

UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM COMPUTAÇÃO

BRUNO OLIVEIRA MARQUES

**Inteligência artificial explicável para análise
de partidas de futebol**

Dissertação apresentada como requisito parcial
para a obtenção do grau de Mestre em Ciência da
Computação

Orientador: Prof. Dr. Dante Augusto Couto
Barone

Porto Alegre
2019

CIP — CATALOGAÇÃO NA PUBLICAÇÃO

Marques, Bruno Oliveira

Inteligência artificial explicável para análise de partidas de futebol / Bruno Oliveira Marques. – Porto Alegre: PPGC da UFRGS, 2019.

45 f.: il.

Dissertação (mestrado) – Universidade Federal do Rio Grande do Sul. Programa de Pós-Graduação em Computação, Porto Alegre, BR–RS, 2019. Orientador: Dante Augusto Couto Barone.

1. Análise de dados. 2. Aprendizado de máquina. 3. Futebol. 4. Inteligência artificial interpretável. I. Barone, Dante Augusto Couto. II. Título.

UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL

Reitor: Prof. Rui Vicente Oppermann

Vice-Reitora: Prof^a. Jane Fraga Tutikian

Pró-Reitor de Pós-Graduação: Prof. Celso Giannetti Loureiro Chaves

Diretora do Instituto de Informática: Prof^a. Carla Maria Dal Sasso Freitas

Coordenador do PPGC: Prof. João Luiz Dihl Comba

Bibliotecária-chefe do Instituto de Informática: Beatriz Regina Bastos Haro

*“Some people believe football is a matter of life and death;
I am very disappointed with that attitude. I can assure you
it is much, much more important than that.”*

— BILL SHANKLY

AGRADECIMENTOS

Gostaria de agradecer ao meu orientador, professor Dante Barone, por todo o apoio, oportunidades e contatos que ele me apresentou ao longo da minha pesquisa.

Agradeço também ao professor Giovani Cunha, da ESEF, por me guiar quanto às tendências atuais de pesquisa no futebol; isso foi fundamental para o rumo do meu trabalho.

Também agradeço aos meus pais, Luisa e Ricardo, por todo o suporte e base que me deram ao longo da minha vida.

Por fim, agradeço muito à, em breve, Dra. Andrea Balle, por ter me incentivado a ingressar no mestrado, me apoiado ao longo de todo ele, e pela contribuição fundamental no desenvolvimento deste trabalho que sua experiência na academia permitiu dar. Te amo muito.

RESUMO

As recentes discussões sobre a ética na implementação e uso de algoritmos de aprendizagem de máquina para tomada de decisões que impactam pessoas de forma significativa, motivadas principalmente pela constatação de que, em diversos casos, eles reproduziram tendências discriminatórias presentes nos dados históricos que os alimentaram, geraram uma demanda por maior transparência nesses mecanismos.

Tais esforços se consolidaram em um tópico emergente chamado de Inteligência Artificial Explicável, também conhecido pela sigla em inglês XAI. Seu objetivo é prover, para pessoas com acesso a um agente de aprendizagem, meios para que elas possam compreender o processo por trás de suas tomadas de decisão, com influência mínima na acurácia ou desempenho computacional. Uma de suas possíveis aplicações está na análise do impacto de fatores individuais, ou combinações destes, no resultado de uma previsão, permitindo assim otimizar melhor os esforços para a realização do objetivo representado por ela.

O objetivo deste trabalho é utilizar tais métodos para investigar a influência exercida por fatores presentes em partidas de futebol em relação aos seus resultados, em um formato que possa ser manipulado e compreendido pelas partes interessadas, incluindo treinadores, preparadores e jogadores, de forma a auxiliar a melhoria de seus treinamentos e preparações táticas.

Palavras-chave: Análise de dados. aprendizado de máquina. futebol. inteligência artificial interpretável.

Explainable artificial intelligence for match analysis in football

ABSTRACT

Recent discussions regarding the ethics on the application of machine learning algorithms for automated decision taking with significant impact on people's lives, particularly given the acknowledgment of several instances in which they reproduced the discriminative tendencies behind the historical data given as input, generated a significant demand for increasing transparency over these mechanisms.

Such efforts consolidated under an emerging field called Explainable Artificial Intelligence, or XAI. Its goal is to provide, for people with access to a learning agent, the means for them to understand the process behind its reasoning, with minimal impact on performance or predictive power. One of its possible applications is analyzing the weight that individual factors, or combinations thereof, have towards the model's outcome, which allows for a better optimization of efforts behind reaching the objective represented by it. The goal behind this work is to use such methods in order to investigate the influence exerted by factors present in association football (soccer) matches towards their respective outcomes, in a comprehensible format that might be further inspected by interested parties such as managers, coaches, and players, in order to facilitate the improvement of their tactics and training sessions.

Keywords: data analysis, machine learning, association football, soccer, explainable artificial intelligence.

LISTA DE ABREVIATURAS E SIGLAS

| | |
|--------|---|
| ALE | Accumulated Local Effects |
| AUC | Area Under Curve |
| CART | Classification And Regression Tree |
| DARPA | Defense Advanced Research Projects Agency |
| GDPR | General Data Protection Regulation |
| GBM | Gradient Boosting Machine |
| GLM | Generalized Linear Model |
| ICE | Individual Conditional Expectation |
| LIME | Local Interpretable Model-Agnostic Explanation |
| PDP | Partial Dependence Plot |
| PRISMA | Preferred Reporting Items for Systematic reviews and Meta-Analyses |
| SAMME | Stagewise Additive Modeling using a Multi-class Exponential loss function |
| SHAP | Shapley Additive Explanation |
| SVM | Support Vector Machine |
| XAI | Explainable Artificial Intelligence |

LISTA DE FIGURAS

| | | |
|------------|--|----|
| Figura 2.1 | Escopo da Inteligência Artificial Explicável | 24 |
| Figura 5.1 | Valores SHAP por oportunidades de gol | 34 |
| Figura 5.2 | Valores SHAP por total de passes..... | 34 |
| Figura 5.3 | Valores SHAP por taxa de acerto de passes | 35 |
| Figura 5.4 | Valores SHAP por posse de bola | 35 |
| Figura 5.5 | Valores SHAP por tempo de recuperação de bola..... | 36 |

LISTA DE TABELAS

| | |
|--|----|
| Tabela 3.1 Estudos voltados à análise preditiva | 29 |
| Tabela 4.1 Distribuição de partidas por campeonato | 30 |
| Tabela 4.2 Variáveis quantitativas | 31 |
| Tabela 4.3 Variáveis qualitativas | 31 |
| Tabela 4.4 Algoritmos de classificação implementados | 32 |
| Tabela 5.1 Desempenho dos classificadores | 33 |

LISTA DE QUADROS

| | |
|------------------------------------|----|
| Quadro 2.1 Regras do Futebol | 16 |
|------------------------------------|----|

SUMÁRIO

| | |
|---|-----------|
| 1 INTRODUÇÃO | 12 |
| 1.1 Problema de pesquisa | 12 |
| 1.2 Solução proposta | 13 |
| 1.3 Estrutura da dissertação | 14 |
| 2 FUNDAMENTAÇÃO | 15 |
| 2.1 História do futebol | 15 |
| 2.2 Técnicas de classificação | 21 |
| 2.2.1 Árvores de decisão | 21 |
| 2.2.2 Classificadores lineares | 21 |
| 2.2.3 Ensemble..... | 22 |
| 2.3 Inteligência artificial explicável | 24 |
| 3 TRABALHOS RELACIONADOS | 27 |
| 3.1 Visão geral | 27 |
| 3.2 Análise descritiva | 28 |
| 3.3 Análise comparativa | 28 |
| 3.4 Análise preditiva | 29 |
| 4 EXPERIMENTO | 30 |
| 4.1 Dados utilizados | 30 |
| 4.2 Pré-processamento | 31 |
| 4.3 Aprendizado e interpretação | 31 |
| 5 RESULTADOS | 33 |
| 6 CONCLUSÕES E POSSIBILIDADES FUTURAS | 37 |
| REFERÊNCIAS | 39 |

1 INTRODUÇÃO

No contexto de aprendizagem de máquina, a demanda por interpretabilidade é gerada pela dissonância entre os objetivos que motivam a criação de um modelo interpretativo e seus mecanismos de avaliação existentes (LIPTON, 2016). Os recentes estudos nesse tópico são movidos por preocupações relativas à equidade da tomada de decisões por esses sistemas no contexto social (LIPTON, 2016; PEDRESHI; RUGGIERI; TURINI, 2008), geradas principalmente com a descoberta de diversos modelos preditivos que reproduziram os preconceitos presentes nas composições de suas bases de dados (LOWRY; MACPHERSON, 1988; CALISKAN; BRYSON; NARAYANAN, 2017). Além disso, as regras atuais sobre utilização de dados da União Europeia presentes no *General Data Protection Regulation*, ou GDPR, garantem aos seus cidadãos o *direito à explicação* no contexto de inteligência artificial, isto é, eles podem requerer a justificativa de decisões tomadas de forma puramente automatizada que os impactem significativamente (GOODMAN; FLAXMAN, 2017).

Com isso, é de interesse de entidades privadas e governamentais que os modelos preditivos criados por elas permitam interpretações intuitivas de seus resultados. Um dos principais fomentadores atuais desse campo de pesquisa é o DARPA (*Defense Advanced Research Projects Agency*), que criou em conjunto com diversos centros de pesquisa nos Estados Unidos o programa XAI (*Explainable Artificial Intelligence*), cujo intuito é desenvolver uma nova geração de modelos interpretativos para uso em aprendizagem de máquina (GUNNING, 2016).

1.1 Problema de pesquisa

Concomitantemente, o uso de técnicas de aprendizado de máquina para prever os resultados finais de partidas de futebol já foi alvo de diversos estudos, tipicamente com o objetivo de explorar ineficiências em plataformas de apostas esportivas (HUANG; CHANG, 2010; CHENG et al., 2003; BAIQ; BLANGIARDO, 2010; ARABZAD et al., 2014; ROTSHTEIN; POSNER; RAKITYANSKAYA, 2005; RUE; SALVESEN, 2001; TSAKONAS et al., 2002; JOSEPH; FENTON; NEIL, 2006). No entanto, não há estudos significativos sobre o uso desse mecanismo computacional para compreensão dos fatores responsáveis pelos resultados previstos. Uma das principais causas consiste nas limitações inerentes aos modelos preditivos mais complexos, cuja eficácia superior é atingida

às custas da perda de interpretabilidade (LOU; CARUANA; GEHRKE, 2012).

Com o aumento significativo na disponibilidade de dados referentes aos eventos ocorridos ao longo de partidas de futebol, tornou-se viável uma vasta gama de oportunidades para pesquisa (REIN; MEMMERT, 2016). Esses dados permitem, por exemplo, calcular de forma independente métricas não-convencionais como o tempo de recuperação de posse de bola, cuja minimização é um componente central da filosofia *gegenpressing* popularizada pelo treinador Jürgen Klopp (WILSON, 2015), e demonstrar a sua correlação com o sucesso de uma equipe em um campeonato como a *Bundesliga* (VOGELBEIN; NOPP; HÖKELMANN, 2014).

