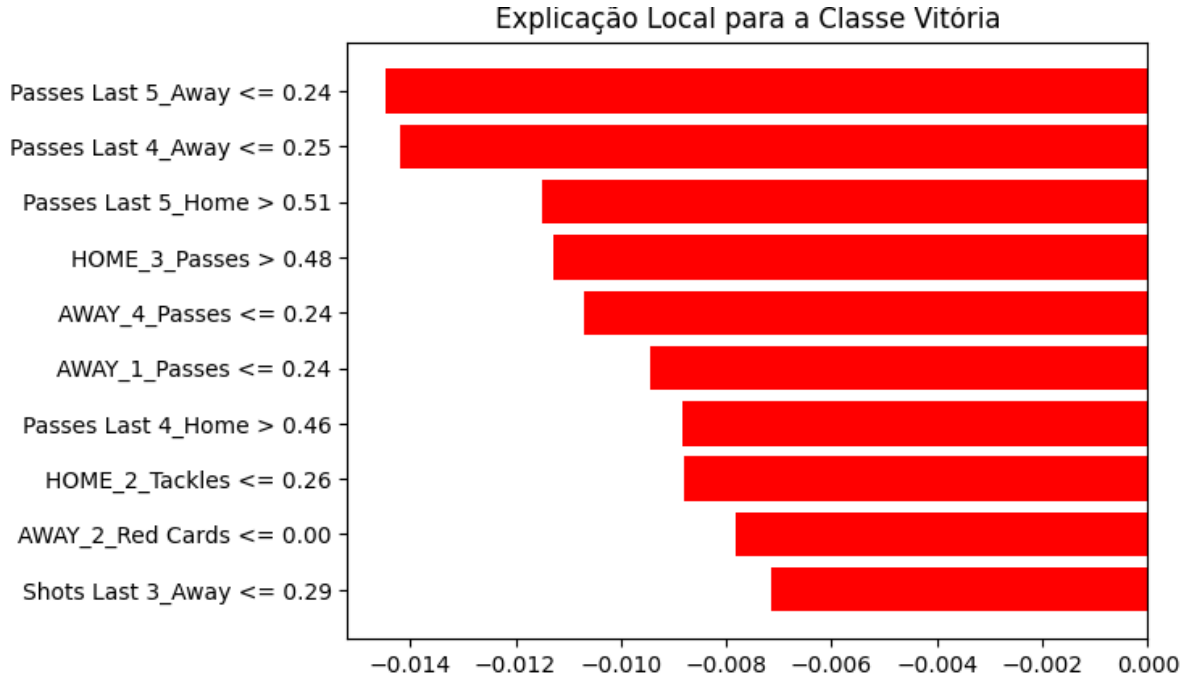


Fonte: Criada pelo autor.

4.1.1.3 Análise dos dados para probabilidades de Vitória

Por fim, a Figura 4.4 apresenta a explicação local gerada pelo LIME para a previsão de vitória no modelo *Random Forest*. Ao analisar o gráfico, pode-se observar que, quando o time adversário realiza poucos passes fora de casa, isso aumenta significativamente as chances de vitória do time analisado, destacando o controle sobre o adversário como um fator crucial em partidas fora de casa. A realização de um elevado número de passes em casa, especialmente nos últimos 5 jogos, também aumenta consideravelmente a probabilidade de vitória, reforçando a posse de bola como uma estratégia essencial para alcançar um resultado positivo. O menor número de desarmes e a ausência de cartões vermelhos para o adversário indicam uma vitória mais controlada, sem que o time precise contar com a vantagem numérica.

Figura 4.4 – Explicação Local de forma mais detalhada para a predição de vitória no modelo *Random Forest*.

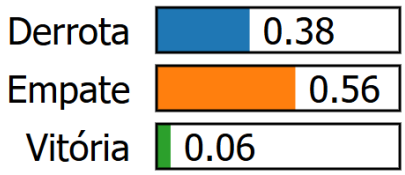


Fonte: Criada pelo autor.

4.1.2 Interpretabilidade XGBoost com 5 jogos anteriores

Na Figura 4.5 é possível visualizar as predições para cada um dos resultados possíveis para o time da casa. Neste exemplo utilizando o *XGBoost*, a maior probabilidade é de empate. Da mesma forma que o *Random Forest*, o *XGBoost* tem uma probabilidade maior ao impacte, seguido pela derrota e vitória respectivamente, o que também foge da distribuição dos dados de teste.

Figura 4.5 – Probabilidade de cada resultado para o time da casa utilizando o modelo XGBoost com 5 jogos anteriores.



