UNIVERSIDADE ESTADUAL DE CAMPINAS
Faculdade de Tecnologia

Leandro Stival

Análise preditiva de fatores que influenciam na entrada na zona de finalização em partidas de futebol

Limeira
2022

Leandro Stival

Análise preditiva de fatores que influenciam na entrada na zona de finalização em partidas de futebol

Dissertação apresentada à Faculdade de Tecnologia da Universidade Estadual de Campinas como parte dos requisitos para a obtenção do título de Mestre em Tecnologia, na área de Sistemas de Informação e Comunicação.

Orientador: Prof. Dr. Ulisses Martins Dias

Este trabalho corresponde à versão final da Dissertação defendida por Leandro Stival e orientada pelo Prof. Dr. Ulisses Martins Dias.

Limeira
2022

Ficha catalográfica
Universidade Estadual de Campinas
Biblioteca da Faculdade de Tecnologia
Felipe de Souza Bueno - CRB 8/8577

St59a

Stival, Leandro, 1997-
Análise preditiva de fatores que influenciam na entrada na zona de finalização em partidas de futebol / Leandro Stival. – Limeira, SP : [s.n.], 2022.

Orientador: Ulisses Martins Dias.
Dissertação (mestrado) – Universidade Estadual de Campinas, Faculdade de Tecnologia.

1. Aprendizado de máquina. 2. Análise preditiva. 3. Ritmo visual. I. Dias, Ulisses Martins, 1983-. II. Universidade Estadual de Campinas. Faculdade de Tecnologia. III. Título.

Informações para Biblioteca Digital

Título em outro idioma: Predictive analysis of factors that influence entry into the end zone in soccer matches
Palavras-chave em inglês:
Machine learning
Predictve analytics
Visual rhythm
Área de concentração: Sistemas de Informação e Comunicação
Titulação: Mestre em Tecnologia
Banca examinadora:
Ulisses Martins Dias [Orientador]
Daniele Cristina Uchoa Maia Rodrigues
Guilherme Palermo Coelho
Data de defesa: 01-02-2022
Programa de Pós-Graduação: Tecnologia

Identificação e informações acadêmicas do(a) aluno(a)
- ORCID do autor: https://orcid.org/0000-0002-3379-6813
- Currículo Lattes do autor: http://lattes.cnpq.br/5747144193762495

FOLHA DE APROVAÇÃO

Abaixo se apresentam os membros da comissão julgadora da sessão pública de defesa de dissertação para o Título de Mestre em Tecnologia na área de concentração Sistemas de Informação e Comunicação, a que se submeteu o aluno Leandro Stival, em 01 de fevereiro de 2022 na Faculdade de Tecnologia – FT/UNICAMP, em Limeira/SP.

Prof. Dr. Ulisses Martins Dias
Presidente da Comissão Julgadora

Profa. Dra. Daniele Cristina Uchoa Maia Rodrigues
PUC/Campinas

Prof. Dr. Guilherme Palermo Coelho
FT/Unicamp

Ata da defesa, assinada pelos membros da Comissão Examinadora, encontra-se no SIGA/Sistema de Fluxo de Dissertação/Tese e na Secretaria de Pós Graduação da Faculdade de Tecnologia.

Agradecimentos

Inicialmente, gostaria de agradecer a Deus e a meus pais: Maria Inalva e José Donizeti, que sempre apoiaram as minhas decisões e forneceram tudo o que eu precisei, e o que eu nem sabia que precisava.

A minha amada companheira Isabela Rafael, pelo apoio incondicional nos momentos em que parecia que nada ia dar certo e pela excelente 'consultoria gramatical' prestada. A minha irmã Priscila, sendo uma ótima amiga que sempre me incentivou e esteve presente durante esse trabalho.

Ao meu orientador Prof. Dr. Ulisses Dias, que se tornou um amigo para mim, que me acolheu na Unicamp e, além de me orientar, me ensinou muito sobre a vida acadêmica, e não desistiu desse trabalho mesmo quando ele parecia sem saída.

Aos colegas que já podem ser considerado amigos que conheci durante essa jornada: Dr. Allan Pinto, que me ajudou muito durante o trabalho; Prof. Dr. Ricardo Torres, que me proporcionou a possibilidade de conhecer pessoas incríveis; Dr. Felipe (Sansão), que me ajudou muito na reta final do trabalho.

E por último queria agradecer a mim, por acreditar em mim e pelo empenho aplicado para a realização desse trabalho, sendo ele meu novo xodó.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001 e agradeço o apoio financeiro fornecido pelos processos nº #2019/17729-0, #2019/22262-3, #2019/16253-1, e #2021/00050-4), Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP).

Resumo

Futebol é um dos esportes mais populares do mundo, sendo essa popularidade refletida nas movimentações financeiras das competições profissionais, que se tornam cada vez mais competitivas a ponto de qualquer vantagem tática influenciar nos resultados. Este trabalho pretende investigar se, observando os primeiros segundos da tomada de posse de bola por uma equipe, seria possível identificar se aquela posse de bola resultaria em uma jogada de perigo na zona de ataque próxima à grande área do oponente, dado que a maneira de vencer uma partida implica em chegar próximo ao gol adversário com a posse de bola. Para responder a essa questão, utilizamos métricas extraídas da atuação dos jogadores e de suas interações em campo, como, por exemplo: posição dos jogadores de ambas as equipes em função do tempo e métricas de redes (centralidade, entropia, vulnerabilidade da rede). Esses dados são tratados modelando o posicionamento dos jogadores como uma série temporal de grafos em que, a cada instante de tempo, o posicionamento dos jogadores gera uma rede. As métricas dessas redes, quando convertidas em imagens de ritmo visual, permitem a extração de características por meio de arquiteturas de redes convolucionais profundas, o que aproxima o nosso estudo da área de aprendizado profundo (do inglês, *deep learning*), agregando possibilidades como o uso de técnicas de transferência de aprendizado (do inglês, *transfer learning*) junto de um refinamento (do inglês, *fine-tune*). A validação foi feita com a arquitetura EfficientNet B0, sendo utilizada para a transferência de aprendizado a partir de uma rede treinada com dados do ImageNet e com seus pesos ajustados com nossos dados. Dessa forma, características de imagens foram obtidas e aplicadas para o treinamento de uma Rede Neural Profunda (do inglês, *Deep Neural Network*) para a classificação das amostras. A técnica de *grid-search* com validação cruzada foi utilizada para otimizar os parâmetros da rede. Essa modelagem pretendem determinar se a posse de bola irá produzir uma jogada ofensiva no campo próximo ao gol do oponente. Além disso, validar quais métricas mais influenciam o resultado do modelo, permitindo assim identificar as mais determinantes para se chegar com a posse de bola na zona de defesa do oponente. Os resultados indicam que as chegadas próximas à grande área adversária podem ser preditas observando somente os primeiros 5 segundos de posse de bola, e que as principais métricas utilizadas pelos modelos encontram embasamento na literatura de análise de partidas de futebol.