1.2 Solução proposta

Os mecanismos existentes de interpretabilidade em aprendizagem de máquina são um campo inexplorado nesse cenário. Com a possibilidade de avaliar o impacto de uma ou mais métricas específicas no resultado final de uma previsão, em um formato acessível para usuários não-técnicos como treinadores e preparadores, é possível desenvolver táticas e treinamentos enfatizando a otimização desses fatores. Além disso, alguns dos modelos interpretativos apresentados são locais, isto é, aplicados a previsões individuais, permitindo assim a análise de partidas específicas, como derrotas recentes ou jogos do próximo adversário.

Esta Dissertação de Mestrado avalia e compara uma série de modelos preditivos baseados em algoritmos populares de aprendizagem de máquina, visando medir a influência de diferentes fatores no resultado final de uma partida de futebol. Tais modelos são posteriormente dissecados através de um conjunto de técnicas interpretativas, para responder à pergunta: como identificar os principais fatores mais significativos por trás de uma vitória, empate ou derrota de uma equipe?

A base de dados deste estudo é disponibilizada publicamente pela empresa *StatsBomb*, especializada no fornecimento de dados de eventos ocorridos em jogos de futebol, para uso não-comercial. A partir dela foram extraídas, para cada combinação de partida e equipe, 10 métricas cuja relevância no placar final foi alvo de estudos significativos. Os modelos produzidos foram baseados em algoritmos dos tipos Modelo Linear Generalizado (regressão logística), *Bagging* (*random forest* e *Extra-Trees*) e *Boosting* (*AdaBoost* e *XGBoost*). Posteriormente, calculou-se para cada modelo as métricas de escore F1, perda logarítmica, acurácia, kappa de Cohen e *Rank Probability Score*, além de visuali-

zações obtidas com a geração de modelos interpretativos do tipo *SHAP* (*Shapley Additive Explanations*).

1.3 Estrutura da dissertação

Os próximos capítulos se dividem da seguinte forma: O capítulo 2 introduz de forma detalhada os conceitos utilizados ao longo da dissertação, referentes ao uso de técnicas estatísticas e computacionais no estudo do futebol profissional, os principais modelos utilizados em aprendizagem de máquina e as técnicas de interpretabilidade aplicáveis a estes. O capítulo 3 traz estudos anteriores referentes ao uso de métodos preditivos aplicados a resultados de partidas de futebol. O capítulo 4 descreve o experimento realizado e o processo de execução, otimização, comparação e interpretação dos modelos obtidos, cujos resultados são exibidos no capítulo 5. Por fim, o capítulo 6 traz as conclusões obtidas, além de ponderar sobre possibilidades de melhora e de futuros trabalhos voltados a estender o tópico da pesquisa.

2 FUNDAMENTAÇÃO

O presente capítulo apresenta os resultados da pesquisa sobre os tópicos que este trabalho abrange, necessária para a sua elaboração. A Seção 2.1 traz um breve sumário da criação do futebol e sua importância no cenário mundial. A Seção ?? apresenta um histórico das técnicas de análise de desempenho aplicadas ao esporte. A Seção 2.2 descreve os algoritmos de classificação utilizados neste trabalho. Por fim, a Seção 2.3 apresenta o campo de inteligência artificial explicável.

2.1 História do futebol

O futebol associação, variante mais popular do esporte e conhecido no Brasil como apenas "futebol", teve seu primeiro conjunto de regras oficial publicado no ano de 1863, quando sua elaboração provocou a cisão definitiva entre esta modalidade e o *rugby* (ALCOCK, 1906). O esporte foi trazido ao Brasil por operários ingleses em 1885 e se tornou popular nos primeiros anos do século XX, alcançando estrondoso sucesso e sendo, até hoje, o principal esporte nacional (GUTERMAN, 2013). É um esporte praticado por homens e mulheres, embora seja notório que o universo do futebol seja caracterizado, “desde sua origem, um espaço eminentemente masculino” (FRANZINI, 2005, p. 316). O futebol é um jogo desportivo coletivo (??), ou seja, representa, segundo Lovatto and Galatti (2007, p. 269)

uma forma de atividade social organizada, uma forma específica de manifestação e de prática, com caráter lúdico e processual, de exercício físico, na qual os participantes (jogadores) estão agrupados em duas equipes numa relação de adversidade típica não hostil (rivalidade desportiva – relação determinada pela disputa através de luta com vista à obtenção da vitória desportiva, com ajuda da bola (ou de outro objeto de jogo) manobrada de acordo com regras pré-estabelecidas.

Atualmente, o futebol é o esporte mais popular do planeta, ganhando espaço inclusive em países onde ele nunca foi expressivo, como China e Estados Unidos (BOUDWAY, 2018). O futebol possui um conjunto de dezessete regras, que estão sumarizadas no Quadro 2.1.

Quadro 2.1: Regras do Futebol

| | |
|----------------|--|
| Regra 1 | <p>O Campo de Jogo</p> <p>A superfície deve ser natural ou artificial, na cor verde. O campo deve ser retangular, marcado com linhas contínuas, dividido em duas metades (meio de campo) por uma linha de meio de campo (linha central) que une os pontos médios das linhas laterais. O centro do campo é marcado com um ponto, em volta do qual há um círculo com um raio de 9,15 metros.</p> <p>Devem ser seguidas as normas estabelecidas para as linhas e posições da área de meta, área penal, área de tiro de canto, postes de bandeiras, área técnica e metas.</p> <p>Pode ser utilizada tecnologia de linhas de meta para suporte ao árbitro.</p> |
| Regra 2 | <p>A Bola</p> <p>Deverá ser esférica, feita com material adequado, ter circunferência de 68 a 70 cm, pesar entre 410 e 450 gramas no começo do jogo e ter pressão equivalente a 0,6 – 1,1 atmosferas ($600 - 1100g/cm^2$) ao nível do mar.</p> |
| Regra 3 | <p>Os Jogadores</p> <p>As partidas são disputadas por duas equipes compostas por onze jogadores, sendo um dos quais o goleiro. O número máximo de substituições é cinco – quando uma partida for para prorrogação, uma substituição adicional pode ser feita. O capitão do time é o responsável pela conduta da sua equipe em campo. Agentes externos (pessoas que não são jogadores, substitutos ou oficiais da equipe) são proibidas no campo de jogo e, se entrarem em campo, a partida é paralisada.</p> |

| | |
|----------------|--|
| Regra 4 | O Equipamento dos Jogadores O equipamento obrigatório dos jogadores é composto por: camiseta com mangas, calções, meias, caneleiras (devem ser cobertas pelas meias) e calçado. Os times devem usar cores diferentes entre si e também cores diferentes dos oficiais de arbitragem. Não é permitido aos jogadores o uso de quaisquer equipamentos eletrônicos ou sistemas de comunicação durante o jogo. |
| Regra 5 | O Árbitro O árbitro do jogo tem a autoridade para fazer as regras do jogo serem cumpridas e deve tomar suas decisões de acordo com as regras e o "espírito do jogo". O árbitro tem a colaboração dos demais oficiais do time de arbitragem. |
| Regra 6 | Os Outros Oficiais de Arbitragem Os outros oficiais de arbitragem são compostos por: dois árbitros assistentes, quarto árbitro, dois árbitros assistentes adicionais, árbitro assistente reserva, árbitro assistente de vídeo, árbitro assistente de vídeo reserva. |
| Regra 7 | A Duração do Jogo O jogo tem duração de dois períodos de 45 minutos, com intervalo de 15 minutos entre os dois períodos. Os períodos podem ser acrescidos do tempo perdido em razão de substituições, avaliações de lesões, sanções disciplinares, entre outros. |

| | |
|-----------------|--|
| Regra 8 | O Início e Reinício do Jogo O jogo começa com um tiro inicial (saída de bola) em cada um dos períodos, nas prorrogações e quando um gol for marcado. Outras formas de reiniciar o jogo são os tiros livres (diretos e indiretos), pênaltis, arremessos laterais, tiros de meta e tiros de canto. |
| Regra 9 | A Bola em Jogo e Fora de Jogo A bola está fora de jogo: (1) quando transpuser completamente uma linha de meta ou a linha lateral ou (2) o jogo for interrompido pelo árbitro. Em todas as outras situações, a bola está em jogo. |
| Regra 10 | Determinação do Resultado de um Jogo O jogo é vencido pela equipe que marcar o maior número de gols. Se o número de gols for o mesmo, o jogo termina empatado. Um gol é marcado quando a bola transpõe completamente a linha de meta, entre os postes e abaixo do travessão, sendo que nenhuma infração tenha sido cometida pela equipe que marcou o gol. Quando um gol acontece, o jogo reinicia. |
| Regra 11 | Impedimento Um jogador está impedido quando estiver na metade do campo adversário e estiver mais próximo da linha de meta adversária do que a bola e o penúltimo adversário (mãos e braços não são considerados na regra do impedimento). Não está impedido o jogador que estiver em linha com o penúltimo adversário. |

| | |
|-----------------|--|
| Regra 12 | Faltas e Incorreções Tiros livres e pênaltis são marcados por faltas e incorreções que acontecem quando a bola estiver em jogo, tais como: fazer carga, saltar, dar ou tentar dar um pontapé em um adversário, empurrar, golpear ou tentar golpear (incluindo cabeçada) um adversário, dar uma entrada, um calço ou uma rasteira em um adversário, morder os cuspir em alguém, pegar a bola com as mãos, entre outros. |
| Regra 13 | Tiros Livres Os tiros livres direto e indiretos são concedidos a favor da equipe adversária do jogador ou do oficial da equipe que cometer falta ou infração. Os tiros livres devem ser executados onde a infração foi cometida, com exceção dos tiros livres indiretos, que são executados do ponto mais próximo da linha da área de meta, paralela à linha de meta. |
| Regra 14 | Tiro Penal (Pênalti) O pênalti é marcado quando o jogador faz uma infração a ser punida com tiro livre direto estando dentro da sua área penal. |
| Regra 15 | O Arremesso Lateral O arremesso lateral é marcado para a equipe adversária do jogador que tocar por último a bola antes dela sair do campo pela linha lateral. Um gol não pode ser marcado diretamente de um arremesso lateral. |

| | |
|-----------------|--|
| Regra 16 | <p>O Tiro de Meta</p> <p>O tiro de meta é marcado para a equipe adversária do jogador atacante quando, sem que um gol tenha sido marcado, a bola ultrapassar totalmente a linha de meta. Um gol pode ser marcado diretamente de um arremesso lateral, mas somente contra a equipe adversária. Se depois do tiro de meta a bola entrar na equipe do executante, será marcado tiro de canto a favor da equipe adversária.</p> |
| Regra 17 | <p>O Tiro de Canto</p> <p>O tiro de canto é marcado para a equipe defensora do jogador quando, sem que um gol tenha sido marcado, a bola ultrapassar totalmente a linha de meta. Um gol pode ser marcado diretamente de um arremesso lateral, mas somente contra a equipe adversária. Se depois do tiro de meta a bola entrar na equipe do executante, será marcado tiro de canto a favor da equipe adversária.</p> |

Fonte: IFAB (2018)

Além destas regras, o regulamento oficial do esporte, denominado *Laws of the Game*, define ainda diretrizes para exclusão temporária e para retorno de substitutos, além de linhas de orientação prática para oficiais de arbitragem quanto a posicionamento, movimentação e trabalho de equipe, linguagem corporal, comunicação e apito, vantagens, recuperação de tempo perdido, agarrar adversário e impedimento (IFAB, 2018). Ainda, a Federação Internacional de Futebol Associação (FIFA), associação internacional que dirige as confederações de futebol do mundo inteiro, define um código disciplinar com condições, sanções e medidas disciplinares aplicáveis em cada caso (FIFA, 2017).