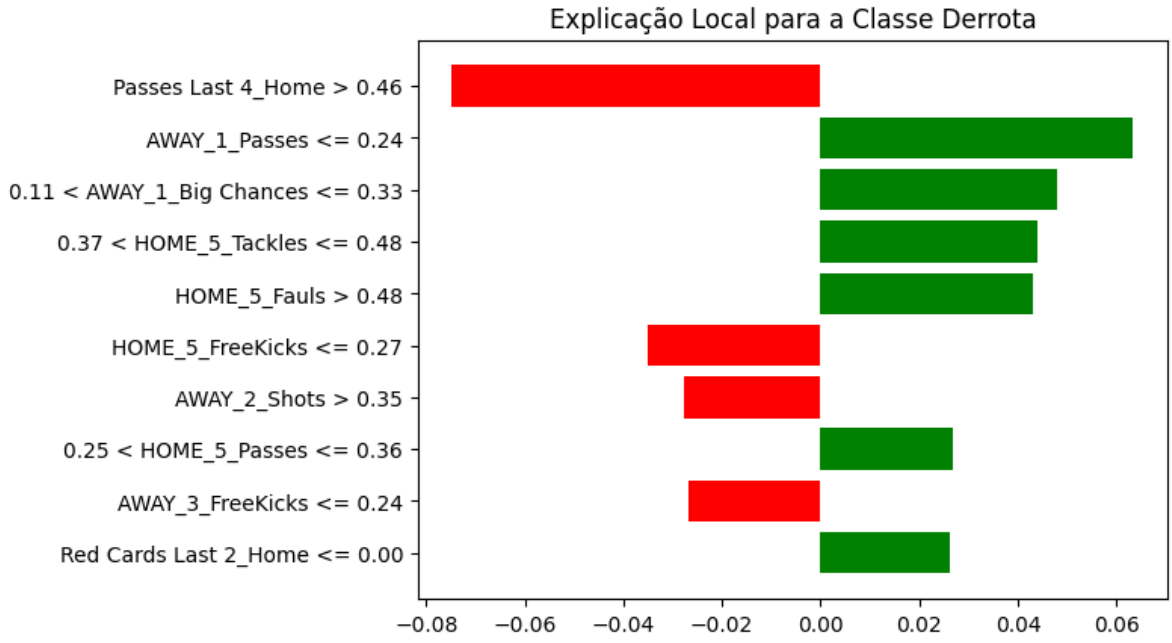
Fonte: Criada pelo autor.

4.1.2.1 Análise dos dados para probabilidades de Derrota

A Figura 4.6 apresenta a explicação local gerada pelo LIME para a previsão de derrota no modelo *XGBoost*. O gráfico analisado mostra que um número elevado de passes em jogos em casa aumenta consideravelmente a probabilidade de derrota. Isso pode indicar que um estilo de jogo mais passivo, focado excessivamente na posse de bola e com pouca agressividade ofensiva,

pode ser prejudicial. Por outro lado, limitar as grandes chances criadas pelo adversário fora de casa reduz a probabilidade de derrota, o que sugere que uma defesa sólida é fundamental para evitar resultados negativos. Além disso, a ausência de cartões vermelhos nos últimos dois jogos em casa favorece o time, indicando que a disciplina em campo contribui para evitar a derrota.

Figura 4.6 – Explicação Local de forma mais detalhada para a predição de derrota no modelo *XGBoost*.



Fonte: Criada pelo autor.

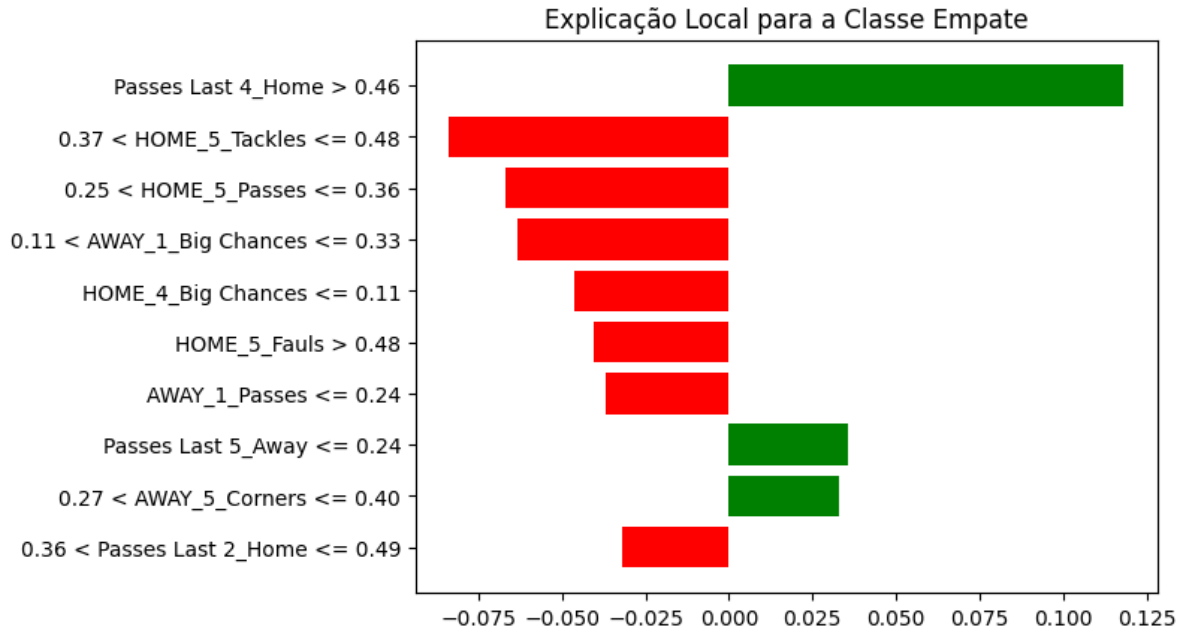
4.1.2.2 Análise dos dados para probabilidades de Empate

A Figura 4.7 mostra a explicação local gerada pelo LIME para a previsão de empate no modelo *XGBoost*. O gráfico revela que a posse de bola, especialmente por meio de um elevado número de passes em casa, aumenta significativamente as chances de empate, sugerindo que o controle do jogo favorece um resultado equilibrado. Por outro lado, fatores como desarmes frequentes, muitas faltas e grandes chances criadas pelo adversário reduzem a probabilidade de empate, indicando que um jogo mais defensivo ou desorganizado tende a levar a um resultado decisivo, seja vitória ou derrota. Assim, manter a posse e limitar as oportunidades do adversário são essenciais para garantir o equilíbrio no placar.