Abstract

The focus of this work is investigated if watching the first seconds of a ball possession interval (BPI) of a team is possible to determine whether the ball possession will arrive in the last fourth of the soccer field, thereafter creating a goal chance situation. For the answer, this question was extracted metrics of player's interactions like the player's positions of both teams in the function of time and the graph metrics (centrality, entropy, vulnerably). The player's positions were modeled in a temporal series of graphs, where each instant of time is represented as a network. These metrics when converted to images of visual rhythm allow the feature extraction through convolutional networks, which allows using of transfer learning techniques with fine-tuning. EfficientNet B0 architecture was used for the transfer of learning with fine-tuning, this network was trained to begin the ImageNet dataset, and your weights adjusted with your data. In this way, the features were obtained and applied to a Deep Neural Network (DNN) that is responsible for classifying the ball possessions. The grid-search technic with cross-validate was utilized to optimize the parameters of the DNN. This modeling intends to determine if the ball possession will produce an offensive play nearest the adversary penalty area. Furthermore, we intend valid which are the metrics that influence the classifier results, thus allowing us to identify the most important metric for the ball possession arriving in the final quarter of the pitch. The results indicate that the arrivals near the opponent box can be predicted only by looking at the first 5 seconds of ball possession and that the main metrics utilized by the models have a foundation on the literature of soccer match analyzes.

Lista de Figuras

Lista de Tabelas

Sumário

Capítulo 1

Introdução

Futebol no Brasil é mais que um esporte, sendo rotulado como uma "paixão nacional". Além de ser apreciado por muitos brasileiros, movimenta um valor de R$52,9 bilhões anualmente no Brasil, representando 0,72% do PIB (CBF, 2019). À medida que esse mercado fica mais expressivo em movimentação financeira, as competições tendem a apresentar uma maior qualidade, aumentando a importância de qualquer vantagem tática.

O que determina o vencedor de uma partida de futebol é a quantidade de gols marcados por cada equipe e, considerando que para um gol ser marcado se faz necessária uma finalização em direção ao gol, as equipes tendem a buscar um posicionando ofensivo baseado na movimentação da bola no campo de defesa do oponente.

Segundo o estudo de Wright et al. (2011), a maioria das finalizações ocorre dentro ou próxima da grande área, sendo essa a região de campo com a maior taxa de conversão dos chutes. Sendo assim, chegar a essa região é um passo importante para obter finalizações. Além disso, o estudo de Ruiz-Ruiz et al. (2013) também mostra a importância dessa região do campo, apresentando que equipes que entram mais na área adversária têm mais chances de vencer partidas.

Sabendo que a chegada próxima à área é importante, o próximo passo é entender como isso acontece. Essa pergunta tem sido objeto de estudo de diversos autores e algumas respostas foram encontradas. Marchiori e Vecchi (2020) analisaram a eficiência das equipes momentos antes de o gol ser realizado, demostrando a importância do controle de determinadas posições dentro de campo e do comportamento da equipe na finalização da jogada. Wright et al. (2011) concluíram que jogadas consideradas "rápidas" (com 4 passes ou menos) apresentam melhor taxa de conversão das finalizações em gols, indicando que a

quantidade de passes apresenta influência no resultado da jogada. Outra conclusão foi que a quantidade de atacantes presentes na grande área melhora os resultados das finalizações, identificando os fatores que mais influenciam as finalizações bem sucedidas: o local do chute, a posição do goleiro e o tipo do chute (chute direto, cabeceio ou com o corpo).

Outro ponto importante, segundo a literatura, é a movimentação dos jogadores que se altera no decorrer do jogo e a situação da partida, como mostrado pelo estudo de Carling et al. (2012). Moura et al. (2011) observaram que as equipes tendem a ficar mais compactas (jogadores mais próximos) quando estão se defendendo, e que quando esse processo não ocorre a equipe tende a sofrer mais finalizações.

Apesar de a movimentação isolada de jogadores fazer diferença durante a partida em jogadas individuais, este trabalho focará no comportamento coletivo das equipes, tanto as que estão atacando quanto as que estão defendendo. Análises do comportamento das equipes também podem ser vistas no trabalho de Baboota e Kaur (2019) com o objetivo de prever o resultado das partidas de futebol e definir quais as métricas mais importantes para o resultado. Frencken, Lemmink et al. (2011) utilizam o centroide da equipe em campo (ponto central considerando a posição de todos os jogadores em campo da equipe) para identificar o ritmo temporal da movimentação dos times.

Além de jogadores e sua movimentação no campo, a posse de bola também foi abordada em pesquisas, como em Barreira et al. (2014), em que os autores apresentam que a região do campo e a forma em que a posse de bola é obtida influenciam no perigo de gol que a jogada pode ter. Esse fator é também estudado por Tenga et al. (2010), que mostram que as tendências de contra-ataques são mais efetivas contra defesas não tão organizadas taticamente.

Contra-ataques também são abordados por Moura et al. (2007), concluindo que retomadas de bola na zona de ataque com poucos passes apresentam mais perigo de gol do que retomadas de bola no campo defensivo seguidas de muitos passes.

Outra forma de visualizar e analisar uma partida de futebol é por meio das medidas de redes complexas (grafos), em que os jogadores em campo são geralmente representados por vértices e as ligações entre eles (geralmente possibilidade de passe) por arestas.