Uma das competições organizadas pela FIFA é a Copa do Mundo de Futebol (FIFA, 2018). A Copa do Mundo, em sua edição com seleções masculinas, é um evento mundial realizado a cada quatro anos, onde seleções de 32 países disputam entre si o prêmio de melhor seleção (GLANVILLE, 2005). A primeira edição da Copa do Mundo aconteceu em 1930, no Uruguai, tendo como campeão o país-sede (GLANVILLE, 2005) e a última edição ocorreu em 2018 na Rússia, sendo a seleção francesa a campeã (FIFA,

2018).

2.2 Técnicas de classificação

O objetivo de um modelo de classificação, ou classificador, é o de, a partir do fornecimento de um dado, obter uma resposta correspondente a uma ou mais variáveis qualitativas (JAMES et al., 2014). Suas implementações mais populares serão descritas a seguir.

2.2.1 Árvores de decisão

Os algoritmos de aprendizagem por árvores de decisão consistem em uma família de sistemas que expressam, a partir de um processo de inferência indutiva por exemplos, um modelo preditivo baseado na travessia de uma estrutura de árvore cujos nodos internos correspondem a testes a serem aplicados até chegar nos resultados representados por suas folhas (QUINLAN, 1986). Suas variantes mais populares são os algoritmos C4.5 (QUINLAN, 1993) e CART (BREIMAN et al., 1984).

A principal vantagem dessa classe de algoritmos é a geração de modelos *caixa-branca* de fácil compreensão, devido à sua estrutura intuitiva; entretanto, o desempenho delas é por vezes insatisfatório pois sua otimização consiste em um problema NP-completo (HYAFIL; RIVEST, 1976). Ainda assim, eles são um componente essencial em diversos modelos mais robustos do tipo *ensembling*, descritos posteriormente.

2.2.2 Classificadores lineares

Classificadores lineares são modelos cujo processo de avaliação se baseia em de uma combinação linear de fatores do dado recebido. Segundo a definição de Ng and Jordan (2002), eles são exemplos de modelos discriminativos, os quais calculam as probabilidades envolvidas diretamente em vez de se basearem em uma função de densidade de probabilidade intermediária.

A variedade mais comum desses modelos utiliza a técnica de regressão logística, a qual modela o problema de classificação como uma função logística com parâmetros correspondentes às variáveis conhecidas, que calcula a probabilidade de pertencimento a

uma determinada categoria, e otimiza seus respectivos coeficientes de forma a minimizar uma função de custo (COX, 1958). Posteriormente, Nelder and Wedderburn (1972) introduziram o conceito de *modelo linear generalizado* como uma forma de unificar os modelos de regressão conhecidos à época, em função das distribuições de probabilidade assumidas por cada um deles; dessa forma, foi possível estabelecer um padrão para a estimativa de seus parâmetros.

Com o objetivo de controlar os coeficientes desses modelos, penalizando aqueles que possam causar efeitos adversos como alta variância (que aumenta o risco de sobreajuste), excesso de complexidade computacional ou difícil interpretabilidade, foram desenvolvidas técnicas de regularização que os tornam parte dos cálculos das estimativas (JAMES et al., 2014).

A regressão *ridge*, ou regularização de Tikhonov, acrescenta à função de estimação a norma L2 do vetor de coeficientes, isto é, a soma de seus quadrados, multiplicada por uma constante α (HOERL; KENNARD, 1970). Tal solução é computacionalmente eficiente devido à possibilidade de obtê-la por métodos analíticos, e coíbe o sobreajuste causado por excesso de variância.

Por sua vez, o método LASSO (Least Absolute Shrinkage and Selection Operator) realiza um processo análogo utilizando a norma L1, que consiste na soma dos valores absolutos dos coeficientes. Como resultado, alguns deles convergem a 0 ao longo do processo de treinamento, gerando assim modelos mais esparsos e, conseqüentemente, com um menor número de variáveis relevantes (TIBSHIRANI, 1996). No entanto, existem limitações: em problemas de alta dimensionalidade, como sequenciamento genético ou processamento de imagens, o número de variáveis selecionadas fica limitado à quantidade de exemplos de treinamento, e quando existe uma correlação entre múltiplas variáveis apenas uma delas é selecionada arbitrariamente (ZOU; HASTIE, 2005).

Como forma de superar essas limitações, Zou and Hastie (2005) desenvolveram uma técnica de penalização que utiliza as normas L1 e L2 em conjunto, chamada de *elastic net*, permitindo assim gerar um modelo esparsos com uma seleção de variáveis mais ampla.

2.2.3 Ensemble

Como uma forma de compensar as deficiências presentes em algoritmos de classificação, uma solução comumente utilizada é a de combinar múltiplos modelos discordan-

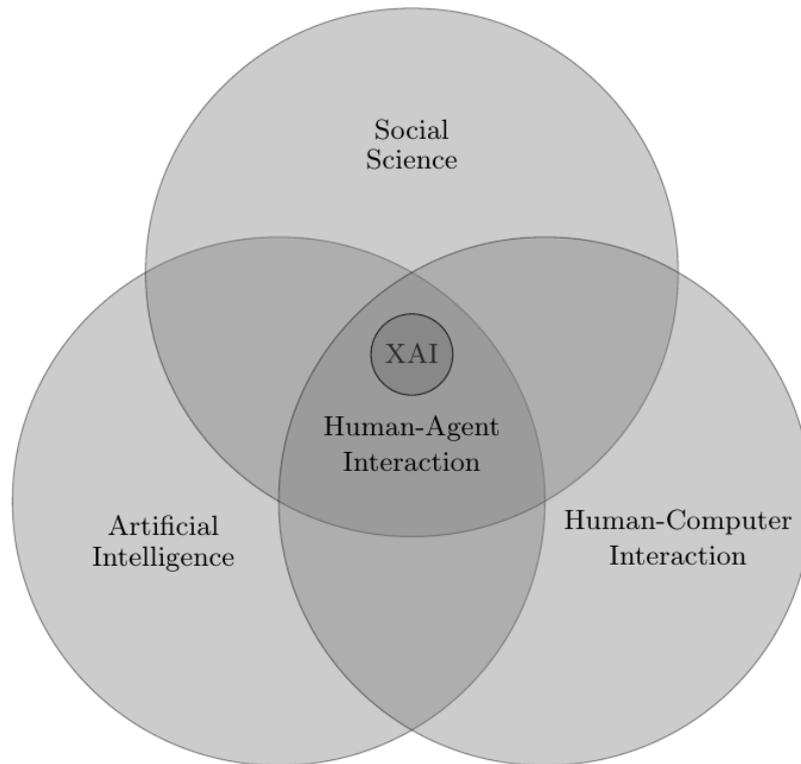
tes de forma que os erros individuais de cada um sejam atenuados no resultado final; ao modelo gerado por essa técnica é dado o nome de *ensemble* (OPITZ; MACLIN, 1999). Classificadores desse tipo podem ser divididos em dependentes e independentes; a diferença entre eles é que, no caso dos métodos dependentes, cada modelo gerado traz a influência de seus antecessores (ROKACH, 2010).

Os métodos dependentes com maior utilização são os do tipo *boosting*, que se baseiam na geração de uma grande quantidade de classificadores pouco eficientes cuja combinação resulta em um modelo robusto. O primeiro algoritmo a aplicar essa heurística com desempenho satisfatório foi o AdaBoost, cujo diferencial estava em priorizar, ao longo do processo de treinamento, a geração de modelos que classifiquem corretamente exemplos onde seus antecessores não obtiveram êxito, de forma que a combinação deles possa identificar uma variedade maior de casos com sucesso (FREUND; SCHAPIRE, 1997).

A posterior abordagem do *boosting* como um algoritmo iterativo de descida de gradiente (MASON et al., 1999), isto é, que percorre o espaço de soluções de forma a minimizar uma função de custo, inspirou o desenvolvimento de diversas implementações. Uma delas em particular, a biblioteca XGBoost (CHEN; GUESTRIN, 2016), obteve grande popularidade com seu uso em diversas soluções vencedoras de competições no campo de aprendizagem de máquina (GORMAN, 2017).

No caso dos métodos independentes, o seu principal representante é o *bootstrap averaging*, ou *bagging*, em que o conjunto de treinamento de cada modelo individual é gerado por amostragem com repetição. Esse método é particularmente eficiente quando aplicado a modelos com bom desempenho porém alta variância, sendo assim suscetíveis a mudanças significativas após pequenas alterações nos dados de entrada (BREIMAN, 1996). Sua implementação mais popular é o *random forest*, que gera árvores de decisão cujo processo de amostragem também engloba a seleção de suas variáveis, de forma a reduzir a correlação entre os modelos e, assim, evitar o risco de sobreajuste (BREIMAN, 2001). A variante Extra-Trees traz outra abordagem para o aumento da variabilidade, substituindo a amostragem das instâncias de treinamento pela geração de árvores com pontos de corte aleatórios (GEURTS; ERNST; WEHENKEL, 2006).

Figura 2.1: Escopo da Inteligência Artificial Explicável



Fonte: Miller (2017)

2.3 Inteligência artificial explicável

A inteligência artificial explicável, ou XAI, consiste em um agente que possui a capacidade de revelar as causas por trás das tomadas de decisões dele ou de um agente externo. Assim, ele consiste em um problema de *interação humano-agente*, que pode ser definida como a intersecção entre os campos de inteligência artificial, ciências sociais e interação humano-computador (MILLER, 2017).

Como alternativa ao uso de modelos preditivos de interpretação trivial, cuja simplicidade por vezes traz impacto em sua acurácia ou desempenho (LOU; CARUANA; GEHRKE, 2012), foram desenvolvidos métodos de interpretação de resultados que independem do modelo de aprendizagem, ou modelo-agnósticos. A principal vantagem trazida por eles é a flexibilidade; por não dependerem do funcionamento interno do sistema-alvo, este pode trabalhar de uma maneira muito mais flexível, além de poder ser comparado diretamente com outros modelos através dos mesmos mecanismos de interpretação (RIBEIRO; SINGH; GUESTRIN, 2016a).

O mais antigo método interpretativo modelo-agnóstico registrado na literatura é o

Partial Dependence Plot, introduzido por Friedman (2001) para a análise do algoritmo de *gradient boosting*. Seu objetivo é o de apresentar o efeito médio de uma variável específica sobre a previsão gerada, através do cálculo da estimativa do seu efeito marginal em relação ao restante dos fatores.