4.1.2.3 Análise dos dados para probabilidades de Vitória

A Figura 4.8 apresenta a explicação local gerada pelo LIME para a previsão de vitória no modelo *XGBoost*. O gráfico destaca que a realização de menos passes e menos desarmes pelo time da casa favorece a vitória, sugerindo que um jogo mais direto e eficiente é crucial para alcançar um resultado positivo.Além disso, o controle das grandes chances criadas pelo adversário e a eficiência das defesas do goleiro adversário também contribuem para aumentar a

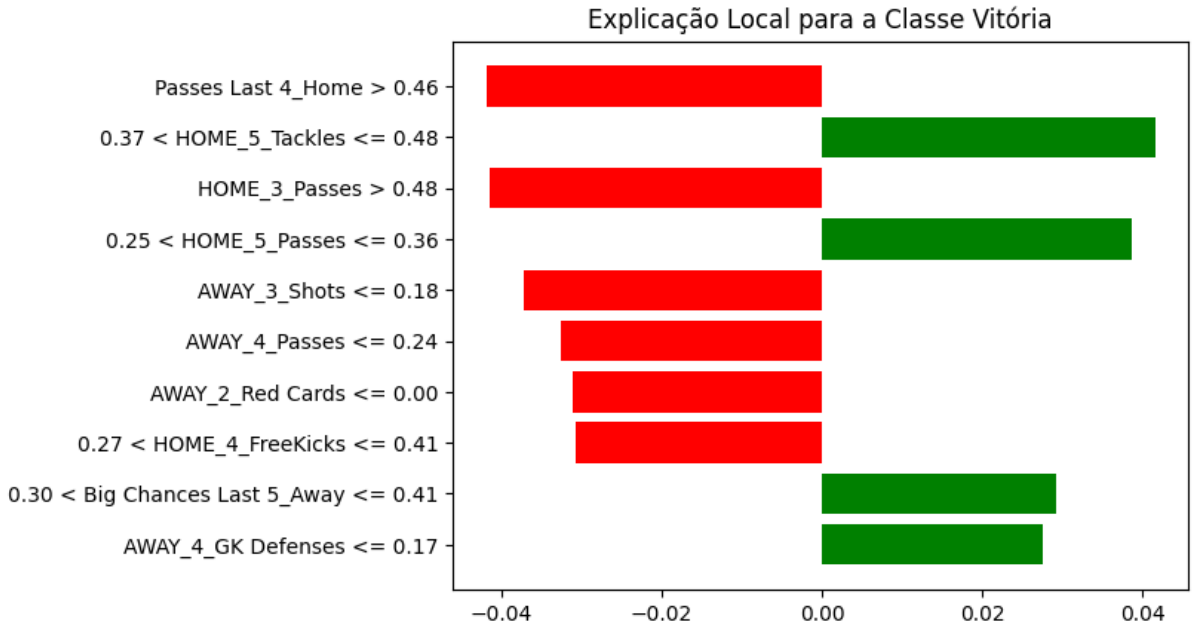
Figura 4.7 – Explicação Local de forma mais detalhada para a predição de empate no modelo *XGBoost*.



Fonte: Criada pelo autor.

probabilidade de vitória. Por outro lado, fatores como um elevado número de passes em casa, faltas cobradas e cartões vermelhos recebidos pelo adversário estão associados a uma diminuição nas chances de vitória, indicando que o jogo pode se tornar mais desafiador ou equilibrado.

Figura 4.8 – Explicação Local de forma mais detalhada para a predição de vitória no modelo *XGBoost*.

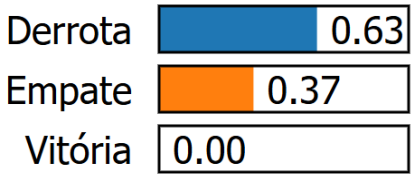


Fonte: Criada pelo autor.

4.1.3 Interpretabilidade *Naive Bayes* com 5 jogos anteriores

Na Figura 4.9 é possível visualizar as predições para cada um dos resultados possíveis para o time da casa. Neste exemplo, utilizando o modelo *Naive Bayes*, a maior probabilidade é de derrota. O *Naive Bayes* foi o único com um comportamento diferente, onde há mais probabilidade de derrota do que empate e vitória.

Figura 4.9 – Probabilidade de cada resultado para o time da casa utilizando o modelo Naive Bayes com 5 jogos anteriores.