Na literatura, é possível encontrar trabalhos que buscam apresentar as métricas obtidas dos grafos de formas mais visuais. Rodrigues et al. (2019) desenvolveram um *framework* para análise de ritmo visual com séries temporais das métricas de redes de uma partida, e Frencken, Lemmink et al. (2011) com o uso de centroides. Clemente, Sarmento e Aquino

(2020) utilizaram a centralidade dos jogadores para determinar qual equipe venceria uma partida, e Malta e Travassos (2014) utilizaram a centralidade e a proximidades dos vértices (jogadores) para identificar a relação entre essas métricas e a chance de ocorrer gol.

1.1 Objetivo

Utilizando a literatura como embasamento e agrupando as métricas de jogadores, movimentação em campo e métrica de redes, levantamos as seguintes hipóteses de pesquisa:

- Um modelo baseado nos primeiros 5 segundos de posse de bola de uma equipe pode prever se, ao final dessa posse, a bola chegará no quarto final do campo?
- Existe um conjunto reduzido de características mais importante para prever se, ao final de uma posse, a bola chegará no quarto final do campo?

Nossos resultados fornecem embasamento para discutir essas questões. Modelamos o problema de modo que pertença à área de Aprendizado Supervisionado, dentro do campo de Aprendizado de Máquina que, por sua vez, se insere no contexto de Inteligência Artificial.

Obter um modelo com bom desempenho para prever o resultado de uma posse de bola é o nosso primeiro objetivo específico, e servirá de base para a análise das características que mais impactam o modelo.

1.2 Materiais e Métodos

Utilizaremos uma base de dados contendo 10 jogos completos. Os dados dos jogos possuem a posição de todos os jogadores em blocos de 30 *frames* por segundo. Além disso, uma série de eventos estão rotulados nos dados: faltas, gols, saídas pelas laterais e outros. Esses eventos foram utilizados para identificar a qual equipe pertencia a posse de bola, analisando a qual equipe pertencia o jogador do último evento rotulado.

A seguir, resumimos a metodologia feita a partir desses jogos sem fornecer detalhes e, no decorrer do texto, forneceremos as definições correspondentes:

1. Cada jogo foi particionado em janelas de posse de bola: intervalos maximais de tempo em que um time mantém a posse de bola, descartando os menores (menos de 15 segundos de duração).

2. Um período de alguns segundos no início de cada janela de posse de bola foi fixado para obter características.

3. Para cada *frame*, geramos um grafo em que os vértices são os jogadores e as arestas representam possibilidades de passes. A janela de posse de bola passa a ser representada como uma série temporal de grafos.

4. Em cada grafo da série, extraímos medidas comuns na área de redes complexas: centralidade, entropia, e outros.

5. As medidas são codificadas em uma imagem utilizando ritmo visual.

6. Características são obtidas a partir da imagem utilizando uma rede convolucional profunda pré-treinada e com seus pesos ajustados à base de dados. Essa rede atua como um extrator automático de características.

7. Classificadores foram treinados para prever se a posse de bola resultaria em uma jogada ofensiva que alcançou o quarto final do campo.

8. Um estudo da rede e das características mais utilizadas por ela permite analisar com que frequência cada característica foi usada, o que fornece um indício da importância de cada uma delas.

A Figura 1.1 apresenta a metodologia, numerando as etapas conforme os tópicos acima.

1.3 Organização do Texto

No Capítulo 2, são apresentados os conceitos sobre os quais este trabalho foi construído, explicando brevemente os conceitos: grafos, redes complexas, aprendizado de máquina, redes neurais convolucionais, descida do gradiente, acurácia balanceada, funcionamento de um teste de significância e explicabilidade de um classificador.

No Capítulo 3, são apresentados os trabalhos relacionados ao futebol e quais atributos estão sendo utilizadas para treinar a rede junto do seu embasamento na literatura.

No Capítulo 4, o desenvolvimento do trabalho é descrito demonstrando como foi realizada a seleção dos intervalos de posse de bola, quais métricas estão sendo extraídas e fornecidas, como a construção das imagens de ritmo visual é feita, como foi realizado o processo de transferência

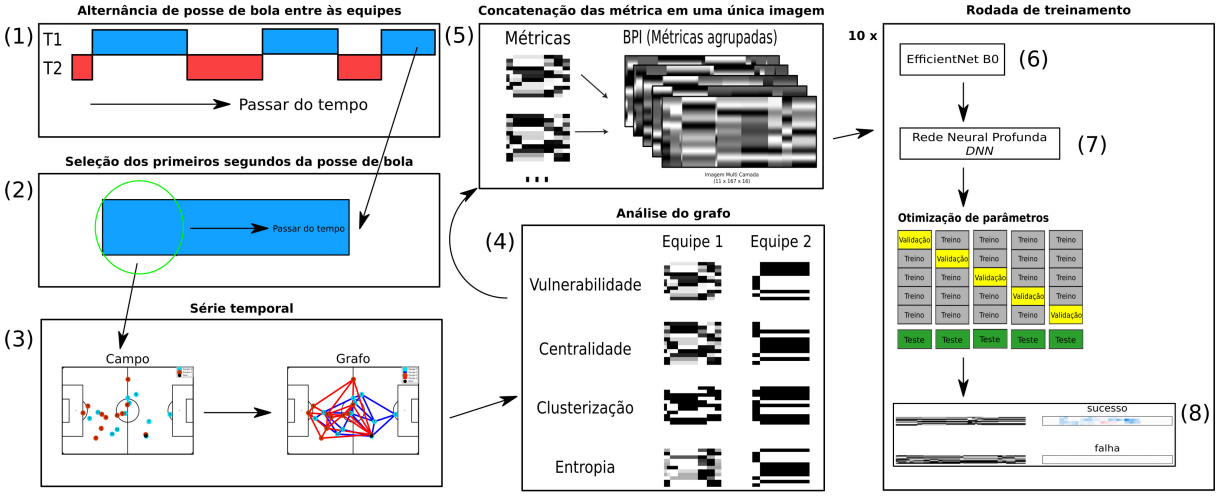


Figura 1.1: Representa o fluxo da metodologia proposta neste trabalho: (1) separação das posses de bola da partida; (2) seleção dos primeiros segundos dessa posse; (3) construção dos grafos a partir da posição dos jogadores e da bola; (4) extração dos atributos do grafo de cada equipe e construção do ritmo visual; (5) concatenação das imagens de ritmo visual em uma só imagem; (6) transferência de aprendizado utilizando EfficientNet B0; (7) classificação por meio de uma DNN; (8) identificação dos principais atributos utilizados para a classificação da rede.

de aprendizado e treinamento dos pesos originais da rede CNN e quais as técnicas para avaliar a importância das métricas.