PDPs são gráficos de compreensão intuitiva e clara, por mostrarem diretamente o impacto das variáveis representadas no resultado final. No entanto, como todo o intervalo observado da variável é levado em consideração no cálculo, são geradas combinações com baixíssima probabilidade de ocorrerem. Além disso, como apenas a tendência média é representada, resultados mais heterogêneos não são representados apropriadamente (GOLDSTEIN et al., 2013). Cada um desses casos motivou o desenvolvimento de uma variação do PDP: gráficos de *Individual Conditional Expectation*, ou ICE, retratam a variação em cada exemplo do conjunto de treinamento de forma individual, permitindo assim identificar múltiplas tendências paralelas (GOLDSTEIN et al., 2013), enquanto o ALE (*Accumulated Local Effects*) calcula para cada valor as variações nas previsões ao longo de um intervalo reduzido (APLEY, 2016).

Outra alternativa é o uso de modelos substitutos, que consistem em sistemas do tipo caixa-branca, de interpretabilidade conhecida e intuitiva, treinados para gerar aproximações do modelo preditivo. Estes podem ser locais, focados na análise de exemplos específicos, ou globais, visando explicar o modelo preditivo como um todo (RIBEIRO; SINGH; GUESTRIN, 2016a). Atualmente se prioriza a criação de modelos locais, pois a interpretação de predições específicas é um objetivo mais comum e exige métodos menos robustos. O primeiro algoritmo desenvolvido para a geração de um modelo substituto local foi o LIME, ou *Local Interpretable Model-Agnostic Explanations* (RIBEIRO; SINGH; GUESTRIN, 2016b). Ele se baseia na geração de um conjunto de dados formado inteiramente por perturbações do exemplo a ser estudado, com pesos correspondentes às distâncias deste, para o treinamento do modelo substituto. Tal método pode ser aplicado não apenas em dados tabulares, mas também em problemas como reconhecimento de imagens, visão computacional e processamento de linguagem natural, delimitando os pontos que contribuiriam de alguma forma para a previsão obtida.

No entanto, ele traz algumas deficiências, como a instabilidade, que introduz o risco de gerar modelos significativamente diferentes para exemplos similares, e a natureza *ad hoc* do processo de amostragem, que assume variáveis não-correlacionadas seguindo uma distribuição normal (ALVAREZ-MELIS; JAAKKOLA, 2018).

O LIME foi posteriormente generalizado com o nome de *modelo de atribuição*

aditiva de fatores, que consiste em um modelo substituto local baseado em um classificador linear aplicado a simplificação do exemplo a ser estudado (LUNDBERG; LEE, 2017). Tal categorização inclui também o modelo de amostragem de valores de Shapley, baseado em uma métrica da teoria dos jogos voltada a medir a distribuição dos lucros para cada participante de um jogo cooperativo (SHAPLEY, 1953); neste contexto, o lucro consiste na diferença entre a previsão e a média geral do modelo, e para cada variável é calculada sua participação nessa diferença. No entanto, como a otimização dessa métrica é um problema de complexidade exponencial em relação ao número de variáveis, exigindo a geração de 2^k candidatos para k valores finais, esse modelo utiliza um processo de amostragem através do método de Monte Carlo (ŠTRUMBELJ; KONONENKO, 2014).

Os valores de Shapley possuem quatro propriedades que favorecem seu uso para o cálculo da importância de cada agente participante:

Eficiência A soma dos valores de todos os agentes é igual ao lucro previsto

Simetria Agentes que contribuem de forma igual em todos os cenários possuem o mesmo valor

Agente irrelevante Agentes que não apresentam nenhuma contribuição possuem um valor igual a 0

Aditividade Quando dois jogos distintos possuem os mesmos agentes, a distribuição da soma dos lucros corresponde à soma dos valores obtidos para cada agente

Assim, foi desenvolvido o método SHAP (Shapley Additive Explanation), que consiste em uma modificação do LIME cuja função de custo aproxima os valores de suas variáveis para seus respectivos valores de Shapley (LUNDBERG; LEE, 2017).

3 TRABALHOS RELACIONADOS

A análise de desempenho é um componente essencial para o sucesso de uma equipe de futebol profissional, por permitir a geração de conhecimento sobre jogadores e táticas que pode ser utilizado para aprimorar treinamentos, guiar a filosofia de jogo empregada, prospectar jogadores com atributos desejáveis ou investigar fraquezas de equipes adversárias, dentre inúmeras possibilidades. Os parâmetros de desempenho de uma equipe também são utilizados no mercado de apostas para melhor determinar as probabilidades de resultados de partidas, e com previsões mais precisas é possível detectar possíveis ineficiências.

Este capítulo faz um resumo de três modalidades distintas de análise que podem ser utilizadas para tal. A Seção 3.1 descreve tais abordagens em linhas gerais. Já a Seção 3.2 apresenta os trabalhos focados na análise descritiva. A Seção 3.3 expõe os estudos voltados à análise comparativa. Por fim, a Seção 3.4 mostra as pesquisas envolvendo a análise preditiva.

3.1 Visão geral

A análise de desempenho voltada ao futebol apresentou um crescimento expressivo nas últimas décadas, atraindo o interesse das entidades envolvidas (Carling et al., 2005; James, 2006) e sendo alvo direto de pesquisas por parte dos clubes (Carling et al., 2005; Groom & Cushion, 2004). Três revisões de literatura recentes, com objetivos distintos, se complementam como um sumário da evolução e do desenvolvimento na área (MACKENZIE; CUSHION, 2013; SARMENTO et al., 2014; SARMENTO et al., 2018).

O estudo realizado por Marcelino, Sampaio and Mesquita (2011) para revisar a literatura envolvendo análise de desempenho para esportes coletivos em geral, dividiu os trabalhos encontrados em três categorias: *análise descritiva*, *análise comparativa* e *análise preditiva*. Com base na mesma categorização, Sarmiento et al. (2014) apresentaram uma revisão com foco específico no futebol, conduzida de acordo com as diretrizes PRISMA (*Preferred Reporting Items for Systematic reviews and Meta-analyses*), a qual foi complementada por um estudo recente com o objetivo de mostrar o desenvolvimento posterior ao período analisado (SARMENTO et al., 2018).

Paralelamente, Mackenzie and Cushion (2013) tiveram um objetivo mais crítico, contestando a abordagem positivista predominante e seu papel na geração de modelos

simplistas e deficientes de contextualização, além da ausência de evidências do impacto desses estudos sob a prática do esporte.

3.2 Análise descritiva

As análises descritivas são voltadas a identificar e sistematizar indicadores globais de rendimento, sem buscar contextualização, e tipicamente categorizam padrões fisiológicos de movimentação dos jogadores de acordo com uma escala de intensidade (MARCELINO; SAMPAIO; MESQUITA, 2011). Embora limitados pela ausência de estratificação entre fatores como posicionamento tático ou resultado parcial da partida, tais estudos permitem visualizar um panorama das demandas fisiológicas geradas.

3.3 Análise comparativa

O objetivo da análise comparativa é estudar a variação das métricas de desempenho entre diferentes categorias de um ou mais fatores (MARCELINO; SAMPAIO; MESQUITA, 2011), sendo os mais comuns no futebol a posição tática dos jogadores e o nível competitivo das equipes (SARMENTO et al., 2014).

Os trabalhos voltados à comparação entre diferentes papéis táticos buscam compreender as diferentes demandas físicas e técnicas de cada posição, seja através de indicadores fisiológicos como os obtidos nos estudos de análise descritiva, seja com as ações realizadas ao longo da partida, com o objetivo de otimizar os programas de treinamento para os requisitos específicos das diferentes posições.

Já a comparação entre níveis competitivos pode se realizar entre equipes que disputam a mesma competição, seguindo a classificação de cada uma, ou mesmo entre níveis de profissionalismo, sejam profissionais, semi-profissionais ou amadoras. Em ambos os casos, é possível distinguir claramente as demandas trazidas pelos diferentes níveis avaliados, com maior esforço físico e ações mais frequentes por parte de equipes de qualidade superior.

Outros aspectos mais raramente estudados envolvem principalmente o resultado final de cada partida, mostrando diferenças nítidas entre as métricas obtidas em casos de vitória, empate e derrota, e os efeitos de variáveis diversas como a densidade de partidas no calendário, substituições, formações táticas, períodos como o Ramadã e condições

Tabela 3.1: Estudos voltados à análise preditiva

| Métrica | Estudos |
|--|--|
| Distância média percorrida por jogador | Lago-Peñas et al. (2011) |
| Número de gols | Hughes and Franks (2005), Tenga et al. (2010b), Tenga, Ronglan and Bahr (2010) |
| Número de oportunidades de gol | Hughes and Franks (2005), Tenga, Ronglan and Bahr (2010) |
| Probabilidade de converter oportunidades de gol | Pollard and Reep (1997), Hughes and Franks (2005) |
| Probabilidade de chegar à grande área adversária | Tenga et al. (2010a), Tenga, Ronglan and Bahr (2010), Lago-Ballesteros, Lago-Peñas and Rey (2012) |
| Probabilidade de vitória, empate ou derrota | Lago-Peñas et al. (2010), Lago-Peñas and Lago-Ballesteros (2011), Castellano, Casamichana and Lago (2012), Harrop and Nevill (2014), Lago-Peñas et al. (2016), Moura, Martins and Cunha (2014) |

Fonte: Sarmento et al. (2014), Sarmento et al. (2018)

ambientais como a altitude e a qualidade do ar.

3.4 Análise preditiva

A análise preditiva visa, com base em dados históricos de partidas, estimar os valores de uma ou mais métricas a partir de determinados indicadores de rendimento. Tais estudos se mostraram extremamente escassos durante muitos anos: Sarmento et al. (2014) listam apenas dois cuja data de publicação precede a presente década (POLLARD; REEP, 1997; HUGHES; FRANKS, 2005).

Nos últimos anos, houve uma predominância de trabalhos voltados à previsão do resultado de partidas: vitória, empate ou derrota. As variáveis analisadas neles, tipicamente, foram números absolutos: oportunidades de gol (totais, no alvo ou fora), cruzamentos, escanteios, faltas (sofridas e cometidas) e impedimentos (também sofridos e cometidos), além do percentual de posse de bola.

4 EXPERIMENTO

Este capítulo descreve o experimento em interpretabilidade no qual o presente trabalho se baseia, cujo objetivo é a detecção de padrões em modelos de classificação por aprendizagem de máquina voltados à predição de resultados de partidas de futebol. A seção 4.1 descreve a base de dados utilizada para o treinamento dos classificadores. Posteriormente, a seção 4.2 descreve o pré-processamento aplicado a esses dados para a geração das métricas de desempenho relevantes. Já a seção 4.3 apresenta os modelos preditivos utilizados, seu processo de treinamento e como os modelos interpretáveis serão usados.