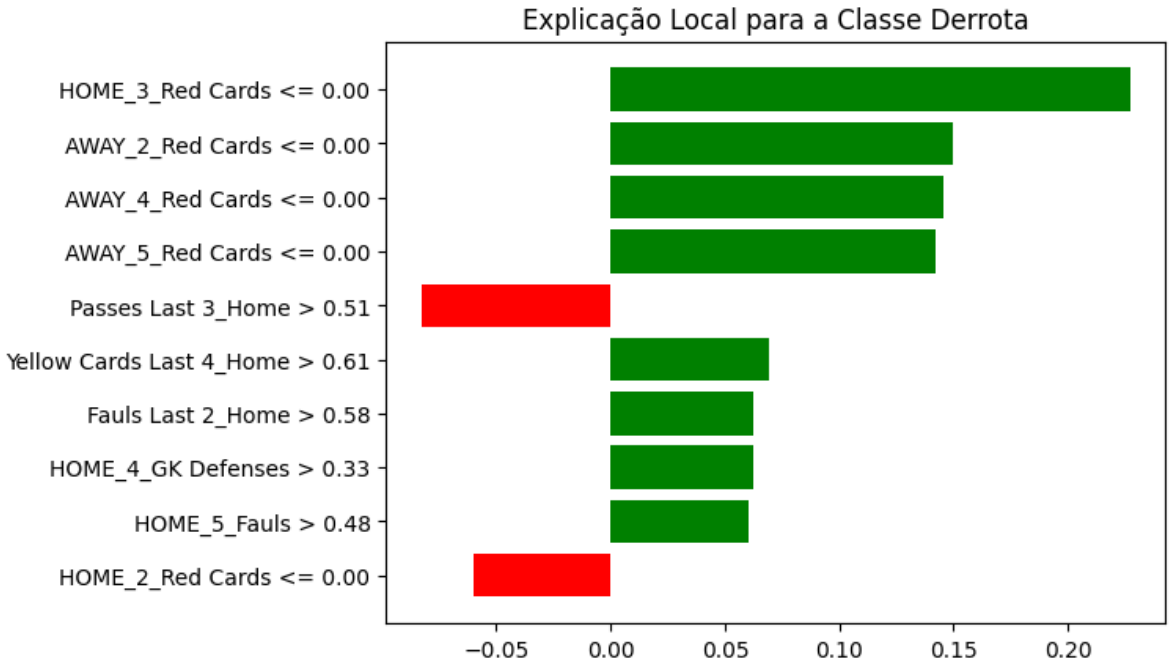


Fonte: Criada pelo autor.

4.1.3.1 Análise dos dados para probabilidades de Derrota

A Figura 4.10 apresenta a explicação local gerada pelo LIME para a previsão de derrota no modelo *Naive Bayes*.

Figura 4.10 – Explicação Local de forma mais detalhada para a predição de derrota no modelo *Naive Bayes*.



Fonte: Criada pelo autor.

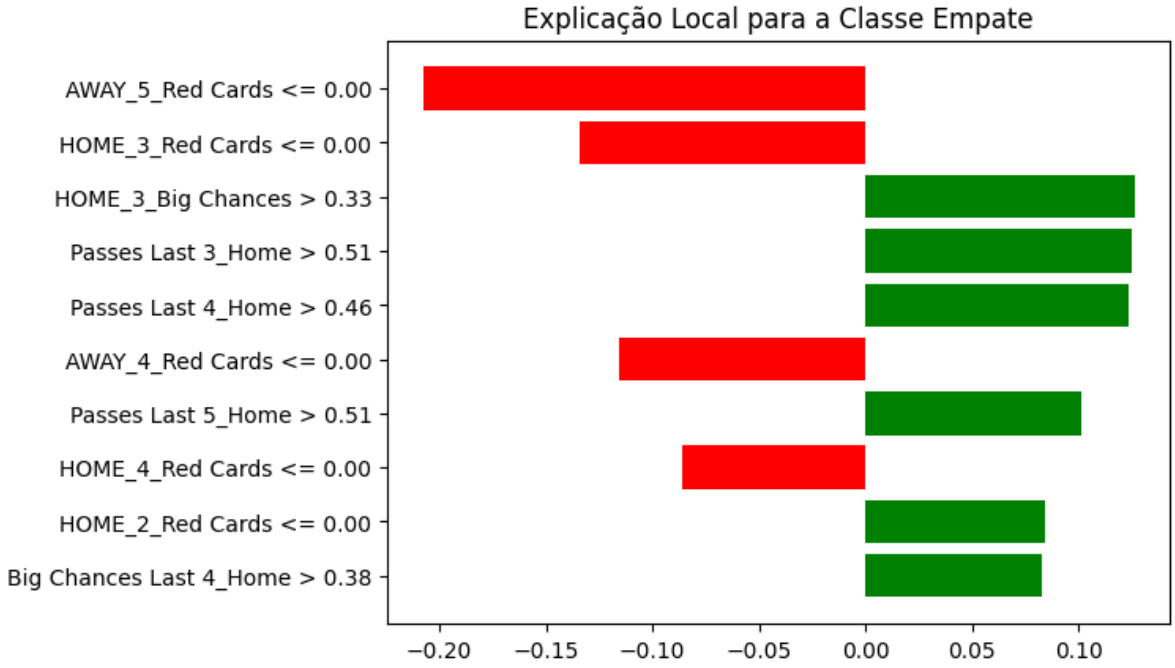
Pode-se observar que a ausência de cartões vermelhos, tanto para o time da casa quanto para o adversário, tem um impacto positivo na redução da probabilidade de derrota, sugerindo que a disciplina em campo é fundamental para evitar um resultado negativo. Além disso, variáveis

como defesas do goleiro e controle de faltas também ajudam a diminuir a chance de derrota. Por outro lado, um número elevado de passes nos últimos jogos em casa e cartões amarelos estão associados a um aumento na probabilidade de derrota, indicando que um jogo mais agressivo e excessivamente focado em posse de bola, sem efetividade, pode prejudicar o time.