No Capítulo 5, são apresentados os resultados das classificações, teste de hipótese e quais foram os principais atributos para a rede. O capítulo também discute os resultados obtidos com o que a literatura apresenta.

O Capítulo 6 apresenta o que foi possível concluir com os resultados atuais e quais serão os próximos passos neste tema de pesquisa.

Capítulo 2

Conceitos Básicos

Este capítulo introduz conceitos que irão ajudar no entendimento do trabalho, apresentando o funcionamento das técnicas utilizadas. Primeiramente, serão apresentados os grafos e as redes complexas, seguidos das métricas dentro da área de redes complexas utilizadas. A segunda parte deste capítulo foca no aprendizado de máquina, de modo particular, no aprendizado de máquina supervisionado. O capítulo também apresenta a métrica "acurácia balanceada" e o teste de significância estatística de Wilcoxon, utilizados para caracterizar a qualidade dos modelos. Por fim, são apresentadas formas de realizar a explicabilidade das predições.

2.1 Grafos e Redes Complexas

Um Grafo $G = (V, A)$ é composto por um conjunto não vazio de vértices $V = \{v_1, v_2, \dots, v_n\}$ e um conjunto de arestas $A = \{a_1, a_2, \dots, a_m\}$, onde cada aresta é um par não ordenado de vértices. Dependendo da aplicação, as arestas podem possuir direção, em que um dos vértices será chamado de origem e o outro vértice será chamado de destino.

Neste trabalho, não houve a necessidade de incluir direção nas arestas, dado que para o contexto de partidas de futebol modelado neste trabalho foi decidido fazer com que as arestas (v_i, v_j) indicassem tanto uma ligação de v_i para v_j, quanto de v_j para v_i, para $1 \leq i, j \leq n$. Na arquitetura de grafos utilizada neste trabalho, vértices representam jogadores e arestas representam a possibilidade de passe entre dois jogadores. Assim, uma aresta demonstra que há uma possibilidade de o jogador A realizar um passe para o jogador B, o que implica que há a mesma possibilidade de o jogador B realizar um passe para o jogador A.

Se (v_i, v_j) é uma aresta no grafo, dizemos que a aresta incide tanto em v_i quanto em v_j. Além disso, dizemos que v_i e v_j são vizinhos, sendo esta definição simétrica, onde v_i é vizinho de v_j se, e somente se, v_j é vizinho de v_i. Alguns termos de grafos serão importantes para este trabalho e uma nomenclatura é, a seguir, estabelecida.

Definição 2.1. *O **grau** do vértice v_i, denotado por $deg(v_i)$, é o número de arestas que incidem nele.*

Definição 2.2. *O **caminho** de um vértice v_i a um vértice v_j é uma sequência de vértices em que, para cada vértice, do primeiro ao penúltimo, há uma aresta ligando esse vértice ao próximo na sequência. O comprimento de um caminho é o número de arestas presentes nele.*

Definição 2.3. *A **distância** entre dois vértices (v_i, v_j), denominada $d(v_i, v_j)$, é o comprimento do menor caminho entre v_i e v_j.*

A quantidade de menores caminhos entre dois vértices v_i e v_j é uma métrica importante, denominada $\sigma(v_i, v_j)$. Em alguns casos, é importante considerar apenas os caminhos que passam por algum vértice v_k, então $\sigma(v_i, v_j|v_k)$ identificará a quantidade de menores caminhos entre v_i e v_k que passam por v_k.

A Figura 2.1 apresenta um grafo com cinco vértices: A, B, C, D e E; e seis arestas não direcionadas: (A, B), (A, C), (A, E), (B, E), (C, D) e (C, E). Os graus dos vértices são: $deg(A) = 3$, $deg(B) = 2$, $deg(C) = 3$, $deg(D) = 1$, $deg(E) = 3$.

Possíveis caminhos entre B e D seriam: (B, E, A, C, D), (B, E, C, D), (B, E, C, A, E, C, D), e outros. A distância entre os vértices B e D é igual a 3, dado este ser o número de arestas dos dois caminhos de menor tamanho: (B, A, C, D) e (B, E, C, D). Além disso, $\sigma(B, D) = 2$ por existirem dois caminhos cujos tamanhos igualam a distância. Por fim, $\sigma(B, D|A) = 1$, dado que apenas o caminho (B, A, C, D) passa por A.

Existe uma grande interseção entre a teoria dos grafos e o que se convencionou chamar de redes complexas. Entretanto, algumas particularidades destas últimas geraram a subdivisão das áreas. De especial importância para este trabalho, existe o fato de redes complexas serem compostas por representações dinâmicas que evoluem em função do tempo, o que decorre diretamente do fato de serem geralmente derivadas da análise de dados empíricos (ALBERT; BARABÁSI, 2002).