4.1 Dados utilizados

Para este estudo foi utilizada uma base de dados disponibilizada publicamente pela provedora de dados StatsBomb, através de um repositório hospedado no serviço GitHub, para uso não-comercial (StatsBomb, 2019). Ela consiste em dados relativos a todos os eventos ocorridos ao longo das partidas representadas, que à época da realização deste trabalho totalizavam 170, distribuídas em três competições: a edição de 2018 da *FIFA World Cup*, a temporada 2018 da *National Women's Soccer League*, nos Estados Unidos, e a temporada 2018—19 da *FA Women's Super League*, na Inglaterra. Ela é a primeira base de dados aberta a disponibilizar partidas de futebol com tamanha granularidade, o que reduz a barreira de entrada para pesquisas nesse campo e facilita a reprodutibilidade dos trabalhos realizados.

Tabela 4.1: Distribuição de partidas por campeonato

| Campeonato | Abrangência | Gênero | Edição | # partidas |
|--------------------------------|----------------|-----------|---------|------------|
| FIFA World Cup | Mundial | Masculino | 2018 | 64 |
| National Women's Soccer League | Estados Unidos | Feminino | 2018 | 34 |
| FA Women's Super League | Inglaterra | Feminino | 2018-19 | 71 |

Fonte: O Autor

Tabela 4.2: Variáveis quantitativas

| Variável | Mínimo | Mediana | Média | Máximo | D.P. | Descrição |
|-----------------------|--------|---------|---------|--------|---------|--|
| avg.recovery.time | 8,884 | 29,926 | 31,569 | 66,876 | 10,379 | Recuperação de posse de bola média (s) |
| avg.player.possession | 0,627 | 1,346 | 1,356 | 2,298 | 0,243 | Posse de bola média por jogador (s) |
| avg.team.possession | 0,425 | 0,503 | 0,521 | 0,757 | 0,067 | Posse de bola (%) |
| shots | 0 | 13 | 13,833 | 43 | 6,845 | Chutes a gol |
| expulsions | 0 | 0 | 0,025 | 1 | 0,157 | Jogadores expulsos |
| passes | 224 | 456,5 | 470,101 | 1157 | 132,845 | Passes iniciados |
| successful.passes | 138 | 405 | 421,714 | 1113 | 139,343 | Passes completados |
| corners | 0 | 4 | 4,516 | 20 | 2,928 | Escanteios cobrados |
| fouls | 2 | 11 | 11,808 | 26 | 4,781 | Faltas cometidas |

Fonte: O Autor

Tabela 4.3: Variáveis qualitativas

| Variável | # níveis | Frequência | Descrição | Valores |
|----------|----------|----------------|-------------------|-----------------------------|
| home | 2 | 0: 223; 1: 95 | Equipe mandataria | 0 = Sim; 1 = Não |
| women | 2 | 0: 128; 1: 190 | Gênero da equipe | 0 = Masculino; 1 = Feminino |

Fonte: O Autor

4.2 Pré-processamento

Para cada combinação de partida e equipe, totalizando 340 exemplos, foi extraída uma instância de treinamento contendo as métricas apresentadas na tabela 4.2 e os fatores presentes na tabela 4.3, além de uma variável-alvo correspondendo ao resultado obtido pela equipe na partida (vitória, empate ou derrota). No caso das partidas da Copa do Mundo, a variável *home* foi considerada sempre falsa, pois a natureza da competição torna esse fator irrelevante.

4.3 Aprendizado e interpretação

A implementação dos modelos preditivos foi realizada na linguagem de programação Python com o auxílio da biblioteca scikit-learn (PEDREGOSA et al., 2011). Aos parâmetros de cada algoritmo foi aplicado o otimizador bayesiano da biblioteca scikit-optimize (HEAD et al., 2018), utilizando como prior um processo gaussiano e como objetivo a minimização do Rank Probability Score, ou RPS (EPSTEIN, 1969). Tal métrica foi escolhida para a otimização deste problema por ser voltada a variáveis qualitativas ordinais e utilizar a distância entre os valores como parte do cálculo (MURPHY, 1970),

Tabela 4.4: Algoritmos de classificação implementados

| Algoritmo | Modelo | Ensemble | Referência |
|---------------------|----------------------|----------------|-----------------------------------|
| Regressão logística | Classificador linear | Não | Cox (1958) |
| Random forest | Árvores de decisão | Sim (bagging) | Breiman (2001) |
| Extra-Trees | Árvores de decisão | Sim (bagging) | Geurts, Ernst and Wehenkel (2006) |
| AdaBoost.SAMME | Árvores de decisão | Sim (boosting) | Zhu et al. (2009) |
| Gradient boosting | Árvores de decisão | Sim (boosting) | Chen and Guestrin (2016) |

Fonte: O Autor

permitindo assim que vitórias e derrotas sejam avaliadas como mais próximas de empates do que uma à outra (CONSTANTINO; FENTON, 2012). Como forma de minimizar o risco de sobreajuste no modelo, e conseqüentemente de perda de capacidade de generalização, foi aplicada nesse processo de otimização a técnica de validação cruzada *3-fold*.

Além do RPS, outras métricas foram calculadas para fins de comparação: kappa de Cohen, logarithmic loss, acurácia, e score F1 macro. Por fim, são gerados os valores SHAP, a partir do modelo preditivo de melhor desempenho, para cada exemplo gerado.

5 RESULTADOS

A análise dos valores SHAP mostra como o modelo preditivo analisado encontrou padrões que merecem uma análise mais aprofundada.

Por exemplo, a figura 5.1 apresenta um impacto positivo nos cenários de vitória e empate apenas quando se ultrapassa a marca de 20 oportunidades de gol. Além disso, os valores para o cenário de empate são significativamente menores em uma situação de dominância da posse de bola.

Já a figura 5.2 apresenta padrões bastante distintos em relação ao impacto da quantidade de passes. A troca reduzida de passes é considerada um fator positivo para chances de vitória, mesmo que isso implique em menos oportunidades de gol, enquanto para valores maiores ela é negativa exceto por um intervalo restrito e crescente em cenários de empate. Uma possível interpretação envolve a facilidade de aplicação da estratégia de bola longa, enquanto táticas que envolvem a manutenção da posse exigem certo equilíbrio para não imobilizar o jogo.

Por outro lado, a figura 5.3 mostra um impacto significativo em vitórias apenas quando a troca de passes apresenta uma taxa de acerto superior a 90%, independente de fatores como posse de bola e pressão defensiva. Quando combinada com a figura anterior, isso aponta uma ausência de superioridade entre a tática de ligação direta e aquela visando a manutenção de posse de bola; no entanto, convém fazer uma escolha entre essas duas estratégias.

Os valores SHAP encontrados em relação à posse de bola na figura 5.4 mostram uma tendência negativa para vitórias e empates abaixo da marca de 50%. Acima desse valor o impacto nos empates é positivo, porém, com altíssima variância, enquanto para vitórias ele atinge o ápice aos 60%, mantendo-se nesse mesmo patamar acima dessa taxa.

Tabela 5.1: Desempenho dos classificadores

| Classificador | RPS | Kappa | Log loss | F1 | Acurácia |
|---------------------|---------|---------|----------|---------|----------|
| Random forest | 0.16626 | 0.44203 | 0.82703 | 0.55966 | 0.66177 |
| Gradient boosting | 0.17058 | 0.39090 | 0.86426 | 0.53790 | 0.63235 |
| Extra-Trees | 0.18142 | 0.37638 | 0.88640 | 0.46181 | 0.63235 |
| Regressão logística | 0.19302 | 0.41505 | 0.90147 | 0.51886 | 0.64706 |
| AdaBoost-SAMME | 0.24060 | 0.43471 | 1.07585 | 0.56909 | 0.64706 |