4.1.3.2 Análise dos dados para probabilidades de Empate

A Figura 4.11 apresenta a explicação local gerada pelo LIME para a previsão de empate no modelo *Naive Bayes*. O gráfico indica que a ausência de cartões vermelhos, tanto para o time da casa quanto para o adversário, diminui significativamente a probabilidade de um empate, sugerindo que quando o jogo é disciplinado e sem expulsões, a tendência é que o resultado seja mais decisivo, seja vitória ou derrota. Por outro lado, fatores como um número elevado de grandes chances criadas pelo time da casa e um maior número de passes realizados, tanto nos últimos 3 como nos últimos 4 jogos em casa, aumentam a probabilidade de empate. Isso indica que um controle mais efetivo da bola e a criação de oportunidades pelo time da casa equilibram o jogo, favorecendo o empate. Assim, o gráfico destaca a importância da disciplina e do controle de jogo para a obtenção desse tipo de resultado.

Figura 4.11 – Explicação Local de forma mais detalhada para a predição de empate no modelo *Naive Bayes*.



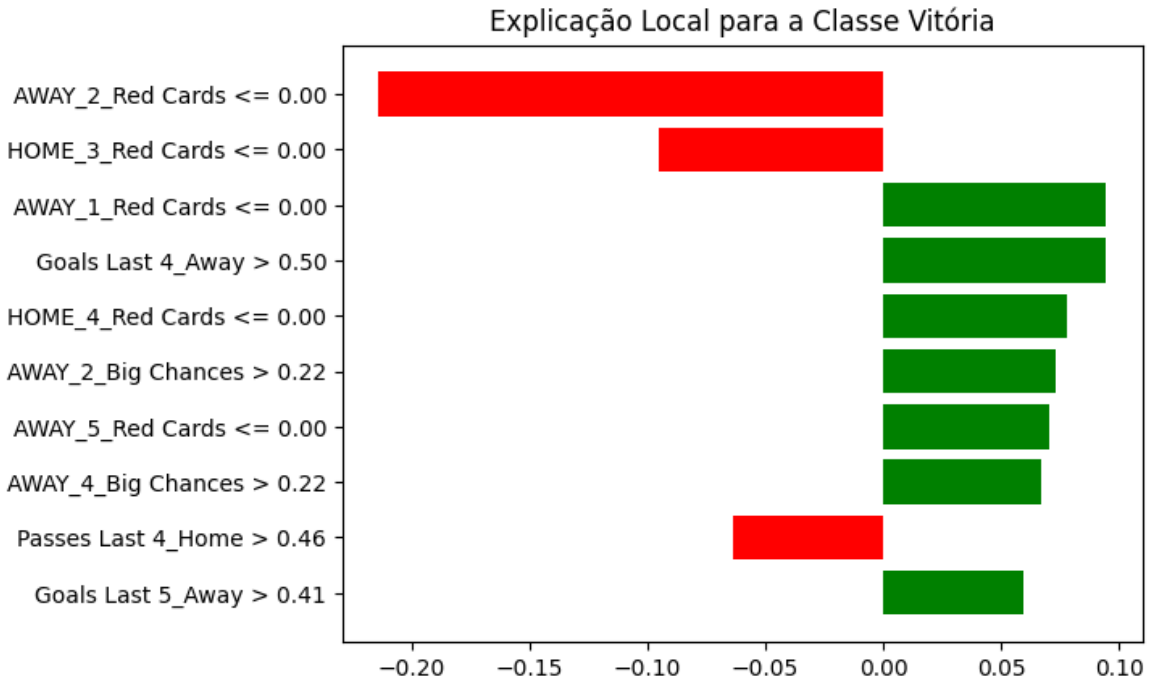
Fonte: Criada pelo autor.

4.1.3.3 Análise dos dados para probabilidades de Vitória

A Figura 4.12 apresenta a explicação local gerada pelo LIME para a previsão de vitória no modelo *Naive Bayes*. E por fim , o gráfico sugere que um jogo mais agressivo, com a criação de grandes chances e finalizações bem-sucedidas, é essencial para aumentar a probabilidade

de vitória, enquanto a ausência de cartões vermelhos, tanto para o time da casa quanto para o adversário, pode dificultar a obtenção de um resultado positivo.

Figura 4.12 – Explicação Local de forma mais detalhada para a predição de vitória no modelo *Naive Bayes*.



Fonte: Criada pelo autor.

4.2 Discussão dos resultados

Um dos fatores mais recorrentes na determinação do resultado é o controle do jogo por meio de passes. Quando o time da casa realiza um número elevado de passes, especialmente nos últimos jogos, isso tende a aumentar a probabilidade de empate. Esse controle indica uma partida equilibrada, onde a posse de bola é fundamental para manter o jogo em ritmo controlado. No entanto, a mesma estratégia pode ser prejudicial quando focada excessivamente na posse sem conversão ofensiva, o que acaba aumentando a probabilidade de derrota, como observado nos gráficos da classe “Derrota”. Por outro lado, a eficiência nos passes e o controle das grandes chances criadas pelo adversário são fundamentais para alcançar a vitória, demonstrando a importância de um equilíbrio entre posse e efetividade.