Redes complexas são representações de grafos que apresentam propriedades topográficas particulares, como a necessidade das arestas representarem algum tipo de relação entre os

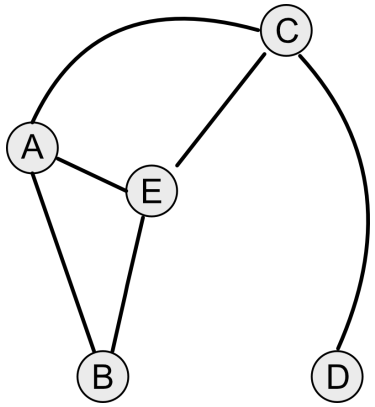


Figura 2.1: Neste exemplo de um grafo, temos $deg(A) = deg(C) = deg(E) = 3$, enquanto que $deg(B) = 2$ e $deg(D) = 1$. Existem dois menores caminhos entre B e D, notadamente (B, A, C, D) e (B, E, C, D), gerando: $\sigma(B, D) = 2$ e $\sigma(B, D|A) = 1$

vértices (METZ et al., 2007). Com base nessa modelagem, as seguintes métricas foram extraídas das redes criadas nesse trabalho:

Definição 2.4. *A **Centralidade de Intermediação** de um vértice (do inglês, Betweenness Centrality) é a proporção de menores caminhos que utilizam esse vértice (BRANDES, 2001), como mostra a Equação 2.1.*

$$c_B(v) = \sum_{s,t \in V} \frac{\sigma(s,t|v)}{\sigma(s,t)}, \text{ onde: } \begin{cases} \sigma(s,t) = 1, & Se\ s = t, \\ \sigma(s,t|v) = 0, & Se\ v \in s,t \end{cases} \tag{2.1}$$

É comum normalizar a centralidade de intermediação, como mostra a Equação 2.2, em que n é o número de vértices.

$$c_{Bn}(v) = \frac{c_B(v)}{(n-1)(n-2)} \tag{2.2}$$

A centralidade de intermediação de um vértice representa o quão os menores caminhos do grafo são dependentes dele. Um valor alto da métrica representa que os menores caminhos dependem desse vértice para conectar o grafo. C_{Bn} representa o valor de centralidade normalizado, ou seja, o valor da métrica no contexto do tamanho da rede.

Exemplo 2.1. *A centralidade de intermediação do vértice A no grafo da Figura 2.1 é dada pela quantidade de menores caminhos que passam por A quando a origem e o destino do caminho difere de A. Como o grafo não é direcionado, temos que $\sigma(s,t) = \sigma(t,s)$ e que $\sigma(s,t|A) = \sigma(t,s|A)$, o que gera: $c_B(A) = \sum_{s,t \in V} \frac{\sigma(s,t|v)}{\sigma(s,t)} = 2\left[\frac{\sigma(B,C|A)}{\sigma(B,C)} + \frac{\sigma(B,D|A)}{\sigma(B,D)} + \frac{\sigma(B,E|A)}{\sigma(B,E)} + \frac{\sigma(C,D|A)}{\sigma(C,D)} + \frac{\sigma(C,E|A)}{\sigma(C,E)} + \frac{\sigma(D,E|A)}{\sigma(D,E)}\right]$.*

Como A está em um menor caminho apenas quando ligamos B a C, e B a D, temos:

- $\sigma(B, C) = 2$, *pelos caminhos* (B, E, C) *e* (B, A, C), *o que implica em* $\sigma(B, C|A) = 1$.
- $\sigma(B, D) = 2$, *pelos caminhos* (B, E, C, D) *e* (B, A, C, D), *o que implica em* $\sigma(B, D|A) = 1$.

Assim, $c_B(A) = 2\left[\frac{\sigma(B,C|A)}{\sigma(B,C)} + \frac{\sigma(B,D|A)}{\sigma(B,D)}\right] = 2\left[\frac{1}{2} + \frac{1}{2}\right] = 2$. *A centralidade de intermediação normalizada é obtida com* $c_{Bn}(A) = \frac{c_B(A)}{(n-1)(n-2)} = \frac{2}{(5-1)(5-2)} = \frac{2}{12} = 0.167$.

Definição 2.5. *A* ***Excentricidade*** *de um vértice (do inglês, Eccentricity) é a maior distância entre o vértice e algum outro, como mostra a Equação 2.3.*

$$e(v) = max(d(v, v_j)), \forall v_j \in V \tag{2.3}$$

A excentricidade indica o quão isolado um vértice pode estar dos demais na rede, métrica importante para identificar se o vértice está localizado mais ao centro da rede (quando a excentricidade é menor).

Exemplo 2.2. *A excentricidade do vértice A no grafo da Figura 2.1 é dada pela maior distância de A até qualquer outro vértice, sendo* $d(A, B) = 1$, $d(A, C) = 1$, $d(A, D) = 2$ *e* $d(A, E) = 1$, *o que gera* $e(A) = 2$.

Definição 2.6. *A* ***Eficiência Global*** *de um vértice v (do inglês, Global Efficiency) mede a capacidade de transmissão de informação do grafo G através de v, considerando o inverso da distância (d) entre v e os demais vértices j (LATORA; MARCHIORI, 2001), como mostra a Equação 2.4, sendo n o número de vértices em G.*

$$E(v) = \frac{\sum_{j \neq v}^{j \in G} \frac{1}{d(v,j)}}{(n-1)} \tag{2.4}$$

A eficiência global permite identificar se os vértices da rede são próximos uns aos outros. Nesse contexto, o termo próximo indica que possuem distância pequena entre si, considerando que quanto maior a eficiência de um vértice menor a sua distância em relação aos demais.

Exemplo 2.3. *A eficiência global do vértice A no grafo da Figura 2.1 é obtida primeiramente com o somatório:* $\frac{1}{d(A,B)} + \frac{1}{d(A,C)} + \frac{1}{d(A,D)} + \frac{1}{d(A,E)} = \frac{1}{1} + \frac{1}{1} + \frac{1}{2} + \frac{1}{1} = 3.5$. *Essa soma é dividida pelo total de vértices subtraído de 1:* $\frac{3.5}{(n-1)} = \frac{3.5}{(4)} = 0.88$ *de eficiência global para A.*

Definição 2.7. *A* ***Eficiência Local*** *de um vértice v (do inglês, Local Efficiency) é calculada para um subgrafo* $\hat{G}_v$ *que contém os vizinhos de v e suas respectivas arestas, como mostra a Equação 2.5.*

$$E_{loc}(v) = \frac{1}{deg(v)} \sum_{v_j \in \hat{G}_v} E(v_j) \tag{2.5}$$

A eficiência local indica como os vizinhos de um vértice são afetados com a remoção dele, por meio do aumento da distância para seus vizinhos se conectarem novamente com toda a rede, sendo que valores altos significam que a ausência de v torna os caminhos mais longos para seus vizinhos (LATORA; MARCHIORI, 2001).