Fonte: O Autor

Figura 5.1: Valores SHAP por oportunidades de gol

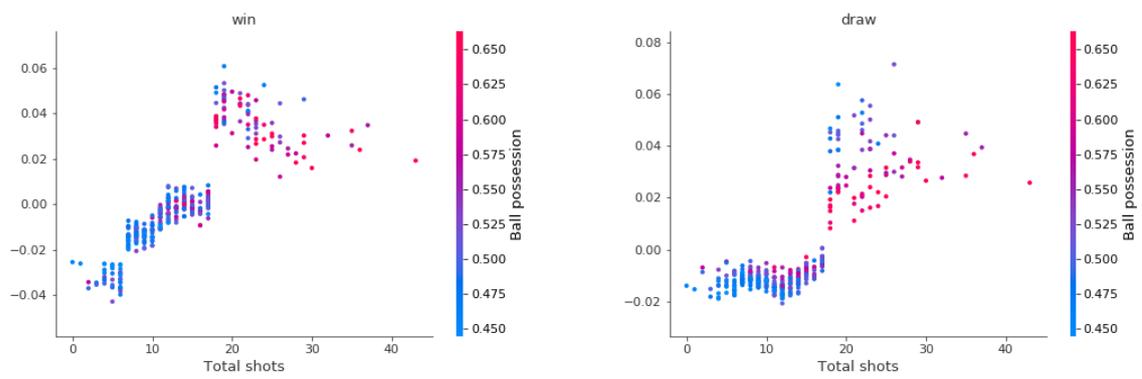


Figura 5.2: Valores SHAP por total de passes

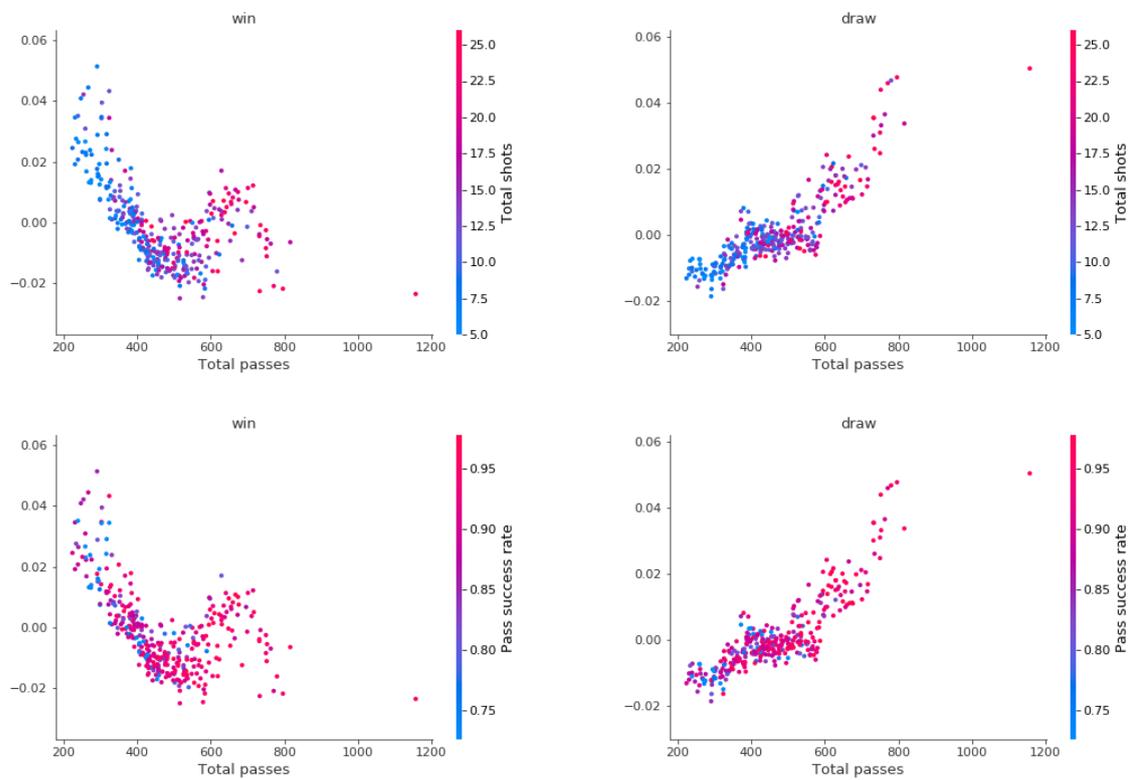


Figura 5.3: Valores SHAP por taxa de acerto de passes

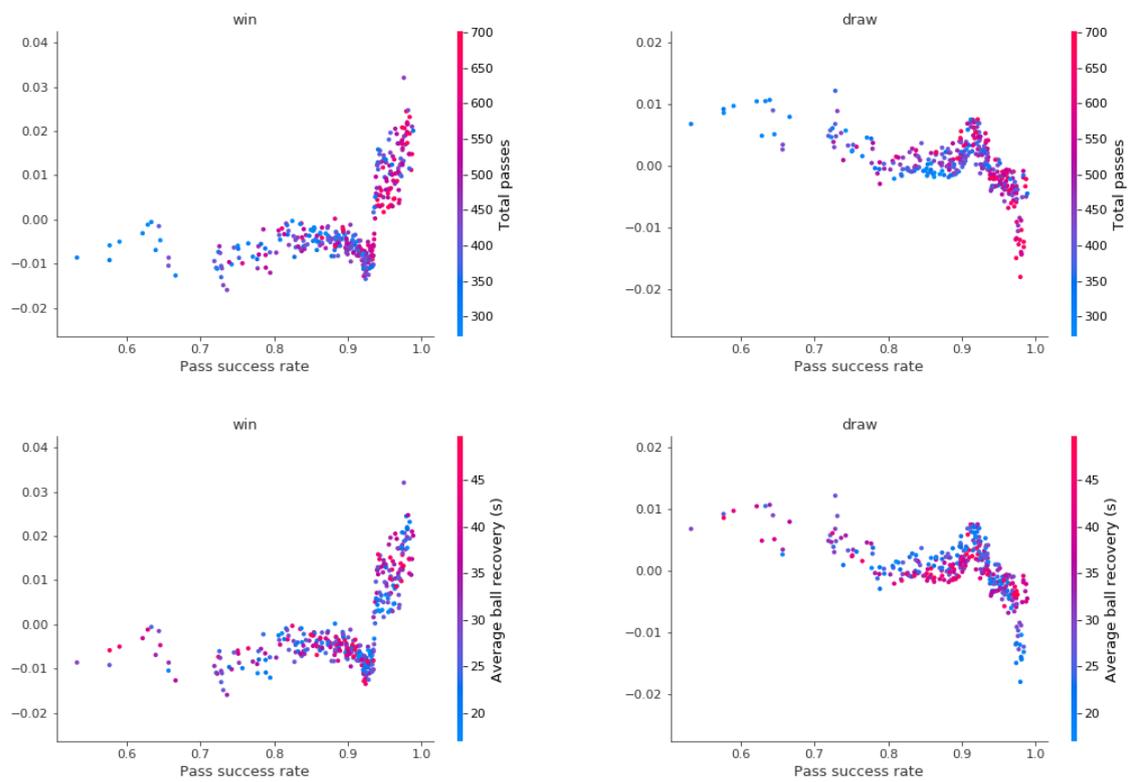


Figura 5.4: Valores SHAP por posse de bola

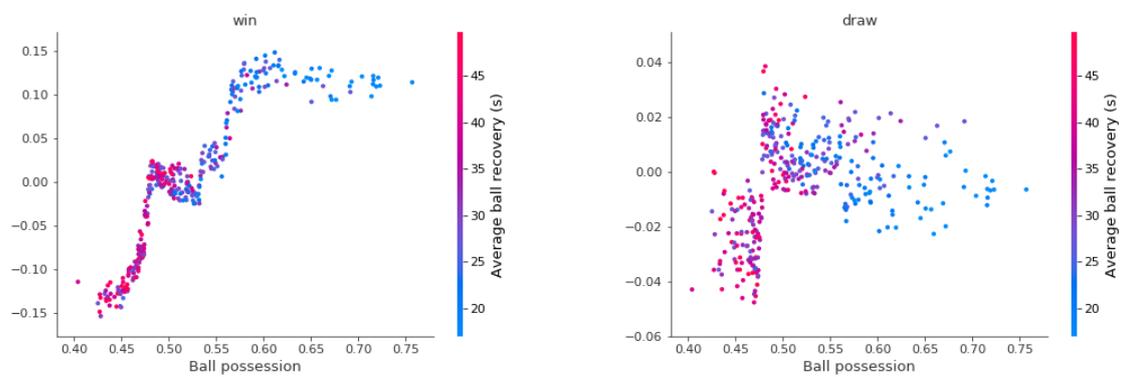
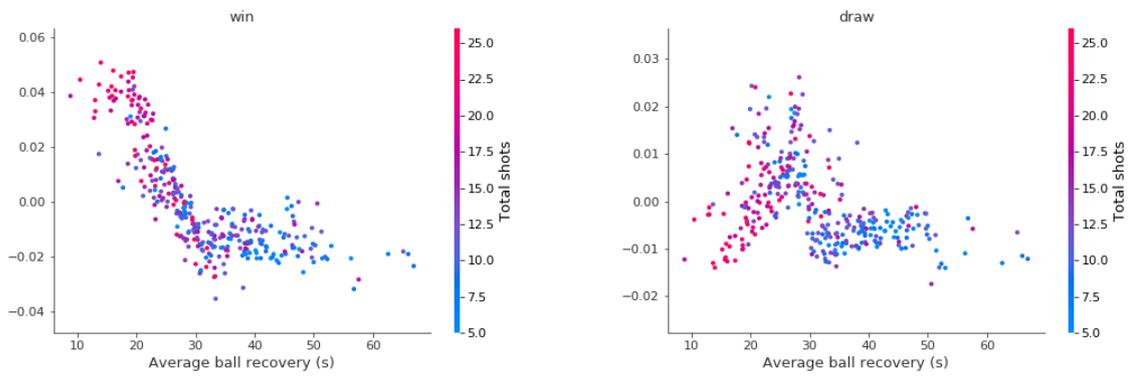


Figura 5.5: Valores SHAP por tempo de recuperação de bola



Por fim, a pressão defensiva, representada na figura 5.5 pelo tempo de recuperação de bola, apresenta uma tendência positiva para vitórias quando inferior a 30 segundos.

6 CONCLUSÕES E POSSIBILIDADES FUTURAS

Este trabalho tinha por objetivo identificar os principais fatores mais significativos por trás de uma vitória, empate ou derrota de uma equipe através de série de modelos preditivos baseados em algoritmos populares de aprendizagem de máquina. Para alcançar este objetivo, o trabalho utilizou uma base de dados que compila resultados de três campeonatos (Copa do Mundo, Campeonato de Futebol Feminino dos Estados Unidos e Campeonato de Futebol Feminino da Inglaterra), disponibilizada publicamente. Dez métricas foram extraídas para cada partida e equipe e estas métricas foram avaliadas, produzindo modelos baseados algoritmos dos tipos Modelo Linear Generalizado (regressão multinomial), Bagging (random forest e Extra-Trees) e Boosting (AdaBoost e XGBoost). Diversas métricas foram calculadas para cada modelo, que possibilitam avaliar a qualidade das previsões. Por fim, geraram-se visualizações baseadas em mecanismos de interpretabilidade, que tornaram possível a definição dos fatores mais importantes para o resultado de uma partida de futebol, desta forma alcançando-se o objetivo proposto.

A partir dos valores obtidos, emergiram padrões de variação de probabilidades condizentes com as análises existentes na literatura. As tendências mutuamente excludentes entre o total de passes trocados ao longo de uma partida e sua taxa de acerto refletem a dicotomia entre os principais estilos de jogo empregados e a, em princípio, ausência de uma relação de superioridade entre eles. Ao mesmo tempo, se apresenta uma clara vantagem na manutenção da posse de bola em um certo patamar e de uma quantidade mínima de tentativas de gol, além de um impacto positivo no uso de táticas de pressão alta imediata.

Por ser uma área de emergência muito recente, a inteligência artificial explicável apresenta uma vasta gama de aplicações a serem exploradas. Através da combinação de modelos complexos do tipo caixa-preta com alto poder preditivo e abstrações de alto nível, ela se torna um mecanismo poderoso para apresentar o conhecimento gerado por um modelo preditivo para quaisquer partes interessadas na interação entre as variáveis analisadas. Assim, este trabalho demonstra a utilidade da inteligência artificial explicável para a inspeção de análises preditivas, algo não encontrado na literatura durante o seu desenvolvimento. No aspecto prático, equipes de futebol que possuem uma massa de dados relevante podem se basear neste trabalho para o desenvolvimento de seus próprios modelos interpretativos voltados a tratar seus problemas específicos.

Diversas limitações estão presentes nesta dissertação de mestrado, em sua maioria

causadas pela imaturidade de um campo de pesquisa tão recente quanto a inteligência artificial explicável. A ausência de uma metodologia para avaliação da interpretabilidade dos modelos obtidos é a principal delas; assim, eles são gerados de uma maneira *ad hoc* e totalmente experimental. Assim, se faz necessária uma maior presença interdisciplinar no tema, em particular com estudos em psicologia e interação humano-computador. Ao mesmo tempo, a divulgação dos resultados, ainda que preliminares, do programa XAI representará um salto significativo nos recursos disponíveis.

O uso exclusivo de indicadores de desempenho agregados para partidas inteiras também é um fator limitante, por subaproveitar a riqueza de detalhes presentes nas bases de dados de eventos como a analisada neste estudo. Uma possível abordagem a ser estudada é a predição de sucesso de eventos individuais, como passes, gols ou pressão defensiva, o que pode ser facilitado pela geração de uma escala relativa para avaliar as habilidades individuais de cada jogador, similar ao empregado por franquias de jogos eletrônicos como *FIFA* and *Football Manager*.