As grandes chances criadas, tanto pelo time da casa quanto pelo adversário, têm um papel crucial na determinação do resultado. Conforme esperado, limitar as oportunidades claras do adversário reduz a probabilidade de derrota e aumenta as chances de empate. Por outro lado, para garantir a vitória, é fundamental converter as chances criadas em gols.Isso demonstra que, além de controlar o jogo por meio de passes, é essencial ter uma ofensiva eficiente, capaz de aproveitar as oportunidades criadas para definir o resultado a favor do time.

Em suma, os resultados indicam que a posse de bola e o controle do ritmo de jogo, quando bem administrados, favorecem o equilíbrio, levando a empates. Contudo, esse controle pode se tornar prejudicial se não for acompanhado de uma estratégia ofensiva eficiente, aumentando a probabilidade de derrota. A disciplina em campo, refletida pela ausência de cartões vermelhos, é um fator decisivo que tanto preserva o equilíbrio quanto dificulta a vitória, dependendo do contexto. Por fim, a capacidade de criar e converter grandes chances, especialmente fora de casa, é o diferencial mais marcante para garantir a vitória. Portanto, a combinação de controle de jogo, disciplina e ofensividade é o caminho ideal para obter um resultado positivo.

5 Considerações Finais

Neste capítulo, são abordadas as conclusões derivadas da aplicação das técnicas e algoritmos anteriormente delineados, avaliando se os objetivos propostos foram atingidos. Além disso, apresentaremos um cronograma de atividades destinadas a trabalhos futuros que serão desenvolvidos. Este segmento proporcionará uma análise reflexiva sobre os resultados obtidos e estabelecerá uma perspectiva para direções potenciais de pesquisa e aprimoramento no futuro.

5.1 Conclusão

O futebol, um dos esportes mais populares do mundo, oferece uma rica base de dados que, quando analisada com técnicas de ciência de dados e aprendizado de máquina, pode proporcionar *insights* valiosos sobre o desempenho das equipes e os resultados das partidas. Este estudo explorou o potencial dessas técnicas para prever o desfecho de jogos da LaLiga, utilizando algoritmos como *Random Forest*, *XGBoost* e *Naive Bayes*.

Ao longo dos testes realizados com diferentes conjuntos de dados, observou-se que não houve uma diferença significativa nos resultados entre os algoritmos testados. No entanto, em casos pontuais, o algoritmo *Naive Bayes* e o *XGBoost* apresentaram resultados ligeiramente inferiores ao *Random Forest*. Quando se utilizou uma base de dados maior, composta pela junção das temporadas 2021-2022, 2022-2023 e 2023-2024 da LaLiga, o algoritmo *Naive Bayes* apresentou o melhor desempenho, alcançando uma taxa de acerto de 53%.

A mineração de dados demonstrou-se fundamental para o processamento e análise de grandes volumes de informações, permitindo a aplicação de métodos avançados de predição. Apesar de as taxas de acerto ainda não serem ideais, este estudo contribui para o avanço na utilização de aprendizado de máquina no futebol, com a possibilidade de refinar os modelos em futuros trabalhos e explorar ainda mais o potencial dessas técnicas em diferentes competições e contextos.

Com a adição das ferramentas de interpretabilidade, foi possível observar que o modelo considera uma grande posse de bola, quando não acompanhada de chutes a gol e grandes chances, como uma métrica negativa para a vitória da equipe. pelo algoritmo de LIME é possível interpretar essa posse de bola como sendo sem objetividade, o que impacta negativamente a previsão de vitória. Além disso, um baixo número de faltas e cartões para ambas as equipes contribui para uma maior probabilidade de predição de empate. Com essas informações, torna-se viável ajustar o estilo de jogo de um time, de tal forma a maximizar as melhores opções durante o treinamento.

5.2 Trabalhos Futuros

Como trabalhos futuros, pode-se conduzir o treinamento dos modelos utilizando outros algoritmos de aprendizado de máquina, como técnicas baseadas em redes neurais. Este planejamento inclui o aumento dos dados, juntamente com pré-processamento, e o uso de um *Ensemble*. Além disso, outras métricas como precisão, revocação e F1-*score* podem ser verificadas, alinhando-se com a meta de oferecer estimativas mais confiáveis do desempenho de cada um dos algoritmos.

Outra estratégia para superar as limitações observadas nos experimentos deste trabalho, sugere-se a adição de mais dados, como as notas dos jogadores no simulador de futebol *EA FC*. Outro ponto é adicionar a informação dos 11 principais escalados para jogar o jogo e considerar informações de lesão e dados dos jogadores.

Referências

AMAZON. **Validação cruzada**. 2023. Disponível em: <https://docs.aws.amazon.com/pt_br/machine-learning/latest/dg/cross-validation.html>.

APOSTOLOU, K.; TJORTJIS, C. Sports analytics algorithms for performance prediction. In: **2019 10th International Conference on Information, Intelligence, Systems and Applications (IISA)**. [s.n.], 2019. p. 1–4. Disponível em: <https://www.researchgate.net/publication/337521686_Sports_Analytics_algorithms_for_performance_prediction>.