Exemplo 2.4. *Para calcular a eficiência local de A no grafo da Figura 2.1, é necessária a eficiência global em $\hat{G}_v$ de todos os vértices vizinhos, sendo elas $B = 0.71$, $C = 0.88$ e $E = 0.88$. Esse valores aplicados na Equação2.5 obtêm: $E_{loc}(A) = 0.33 \times (0.71 + 0.88 + 0.88) = 0.82$ de eficiência local para A.*

Definição 2.8. *A **Vulnerabilidade** de um vértice v (do inglês, Vulnerability) usa a eficiência global de v e a eficiência local de v, como mostra a Equação 2.6.*

$$V(v) = 1 - \frac{E_{loc}(v)}{E(v)} \tag{2.6}$$

A vulnerabilidade indica a relação entre a eficiência global e a local de um vértice, o que informa como a ausência deste vértice impacta a rede, sendo que valores negativos demonstram que a eficiência da rede é maior com a presença do vértice.

Exemplo 2.5. *A vulnerabilidade de A na Figura 2.1 é: $V(v) = 1 - \frac{E_{loc}(v)}{E(v)} = 1 - \frac{0.82}{0.88} =$ -0,07.*

Definição 2.9. *O **Coeficiente de Clusterização** usa a quantidade de triângulos formados no grafo como métrica. Seja $T(v)$ o número de triângulos formados pelo vértice v, o coeficiente de clusterização é computado como mostra a Equação 2.7 (SARAMÄKI et al., 2007).*

$$C(v) = \frac{2T(v)}{deg(v)(deg(v) - 1)} \tag{2.7}$$

O coeficiente de clusterização apresenta a importância do vértice para gerar triangulações entre seus vizinhos, servindo como 'ponte' entre eles. Sendo que valores mais próximos a 1 representam a necessidade deste vértice para a triangulação.

Exemplo 2.6. *O coeficiente de clusterização do vértice A no grafo da Figura 2.1 é obtido com a quantidade de triângulos formados a partir de A. Pela topografia do grafo temos os triângulos (A, E, C) e (A, E, B), sendo $T(A) = 2$. Dessa forma: $C(a) = \frac{2T(A)}{deg(A)(deg(A)-1)} = \frac{2\times2}{3(3-1)} = 0.67$.*

A próxima métrica utiliza o conceito de entropia. Em teoria da informação, a entropia está relacionada com o nível médio de aleatoriedade e incerteza sobre os possíveis resultados de uma variável aleatória. Seja $\{P(1), P(2), \ldots, P(n)\}$ a distribuição de probabilidade dos possíveis eventos de uma variável aleatória, a entropia é computada como: $-\sum_{i=1}^{n} P(i)\log(P(i))$, onde a base do logaritmo pode variar dependendo da aplicação (SHANNON, 1948).

O primeiro passo para se obter a entropia é computar uma distribuição de probabilidade a partir da rede. A distribuição de probabilidade representará a transição entre os vértices da rede em um trajeto aleatório (*random walk*) de h passos. Nesse caso, iniciando em um vértice i, qual a probabilidade de alcançar o vértice j após h passos? Para simplificar, assumiremos nesta primeira explicação que os caminhos não podem utilizar a mesma aresta duas vezes (*self-avoiding random walk*). Assim, partindo-se do vértice A no grafo da Figura 2.1, podemos acessar em apenas um passos os vértices B, C e E com probabilidade $\frac{1}{3}$. Feito isso, temos as possibilidades:

- A partir de B, poderemos apenas alcançar E (dado que não podemos voltar até A).
- A partir de C, poderemos alcançar E ou D no segundo passo, sendo que probabilidade de fazer cada uma dessas escolhas é $\frac{1}{2}$.
- A partir de E, poderemos alcançar tanto B quanto C, com probabilidade $\frac{1}{2}$ para cada.

Dito isso, a probabilidade de sair de A e alcançar cada um dos outros vértices em dois passos poderia ser computada:

- $P_2(A, B) = \frac{1}{3} \times \frac{1}{2} = \frac{1}{6}$; usando o caminho (A, E, B) apenas.
- $P_2(A, C) = \frac{1}{3} \times \frac{1}{2} = \frac{1}{6}$; usando o caminho (A, E, C) apenas.
- $P_2(A, D) = \frac{1}{3} \times \frac{1}{2} = \frac{1}{6}$; usando o caminho (A, C, D) apenas.
- $P_2(A, E) = \frac{1}{3} \times \frac{1}{1} + \frac{1}{3} \times \frac{1}{2} = \frac{1}{2}$; usando respectivamente (A, B, E) e (A, C, E).

Este exemplo ilustra o conceito de trajeto aleatório e apresenta como podemos gerar as probabilidades em uma configuração sem repetição de arestas. Entretanto, a implementação no trabalho considerou que as arestas podem ser percorridas várias vezes, o que gera a distribuição de probabilidade computada conforme a recorrência apresentada na Definição 2.10.

Definição 2.10. *Seja h o número de passos em um trajeto aleatório, a probabilidade de, partindo de um vértice i, alcançar um vértice j em um grafo é dada por:*

$$P_h(i,j) = \sum_k \frac{M_{kj}}{deg(k)} P_{h-1}(i,k) \tag{2.8}$$

Onde k representa os vizinhos de i e M representa a matriz de adjacência que contem as conexões entre os vértices, sendo $M_{i,j} = 1$ quando i for vizinho de j e $M_{i,j} = 0$ nos demais casos (NOH; RIEGER, 2004).

Com a probabilidade de transição de todos os vértices calculada, é possível obter a entropia.

Definição 2.11. *A **Entropia** de um vértice i sabendo as probabilidades de transição de i para todos os outros vértices é computada com (TRAVENÇOLO; DA F. COSTA, 2008):*

$$E_h(i) = -\sum_{vj} P_h(i,j) \log(P_h(i,j)) \tag{2.9}$$

Neste trabalho, *h* foi definido como 2, dessa forma foram contabilizadas as probabilidades somente de vértices adjacentes, pois somente nos deslocamos 2 vezes.