REFERÊNCIAS

- ALCOCK, C. W. **Football: The Association Game**. New edition, revisited. London: George Bell & Sons, 1906. OCLC: 933863822. ISBN 978-0-217-80557-5.
- ALVAREZ-MELIS, D.; JAAKKOLA, T. S. On the Robustness of Interpretability Methods. **arXiv:1806.08049 [cs, stat]**, jun. 2018. ArXiv: 1806.08049. Available from Internet: <<http://arxiv.org/abs/1806.08049>>.
- APLEY, D. W. Visualizing the Effects of Predictor Variables in Black Box Supervised Learning Models. **arXiv:1612.08468 [stat]**, dec. 2016. ArXiv: 1612.08468. Available from Internet: <<http://arxiv.org/abs/1612.08468>>.
- ARABZAD, S. M. et al. Football Match Results Prediction Using Artificial Neural Networks; The Case of Iran Pro League. **Journal of Applied Research on Industrial Engineering**, v. 1, n. 3, p. 159–179, sep. 2014. ISSN 2538-5100. Available from Internet: <http://www.journal-aprie.com/article_43050.html>.
- BAIO, G.; BLANGIARDO, M. Bayesian hierarchical model for the prediction of football results. **Journal of Applied Statistics**, v. 37, n. 2, p. 253–264, 2010. Available from Internet: <<http://www.tandfonline.com/doi/abs/10.1080/02664760802684177>>.
- BOUDWAY, I. Soccer Is the World’s Most Popular Sport and Still Growing. **Bloomberg**, jun. 2018. Available from Internet: <<https://www.bloomberg.com/news/articles/2018-06-12/soccer-is-the-world-s-most-popular-sport-and-still-growing>>.
- BREIMAN, L. Bagging predictors. **Machine Learning**, v. 24, n. 2, p. 123–140, aug. 1996. ISSN 1573-0565. Available from Internet: <<https://doi.org/10.1007/BF00058655>>.
- BREIMAN, L. Random Forests. **Machine Learning**, v. 45, n. 1, p. 5–32, oct. 2001. ISSN 1573-0565. Available from Internet: <<https://doi.org/10.1023/A:1010933404324>>.
- BREIMAN, L. et al. **Classification and Regression Trees**. [S.l.]: Taylor & Francis, 1984. ISBN 978-0-412-04841-8.
- CALISKAN, A.; BRYSON, J. J.; NARAYANAN, A. Semantics derived automatically from language corpora contain human-like biases. **Science**, v. 356, n. 6334, p. 183–186, abr. 2017. ISSN 0036-8075, 1095-9203. ArXiv: 1608.07187. Available from Internet: <<http://arxiv.org/abs/1608.07187>>.
- CASTELLANO, J.; CASAMICHANA, D.; LAGO, C. The Use of Match Statistics that Discriminate Between Successful and Unsuccessful Soccer Teams. **Journal of Human Kinetics**, v. 31, p. 139–147, abr. 2012. ISSN 1640-5544. Available from Internet: <<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3588662/>>.
- 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 - KDD '16**. San Francisco, California, USA: ACM Press, 2016. p. 785–794. ISBN 978-1-4503-4232-2. Available from Internet: <<http://dl.acm.org/citation.cfm?doid=2939672.2939785>>.

CHENG, T. et al. A new model to forecast the results of matches based on hybrid neural networks in the soccer rating system. In: **Proceedings Fifth International Conference on Computational Intelligence and Multimedia Applications. ICCIMA 2003**. [S.l.: s.n.], 2003. p. 308–313.

CONSTANTINO, A. C.; FENTON, N. E. Solving the Problem of Inadequate Scoring Rules for Assessing Probabilistic Football Forecast Models. **Journal of Quantitative Analysis in Sports**, v. 8, n. 1, jan. 2012. ISSN 1559-0410. Available from Internet: <<https://www.degruyter.com/view/j/jqas.2012.8.issue-1/1559-0410.1418/1559-0410.1418.xml>>.

COX, D. R. The Regression Analysis of Binary Sequences. **Journal of the Royal Statistical Society. Series B (Methodological)**, v. 20, n. 2, p. 215–242, 1958. ISSN 0035-9246. Available from Internet: <<https://www.jstor.org/stable/2983890>>.

EPSTEIN, E. S. A Scoring System for Probability Forecasts of Ranked Categories. **Journal of Applied Meteorology**, v. 8, n. 6, p. 985–987, dec. 1969. ISSN 0021-8952. Available from Internet: <<http://journals.ametsoc.org/doi/abs/10.1175/1520-0450%281969%29008%3C0985%3AASSFPF%3E2.0.CO%3B2>>.

FIFA. **FIFA Disciplinary Code**. 2017. Available from Internet: <<https://resources.fifa.com/image/upload/fifa-disciplinary-code-500276.pdf?cloudid=koyeb3cvhxnwy9yz4aa6>>.

FIFA. **FIFA World Cup Russia**. 2018. Available from Internet: <<https://www.fifa.com/worldcup/>>.

FRANZINI, F. Futebol é "coisa para macho"?: Pequeno esboço para uma história das mulheres no país do futebol. **Revista Brasileira de História**, v. 25, n. 50, p. 315–328, dec. 2005. ISSN 0102-0188. Available from Internet: <http://www.scielo.br/scielo.php?script=sci_abstract&pid=S0102-01882005000200012&lng=en&nrm=iso&tlng=pt>.

FREUND, Y.; SCHAPIRE, R. E. A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. **Journal of Computer and System Sciences**, v. 55, n. 1, p. 119–139, aug. 1997. ISSN 0022-0000. Available from Internet: <<http://www.sciencedirect.com/science/article/pii/S002200009791504X>>.

FRIEDMAN, J. H. Greedy function approximation: A gradient boosting machine. **The Annals of Statistics**, v. 29, n. 5, p. 1189–1232, oct. 2001. ISSN 0090-5364, 2168-8966. Available from Internet: <<https://projecteuclid.org/euclid.aos/1013203451>>.

GEURTS, P.; ERNST, D.; WEHENKEL, L. Extremely randomized trees. **Machine Learning**, v. 63, n. 1, p. 3–42, abr. 2006. ISSN 1573-0565. Available from Internet: <<https://doi.org/10.1007/s10994-006-6226-1>>.

GLANVILLE, B. **The story of the World Cup**. London: Faber and Faber, 2005. OCLC: 62177631. ISBN 978-0-571-22944-4.

GOLDSTEIN, A. et al. Peeking Inside the Black Box: Visualizing Statistical Learning with Plots of Individual Conditional Expectation. **arXiv:1309.6392 [stat]**, sep. 2013. ArXiv: 1309.6392. Available from Internet: <<http://arxiv.org/abs/1309.6392>>.

GOODMAN, B.; FLAXMAN, S. European Union regulations on algorithmic decision-making and a "right to explanation". **AI Magazine**, v. 38, n. 3, p. 50, oct. 2017. ISSN 0738-4602, 0738-4602. ArXiv: 1606.08813. Available from Internet: <<http://arxiv.org/abs/1606.08813>>.

GORMAN, B. **A Kaggle Master Explains Gradient Boosting**. 2017. Available from Internet: <<http://blog.kaggle.com/2017/01/23/a-kaggle-master-explains-gradient-boosting/>>.

GUNNING, D. **Explainable Artificial Intelligence Research at DARPA**. DARPA: [s.n.], 2016. Available from Internet: <https://sites.nationalacademies.org/cs/groups/pgasite/documents/webpage/pga_184754.pdf>.

GUTERMAN, M. **O Futebol explica o Brasil: uma história da maior expressão popular do país**. [S.l.]: Editora Contexto, 2013. Google-Books-ID: 1s1nAwAAQBAJ. ISBN 978-85-7244-522-1.

HARROP, K.; NEVILL, A. Performance indicators that predict success in an English professional League One soccer team. **International Journal of Performance Analysis in Sport**, v. 14, n. 3, p. 907–920, dec. 2014. ISSN 2474-8668. Available from Internet: <<https://doi.org/10.1080/24748668.2014.11868767>>.

HEAD, T. et al. **scikit-optimize/scikit-optimize: v0.5.2**. Zenodo, 2018. Available from Internet: <<https://zenodo.org/record/1207017>>.

HOERL, A. E.; KENNARD, R. W. Ridge Regression: Biased Estimation for Nonorthogonal Problems. **Technometrics**, v. 12, n. 1, p. 55–67, feb. 1970. ISSN 0040-1706. Available from Internet: <<https://www.tandfonline.com/doi/abs/10.1080/00401706.1970.10488634>>.

HUANG, K. Y.; CHANG, W. L. A neural network method for prediction of 2006 World Cup Football Game. In: **The 2010 International Joint Conference on Neural Networks (IJCNN)**. [S.l.: s.n.], 2010. p. 1–8.

HUGHES, M.; FRANKS, I. Analysis of passing sequences, shots and goals in soccer. **Journal of Sports Sciences**, v. 23, n. 5, p. 509–514, 2005. ISSN 0264-0414. Available from Internet: <<http://dx.doi.org/10.1080/02640410410001716779>>.

HYAFIL, L.; RIVEST, R. L. Constructing optimal binary decision trees is NP-complete. **Information Processing Letters**, v. 5, n. 1, p. 15–17, may 1976. ISSN 0020-0190. Available from Internet: <<http://www.sciencedirect.com/science/article/pii/0020019076900958>>.

IFAB. **Laws of the Game**. 2018. Available from Internet: <<http://www.theifab.com/laws>>.

JAMES, G. et al. **An Introduction to Statistical Learning: With Applications in R**. [S.l.]: Springer Publishing Company, Incorporated, 2014. ISBN 978-1-4614-7137-0.

JOSEPH, A.; FENTON, N. E.; NEIL, M. Predicting football results using Bayesian nets and other machine learning techniques. **Knowledge-Based Systems**, v. 19, n. 7, p. 544–553, nov. 2006. ISSN 0950-7051. Available from Internet: <<http://www.sciencedirect.com/science/article/pii/S0950705106000724>>.

LAGO-BALLESTEROS, J.; LAGO-PEÑAS, C.; REY, E. The effect of playing tactics and situational variables on achieving score-box possessions in a professional soccer team. **Journal of Sports Sciences**, v. 30, n. 14, p. 1455–1461, 2012. ISSN 1466-447X.

LAGO-PEÑAS, C. et al. Home advantage in football: Examining the effect of scoring first on match outcome in the five major European leagues. **International Journal of Performance Analysis in Sport**, v. 16, n. 2, p. 411–421, aug. 2016. ISSN 2474-8668. Available from Internet: <<https://doi.org/10.1080/24748668.2016.11868897>>.

LAGO-PEÑAS, C.; LAGO-BALLESTEROS, J. Game Location and Team Quality Effects on Performance Profiles in Professional Soccer. **Journal of Sports Science & Medicine**, v. 10, n. 3, p. 465–471, sep. 2011. ISSN 1303-2968. Available from Internet: <<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3737821/>>.

LAGO-PEÑAS, C. et al. Game-Related Statistics that Discriminated Winning, Drawing and Losing Teams from the Spanish Soccer League. **Journal of Sports Science & Medicine**, v. 9, n. 2, p. 288–293, jun. 2010. ISSN 1303-2968. Available from Internet: <<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3761743/>>.

LAGO-PEÑAS, C. et al. The influence of a congested calendar on physical performance in elite soccer. **Journal of Strength and Conditioning Research**, v. 25, n. 8, p. 2111–2117, aug. 2011. ISSN 1533-4287.

LIPTON, Z. C. The Mythos of Model Interpretability. **arXiv:1606.03490 [cs, stat]**, jun. 2016. ArXiv: 1606.03490. Available from Internet: <<http://arxiv.org/abs/1606.03490>>.

LOU, Y.; CARUANA, R.; GEHRKE, J. Intelligible Models for Classification and Regression. In: **Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining**. New York, NY, USA: ACM, 2012. (KDD '12), p. 150–158. ISBN 978-1-4503-1462-6. Available from Internet: <<http://doi.acm.org/10.1145/2339530.2339556>>.