AWAD, M.; KHANNA, R. Machine learning. In: ____. **Efficient Learning Machines**. Berkeley, CA: Apress, 2015. cap. 1. Disponível em: <https://doi.org/10.1007/978-1-4302-5990-9_1>.

Awari. **A importância da ciência de dados no futebol**. 2023. Disponível em: <https://awari.com.br/a-importancia-da-ciencia-de-dados-no-futebol-como-a-analise-estatistica-revoluciona-o-esporte/>.

Becker. **Algoritmo de Classificação Naive Bayes**. 2019. Disponível em: <https://www.organicadigital.com/blog/algoritmo-de-classificacao-naive-bayes/>.

BERNARDES, D. Predizendo os vencedores dos playoffs: Um estudo de caso com aprendizado de máquina em partidas de futebol americano. In: **Anais Estendidos do XXXVIII Simpósio Brasileiro de Bancos de Dados**. Porto Alegre, RS, Brasil: SBC, 2023. p. 22–28. ISSN 0000-0000. Disponível em: <https://sol.sbc.org.br/index.php/sbbd_estendido/article/view/25609>.

CARPITA. Predicting football match results in the english premier league. In: **2015 IEEE International Conference on Data Science and Advanced Analytics (DSAA)**. [s.n.], 2015. p. 1–10. Disponível em: <https://www.researchgate.net/publication/349367241_Predicting_the_Outcome_of_English_Premier_League_Matches_using_Machine_Learning>.

CYPRIANO, E. **O teorema de Bayes**. 2015. Disponível em: <https://edisciplinas.usp.br/pluginfile.php/3110753/mod_resource/content/1/aula2.pdf>.

Data Geeks. **Fundamentos do Web Scraping**. 2024. <https://www.datageeks.com.br/web-scraping/#:~:text=Fundamentos%20do%20Web%20Scraping,-O%20que%20%C3%A9&text=Web%20Scraping%20%C3%A9%20o%20processo,esses%20dados%20para%20gerar%20insights.> Accessed: 2024-09-16.

DEUS, G. A. **Utilização de Aprendizado de Máquina para Previsão de Resultados de Jogos de Futebol**. Trabalho de Conclusão de Curso — Universidade Tecnológica Federal do Paraná, 2019. Disponível em: <http://repositorio.utfpr.edu.br/jspui/handle/1/12482>.

Diniz. **Esquadrão imortal Leicester City**. 2018. Disponível em: <https://imortaisdofutebol.com/esquadrao-imortal-leicester-city-2015-2017/>.

DIO.ME. **Interpretabilidade em modelos de Machine Learning**. 2024. Acesso em: 18 set. 2024. Disponível em: <https://www.dio.me/articles/interpretabilidade-em-modelos-de-machine-learning>.

DNC, E. **A importância da interpretabilidade de modelos de aprendizado de máquina**. 2023. Acesso em: 18 set. 2024. Disponível em: <https://www.escoladnc.com.br/blog/a-importancia-da-interpretabilidade-de-modelos-de-aprendizado-de-maquina/>.

ESPORTE, M. do. **Gráfico 9. Esportes praticados em 2013**. 2013. Disponível em: <http://www.diesporte.gov.br/images/graficos/Grafico_9_Esportes_praticados_em_2013.jpg>.

FERNANDES, F. A. P. Previsão de resultados no futebol por meio de técnicas de aprendizado de máquina. 2019. Disponível em: <https://sig.cefetmg.br/sigaa/verArquivo?idArquivo=2256419&key=6a8ee9c8bd115b18683f4b40fb5ec585>.

FONTANA. **Introdução aos Algoritmos de Aprendizagem Supervisionada**. 2018. <https://fontana.paginas.ufsc.br/files/2018/03/apostila_ML_pt2.pdf>. Apostila, Universidade Federal de Santa Catarina.

FONTANA, F. L. **Apostila de Aprendizado de Máquina (Parte 2)**. 2018. <https://fontana.paginas.ufsc.br/files/2018/03/apostila_ML_pt2.pdf>.

Gazeta Esportiva. **Como a tecnologia no futebol auxilia atletas e clubes**. 2021. Disponível em: <https://www.gazetaesportiva.com/institucional/como-a-tecnologia-no-futebol-auxilia-atletas-e-clubes/>.

Gomes. **Classificação com Naive Bayes**. 2019. Disponível em: <https://www.datageeks.com.br/naive-bayes/>.

GOMES, P. C. T. **Conheça o algoritmo XGBoost**. jun. 2019. Disponível em: <https://www.datageeks.com.br/xgboost/#:~:text=O%20XGBoost%20%C3%A9%20um%20algoritmo,os%20outros%20algoritmos%20ou%20frameworks>.

HABEHH, H.; GOHEL, S. Machine learning in healthcare. **Current genomics**, v. 22, n. 4, p. 291–300, 2021.

IBM. **O que é boosting?** 2023. Disponível em: <https://www.ibm.com/br-pt/topics/boosting>.

IBM. **O que é Mineração de dados?** 2023. Disponível em: <https://www.ibm.com/br-pt/topics/data-mining>.

JUNIOR, I. **O que é Random Forest?** 2021. Disponível em: <https://icmcjunior.com.br/random-forest/#:~:text=O%20que%20%C3%A9%20Random%20Forest,para%20chegar%20no%20resultado%20final.>

KRAVCHYCHYN, C. *et al.* Estudos brasileiros sobre o esporte: ênfase no esporte-educação. **Movimento**, v. 18, n. 2, p. 339–350, 2012. Disponível em: <https://seer.ufrgs.br/index.php/Movimento/article/view/27920>.