A entropia representa a heterogeneidade e a resiliência da rede a falhas, onde os valores da métrica representam a facilidade de se chegar aos vértices caminhando aleatoriamente na rede. Valores maiores demonstram uma maior importância do vértice para a construção dos caminhos na rede.

Exemplo 2.7. *Para calcular a entropia do vértice A, é necessário construir a matriz de adjacência para o grafo da Figura 2.1. Essa matriz pode ser observada na Tabela 2.1.*

	A	B	C	D	E
A	0	1	1	0	1
B	1	0	0	0	1
C	1	0	0	1	1
D	0	0	1	0	0
E	1	1	1	0	0

Tabela 2.1: Matriz de adjacências, demonstrando as conexões entre os vértices do grafo.

Com a matriz definida, computamos a probabilidade de transição de A para os demais vértices utilizando a Equação 2.8. Os resultados para todos os vértices podem ser observados na Tabela 2.2. Para ilustrar, temos $P_2(A,B) = \frac{M_{BB}}{deg(B)} P_1(A,B) + \frac{M_{EB}}{deg(E)} P_1(A,E) + \frac{M_{CB}}{deg(C)} P_1(A,C)$. Substituindo os

valores, é possível perceber que $P_{AB} = \frac{0}{2} \times \frac{1}{3} + \frac{1}{3} \times \frac{1}{3} + \frac{0}{3} \times \frac{1}{3} = 0.11$*, sendo então* 0.11 *a probabilidade de, saindo do vértice A, chegar ao vértice B após dois passos no grafo (saindo de A para algum vizinho, e desse vizinho para B).*

	Vizinhos de A			**Total (P_h)**
Vértices	B	E	C	
A	0.16	0.11	0.11	0.38
B		0.11		0.11
C		0.11		0.11
D			0.11	0.11
E	0.16		0.11	0.27

Tabela 2.2: Soma da probabilidade de transição iniciando no vértice A

Computando a entropia de A temos: $E_A = -[P_2(A,A)\log(P_2(A,A)) + P_2(A,B)\log(P_2(A,B)) + P_2(A,C)\log(P_2(A,C)) + P_2(A,D)\log(P_2(A,D)) + P_2(A,E)\log(P_2(A,E))]$*. Substituindo, temos* $E_A = -[0.38\log(0.38)+0.11\log(0.11)+0.11\log(0.11)+0.11\log(0.11)+0.27\log(0.27)]$*, que no final resulta em* 1.46*.*

Definição 2.12. *O* ***PageRank*** *em um grafo não direcionado mede a importância de um vértice com base na importância dos seus vizinhos (LANGVILLE; MEYER, 2004). Seja i um vértice no grafo e* $nei(i)$ *o conjunto de todos os vizinhos de i, o pagerank é computado como segue:*

$$p(i) = \frac{1-q}{n} + q \sum_{j \in nei(i)} \frac{p(j)}{deg(j)} \tag{2.10}$$

Na equação, q é uma constante chamada de fator de amortização definida normalmente como 0.85, n é a quantidade de vértices no grafo. Note que a equação atende o caso de grafos não direcionados, sendo escrita de forma diferente em grafos direcionados.

A Equação 2.10 é uma recorrência, dado ser definida em função de si mesma. A maneira mais comum de computar os valores é executando um número suficiente de iterações. Nesse caso, a resolução seguiria:

Inicialização: $p_0(i) = \frac{1}{n}$, onde n é o número de vértices.

Recursão: $p_t(i) = \frac{1-q}{n} + q \sum_{j \in nei(i)} \frac{p_{t-1}(j)}{deg(j)}$

Assim, $p_t(i)$ é o valor computado para $p(i)$ na t-ésima iteração.

Exemplo 2.8. *No grafo da Figura 2.1, antes da primeira iteração, temos que todos os vértices possuem pagerank* $p_0(A) = p_0(B) = p_0(C) = p_0(D) = p_0(E) = \frac{1}{5} = 0.20$. *Na primeira iteração, verificamos inicialmente quanto cada vértice recebe de pagerank de seus vizinhos. Assim, computando apenas para o vértice A, percebemos que ele recebe conexões de B, C e E, gerando então:* $p'_1(A) = \frac{p_0(B)}{2} + \frac{p_0(C)}{3} + \frac{P(E)}{3} = \frac{0.20}{2} + \frac{0.20}{3} + \frac{0.20}{3} = 0.233$. *Aplicando a amortização, temos* $p_1(A) = \frac{1-q}{n} + q \times p'_1(A) = \frac{1-0.85}{5} + 0.233 \times 0.85 = 0.228$. *Repetindo esse processo para todos os vértices temos:* $p_1(B) = 0.13$, $p_1(C) = 0.30$, $p_1(D) = 0.08$ *e* $p_1(E) = 0.23$.

Dados os novos valores de pagerank, temos $p'_2(A) = \frac{0.13}{2} + \frac{0.30}{3} + \frac{0.23}{3} = 0.35$, *o que amortizando gera o novo valor* $p_2(A) = 0.235$. *As iterações prosseguem até a convergência, o que garante-se obter após um número constante de passos, mesmo para grafos muito grandes.*

A Tabela 2.3 apresenta os valores obtidos das métricas descritas acima para todos os vértices da Figura 2.1.

	A	B	C	D	E
Centralidade	0.17	0.0	0.5	0.0	0.17
Excentricidade	2.0	3.0	2.0	3.0	2.0
Eficiência Global	0.88	0.71	0.88	0.58	0.88
Eficiência Local	0.82	0.84	0.82	0.82	0.82
Vulnerabilidade	-0.07	0.16	-0.07	0.29	-0.06
Clusterização	0.67	1.0	0.33	0.0	0.67
Entropia	1.46	1.33	1.15	1.1	1.46
Pagerank	0.23	0.14	0.31	0.09	0.23

Tabela 2.3: Tabela com os valores das métricas para o grafo de exemplo da Figura 2.1

2.2 Aprendizado de Máquina

Segundo Samuel (1959), aprendizado de máquina (do inglês, *machine learning*) é o campo de estudo em que os computadores possuem a habilidade de aprender a resolver problemas sem serem explicitamente programados. O aprendizado geralmente ocorre em uma das seguintes formas: supervisionado, não supervisionado e por reforço. Neste trabalho, foi utilizado o aprendizado supervisionado, através de uma rede neural, sendo essa uma técnica de treinamento baseada na otimização dos pesos das ligações da rede através de amostras rotuladas do problema.