LOVATTO, D. L.; GALATTI, L. R. Pedagogia do Esporte e Jogos Esportivos Coletivos: das Teorias Gerais para a Iniciação Esportiva em Basquetebol. **Movimento e Percepção**, Vol. 8, Nº 11, 2007, v. 8, n. 11, nov. 2007. Available from Internet: <<http://ferramentas.unipinhal.edu.br/movimentoepercepcao/viewarticle.php?id=146>>.

LOWRY, S.; MACPHERSON, G. A blot on the profession. **British Medical Journal (Clinical research ed.)**, v. 296, n. 6623, p. 657–658, mar. 1988. ISSN 0267-0623. Available from Internet: <<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2545288/>>.

LUNDBERG, S.; LEE, S.-I. A Unified Approach to Interpreting Model Predictions. **arXiv:1705.07874 [cs, stat]**, may 2017. ArXiv: 1705.07874. Available from Internet: <<http://arxiv.org/abs/1705.07874>>.

MACKENZIE, R.; CUSHION, C. Performance analysis in football: A critical review and implications for future research. **Journal of Sports Sciences**, v. 31, n. 6, p. 639–676, mar. 2013. ISSN 0264-0414, 1466-447X. Available from Internet: <<http://www.tandfonline.com/doi/abs/10.1080/02640414.2012.746720>>.

MARCELINO, R.; SAMPAIO, J.; MESQUITA, I. Investigação centrada na análise do jogo: Da modelação estática à modelação dinâmica. **Revista Portuguesa de Ciências do Desporto**, v. 11, n. 1, p. 481–499, 2011.

MASON, L. et al. Boosting Algorithms As Gradient Descent. In: **Proceedings of the 12th International Conference on Neural Information Processing Systems**. Cambridge, MA, USA: MIT Press, 1999. (NIPS'99), p. 512–518. Event-place: Denver, CO. Available from Internet: <<http://dl.acm.org/citation.cfm?id=3009657.3009730>>.

MILLER, T. Explanation in Artificial Intelligence: Insights from the Social Sciences. **arXiv:1706.07269** [cs], jun. 2017. ArXiv: 1706.07269. Available from Internet: <<http://arxiv.org/abs/1706.07269>>.

MOURA, F. A.; MARTINS, L. E. B.; CUNHA, S. A. Analysis of football game-related statistics using multivariate techniques. **Journal of Sports Sciences**, v. 32, n. 20, p. 1881–1887, dec. 2014. ISSN 0264-0414, 1466-447X. Available from Internet: <<http://www.tandfonline.com/doi/abs/10.1080/02640414.2013.853130>>.

MURPHY, A. H. The ranked probability score and the probability score: A comparison. **Monthly Weather Review**, v. 98, n. 12, p. 917–924, dec. 1970. ISSN 0027-0644, 1520-0493.

NELDER, J. A.; WEDDERBURN, R. W. M. Generalized Linear Models. **Journal of the Royal Statistical Society. Series A (General)**, v. 135, n. 3, p. 370–384, 1972. ISSN 0035-9238. Available from Internet: <<https://www.jstor.org/stable/2344614>>.

NG, A. Y.; JORDAN, M. I. On Discriminative vs. Generative Classifiers: A comparison of logistic regression and naive Bayes. In: DIETTERICH, T. G.; BECKER, S.; GHAHRAMANI, Z. (Ed.). **Advances in Neural Information Processing Systems 14**. Vancouver, BC, Canada: MIT Press, 2002. p. 841–848. Available from Internet: <<http://papers.nips.cc/paper/2020-on-discriminative-vs-generative-classifiers-a-comparison-of-logistic-regression-and-naive-bayes.pdf>>.

OPITZ, D.; MACLIN, R. Popular Ensemble Methods: An Empirical Study. **Journal of Artificial Intelligence Research**, v. 11, p. 169–198, aug. 1999. ISSN 1076-9757. Available from Internet: <<https://jair.org/index.php/jair/article/view/10239>>.

PEDREGOSA, F. et al. Scikit-learn: Machine Learning in Python. **Journal of Machine Learning Research**, v. 12, p. 2825–2830, nov. 2011. ISSN 1532-4435. Available from Internet: <<http://dl.acm.org/citation.cfm?id=1953048.2078195>>.

PEDRESHI, D.; RUGGIERI, S.; TURINI, F. Discrimination-aware Data Mining. In: **Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining**. New York, NY, USA: ACM, 2008. (KDD '08), p. 560–568. ISBN 978-1-60558-193-4. Available from Internet: <<http://doi.acm.org/10.1145/1401890.1401959>>.

POLLARD, R.; REEP, C. Measuring the effectiveness of playing strategies at soccer. **Journal of the Royal Statistical Society Series D: The Statistician**, v. 46, n. 4, p. 541–550, 1997. ISSN 0039-0526.

QUINLAN, J. R. Induction of decision trees. **Machine Learning**, v. 1, n. 1, p. 81–106, mar. 1986. ISSN 1573-0565. Available from Internet: <<https://doi.org/10.1007/BF00116251>>.

QUINLAN, J. R. **C4.5: Programs for Machine Learning**. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1993. ISBN 978-1-55860-238-0.

REIN, R.; MEMMERT, D. Big data and tactical analysis in elite soccer: future challenges and opportunities for sports science. **SpringerPlus**, v. 5, n. 1, p. 1410, aug. 2016. ISSN 2193-1801. Available from Internet: <<http://springerplus.springeropen.com/articles/10.1186/s40064-016-3108-2>>.

RIBEIRO, M. T.; SINGH, S.; GUESTRIN, C. Model-Agnostic Interpretability of Machine Learning. **arXiv:1606.05386 [cs, stat]**, jun. 2016. ArXiv: 1606.05386. Available from Internet: <<http://arxiv.org/abs/1606.05386>>.

RIBEIRO, M. T.; SINGH, S.; GUESTRIN, C. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. **arXiv:1602.04938 [cs, stat]**, feb. 2016. ArXiv: 1602.04938. Available from Internet: <<http://arxiv.org/abs/1602.04938>>.

ROKACH, L. Ensemble-based classifiers. **Artificial Intelligence Review**, v. 33, n. 1, p. 1–39, feb. 2010. ISSN 1573-7462. Available from Internet: <<https://doi.org/10.1007/s10462-009-9124-7>>.

ROTSHTEIN, A. P.; POSNER, M.; RAKITYANSKAYA, A. B. Football Predictions Based on a Fuzzy Model with Genetic and Neural Tuning. **Cybernetics and Systems Analysis**, v. 41, n. 4, p. 619–630, jul. 2005. ISSN 1060-0396, 1573-8337. Available from Internet: <<http://link.springer.com/10.1007/s10559-005-0098-4>>.

RUE, H.; SALVESEN, O. Prediction and Retrospective Analysis of Soccer Matches in a League. **Journal of the Royal Statistical Society: Series D (The Statistician)**, v. 49, n. 3, p. 399–418, sep. 2001. ISSN 1467-9884. Available from Internet: <<https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/1467-9884.00243>>.

SARMENTO, H. et al. What Performance Analysts Need to Know About Research Trends in Association Football (2012–2016): A Systematic Review. **Sports Medicine**, v. 48, n. 4, p. 799–836, abr. 2018. ISSN 1179-2035. Available from Internet: <<https://doi.org/10.1007/s40279-017-0836-6>>.

SARMENTO, H. et al. Match analysis in football: a systematic review. **Journal of Sports Sciences**, v. 32, n. 20, p. 1831–1843, dec. 2014. ISSN 0264-0414. Available from Internet: <<http://shapeamerica.tandfonline.com/doi/abs/10.1080/02640414.2014.898852>>.

SHAPLEY, L. S. A Value for N-Person Games. In: KUHN, H. W.; TUCKER, A. W. (Ed.). **Contributions to the Theory of Games**. Princeton University Press, 1953, (Annals of Mathematics Studies, 28). p. 307–318. ISBN 978-0-691-07935-6. Available from Internet: <<http://www.jstor.org/stable/j.ctt1b9x1zv.24>>.

StatsBomb. **StatsBomb Open Data repository**. 2019. Available from Internet: <<https://github.com/statsbomb/open-data>>.

TENGA, A. et al. Effect of playing tactics on achieving score-box possessions in a random series of team possessions from Norwegian professional soccer matches. **Journal of Sports Sciences**, v. 28, n. 3, p. 245–255, feb. 2010. ISSN 1466-447X.

TENGA, A. et al. Effect of playing tactics on goal scoring in Norwegian professional soccer. **Journal of Sports Sciences**, v. 28, n. 3, p. 237–244, feb. 2010. ISSN 1466-447X.

TENGA, A.; RONGLAN, L. T.; BAHR, R. Measuring the effectiveness of offensive match-play in professional soccer. **European Journal of Sport Science**, v. 10, n. 4, p. 269–277, jul. 2010. ISSN 1746-1391. Available from Internet: <<https://doi.org/10.1080/17461390903515170>>.

TIBSHIRANI, R. Regression Shrinkage and Selection via the Lasso. **Journal of the Royal Statistical Society. Series B (Methodological)**, v. 58, n. 1, p. 267–288, 1996. ISSN 0035-9246. Available from Internet: <<https://www.jstor.org/stable/2346178>>.

TSAKONAS, A. et al. Soft Computing-Based Result Prediction of Football Games. p. 9, 2002.

VOGELBEIN, M.; NOPP, S.; HÖKELMANN, A. Defensive transition in soccer - are prompt possession regains a measure of success? A quantitative analysis of German Fußball-Bundesliga 2010/2011. **Journal of Sports Sciences**, v. 32, n. 11, p. 1076–1083, 2014. ISSN 1466-447X.

WILSON, J. Jürgen Klopp's Liverpool and the importance of 'gegenpressing'. **The Guardian**, oct. 2015. ISSN 0261-3077. Available from Internet: <<https://www.theguardian.com/football/blog/2015/oct/16/jurgen-klopp-liverpool-gegenpressing-borussia-dortmund>>.

ZHU, J. et al. Multi-class AdaBoost. **Statistics and Its Interface**, v. 2, n. 3, p. 349–360, 2009. ISSN 19387989, 19387997. Available from Internet: <<http://www.intlpress.com/site/pub/pages/journals/items/sii/content/vols/0002/0003/a008/>>.

ZOU, H.; HASTIE, T. Regularization and variable selection via the elastic net. **Journal of the Royal Statistical Society: Series B (Statistical Methodology)**, v. 67, n. 2, p. 301–320, abr. 2005. ISSN 1369-7412, 1467-9868. Available from Internet: <<http://doi.wiley.com/10.1111/j.1467-9868.2005.00503.x>>.

ŠTRUMBELJ, E.; KONONENKO, I. Explaining prediction models and individual predictions with feature contributions. **Knowledge and Information Systems**, v. 41, n. 3, p. 647–665, dec. 2014. ISSN 0219-1377, 0219-3116. Available from Internet: <<http://link.springer.com/10.1007/s10115-013-0679-x>>.