Lima e Amorim. **Random Forest**. 2020. Disponível em: <https://lamfo-unb.github.io/2020/07/08/Random-Forest/#:~:text=Em%20suma%2C%20*Random%20Forest*,uma%20floresta%20com%20baixa%20correla%C3%A7%C3%A3o.>

MATIAS, P. J. G. Cristino Júlio Alves da S. AnÁlise de jogo nos jogos esportivos coletivos: A exemplo do voleibol. **Revistas UFG**, 2009. Disponível em: <https://revistas.ufg.br/fef/article/view/6726/6199>.

MATRENIN, P.; ANTONENKOV, D.; ARESTOVA, A. Energy efficiency improvement of industrial enterprise based on machine learning electricity tariff forecasting. In: **2021 XV International Scientific-Technical Conference on Actual Problems Of Electronic Instrument Engineering (APEIE)**. [s.n.], 2021. p. 185–189. Disponível em: <https://www.researchgate.net/publication/357359372_Energy_Efficiency_Improvement_of_Industrial_Enterprise_Based_on_Machine_Learning_Electricity_Tariff_Forecasting>.

MOREIRA, A. **Futebol movimenta o equivalente ao PIB da Finlândia, diz presidente da Fifa**. 2022. Acessado em Novembro de 2023. Disponível em: <https://valor.globo.com/mundo/noticia/2022/09/27/futebol-movimenta-o-equivalente-ao-pib-da-finlandia-diz-presidente-da-fifa.ghtml>.

NABINGER, A. M. **Utilização de Algoritmos do Tipo Machine Learning Supervisionado para a Caracterização dos Resultados da Copa do Mundo de Futebol de 2018**. Trabalho de Conclusão de Curso — Universidade Federal do Rio Grande do Sul, 2018. Disponível em: <https://lume.ufrgs.br/handle/10183/199214>.

NEVES, B. O. **Mineração de Dados Aplicada à Predição de Resultados de Jogos de Basquete**. Trabalho de Conclusão de Curso — Universidade Federal de Ouro Preto, 2022. Disponível em: <https://monografias.ufop.br/bitstream/35400000/4350/6/MONOGRAFIA_Minera%C3%A7%C3%A3oDadosAplicada.pdf>.

PESSANHA, C. **Random Forest: como funciona um dos algoritmos mais populares de ML**. nov. 2020. Disponível em: <https://medium.com/cinthiabpessanha/random-forest-como-funciona-um-dos-algoritmos-mais-populares-de-ml-cc1b8a58b3b4>.

ROSAEN, K. **K-fold cross-validation**. 2016. Disponível em: <http://karlrosaen.com/ml/learning-log/2016-06-20/>.

Sacramento. **NAIVE BAYES: COMO FUNCIONA ESSE ALGORITMO DE CLASSIFICAÇÃO**. 2023. Disponível em: <https://blog.somostera.com/data-science/naive-bayes#:~:text=O%20classificador%20Naive%20Bayes%20%C3%A9,features%20s%C3%A3o%20independentes%20entre%20si.>

SANTOS, E. **Como Cerveja e Fralda viraram ouro em uma grande rede de supermercados**. 2023. Disponível em: <https://abemd.org.br/noticia/como-cerveja-e-fralda-viraram-ouro-em-uma-grande-rede-de-supermercados/>.

SAS. **O que é mineração de dados?** 2023. Disponível em: <https://www.sas.com/pt_br/insights/analytics/mineracao-de-dados.html>.

SAVIETTO, J. V. **Validação cruzada**. 2021. Disponível em: <https://medium.com/@jvsavietto6/machine-learning-m%C3%A9tricas-valida%C3%A7%C3%A3o-cruzada-bias-e-vari%C3%A2ncia-380513d97c95>.

SCHMIDT, H. L. Uso de técnicas de aprendizado de máquina no auxílio em previsão de resultados de partidas de futebol. 2017. Disponível em: <http://hdl.handle.net/11624/2157>.

SILVA, J. **Uma breve introdução ao algoritmo de Machine Learning Gradient Boosting utilizando a biblioteca Scikit-Learn**. 2020. Disponível em: <https://medium.com/equals-lab/uma-breve-introdu%C3%A7%C3%A3o-ao-algoritmo-de-machine-learning-gradient-boosting-utilizando-a-biblioteca-311285783099#:

~:text=Gradient%20Boosting,-O%20algoritmo%20Gradient&text=O%20objetivo%20do%20algoritmo%20%C3%A9,chamado%20de%20taxa%20de%20aprendizagem.>

SOUZA, S. de. **Uma breve história do futebol**. 2013. Disponível em: <https://www.yumpu.com/pt/document/read/12816708/uma-breve-historia-do-futebol-prof-sandro-de-souza>.

TüFEKCI, P. Predicting football match results in the english premier league: A machine learning approach. In: **2016 24th Signal Processing and Communication Application Conference (SIU)**. [s.n.], 2016. p. 1–4. Disponível em: <https://www.researchgate.net/publication/352940839_A_Machine_Learning_Approach_to_Football_Match_Result_Prediction>.