Seja $x = x_1, x_2, \ldots, x_k$ um conjunto de amostras do problema, onde cada amostra, para $1 \leq i \leq k$, pode ser representada por um vetor de características $\vec{x_i} = x_{i1}, x_{i2}, \ldots, x_{in}$. Temos para cada amostra x_i um rótulo y_i associado, o que implica que, juntamente com as características de cada amostra, temos os rótulos $\vec{y} = y_1, y_2, \ldots, y_k$ que representam a solução esperada para as amostras. Dessa maneira, indicamos para o modelo qual deve ser a saída por meio do conjunto y e o classificador deve ser capaz de estimar uma saída $\hat{y}_i$ que seja a mais próxima possível de y_i para a amostra de treino x_i (GÉRON, 2019).

É importante ressaltar que o objetivo é um classificador capaz de fornecer boas estimativas para amostras que não fizeram parte do processo de treinamento. Quando isso acontece, é dito que o classificador consegue generalizar para novas amostras, ao contrário de memorizar os dados de treino. Nesse caso, uma metodologia baseada na divisão das amostras em conjuntos de treino, validação e teste é capaz de validar o modelo e obter acurácia próxima do que será obtido quando o modelo entrar em produção.

2.2.1 Descida do Gradiente

A descida do gradiente é um algoritmo de otimização utilizado para encontrar o valor mínimo de uma função. Esse processo é realizado atualizando os parâmetros da função objetivo na direção oposta à do gradiente, junto a uma taxa aprendizado que determina o tamanho do 'passo' na direção indicada. Dessa maneira, a descida do gradiente tende a minimizar os valores da função objetivo (RUDER, 2016).

A atualização dos parâmetros pode ser realizada de diversas formas: (i) após cada amostra, (ii) após todas as amostras, ou (iii) após um lote de amostras. Para o aprendizado de máquina, é geralmente utilizada a descida do gradiente estocástica, em que um lote de amostras é escolhido ao acaso, o que retira a necessidade de carregar o conjunto inteiro de amostras na memória.

Algoritmos de aprendizado de máquina geralmente utilizam descida do gradiente para minimizar a função que calcula o erro do modelo (*loss function*), permitindo atualizar os parâmetros do modelo de modo que o erro (divergência entre os rótulos e as previsões) seja diminuído gradativamente. A técnica da descida do gradiente é amplamente utilizada por classificadores como a regressão logística e as redes neurais artificiais, o que inclui redes *multilayer perceptron* (MLP) e redes com camadas convolucionais (CNN, do inglês *convolutional neural network*).

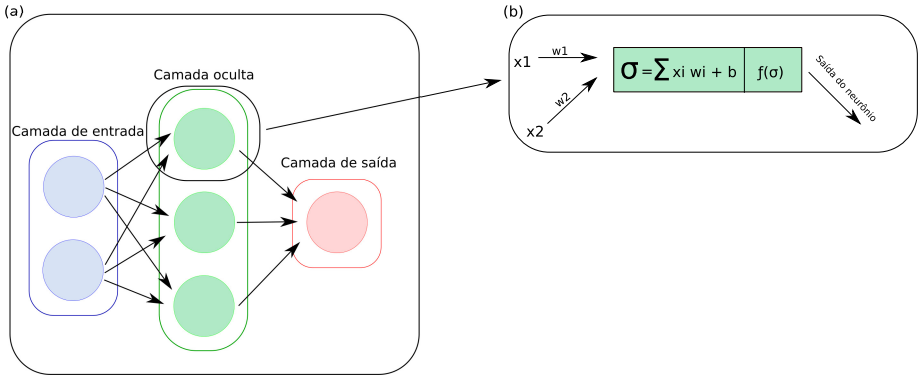


Figura 2.2: (a) Apresenta a arquitetura de uma MLP, (b) Exemplifica o funcionamento do neurônio destacado em (a), onde os pesos w são aplicados às saídas de cada neurônio da camada anterior $x1$ e $x2$ são somados e fornecidos a função $f(\sigma)$ de ativação sigmoidal.

2.2.2 Redes Neurais *Multilayer Perceptron*

Redes neurais *multilayer perceptron* são compostas por unidades de processamento chamadas neurônios, que funcionam como funções, combinando os valores recebidos como entrada e através dessa combinação é realizada a ativação do neurônio ou não.

A ativação se dá por meio da combinação linear das entradas do neurônio que são fornecidas a uma função sigmoidal. A Figura 2.2 esquematiza o funcionamento de um neurônio.

Dessa maneira, o aprendizado desse tipo de rede neural é realizado propagando os valores das entradas através das camadas até a última, que apresenta o resultado da rede. Esse resultado é comparado com o valor esperado da saída (rótulo original da amostra) utilizando alguma função de perda (do inglês, *loss function*). Dessa forma, é obtido o erro da rede. O erro encontrado é retropropagado para atualizar os pesos da rede para otimizar os seus resultados (HAYKIN, 2007).

O aprendizado profundo (do inglês, *Deep Learning*) é realizado em redes que possuem mais de uma camada oculta. Essas redes são conhecidas como redes profundas (do inglês, *Deep Neural Netwotk*), sendo geralmente compostas por uma camada de entrada, responsável por receber os dados iniciais da rede, seguida por diversas outras camadas (por isso a denominação de rede profunda), e, por fim, a última camada, conhecida como camada de saída, que apresenta o resultado do modelo (GÉRON, 2019). Entretanto, vale ressaltar que essa descrição pode não contemplar todas as arquiteturas de redes neurais existentes.

A Figura 2.3(a) apresenta uma rede neural *multilayer perceptron* com apenas uma camada oculta. A Figura 2.3(b) apresenta uma rede profunda com três camadas ocultas, cada uma delas com três neurônios totalmente conectados.

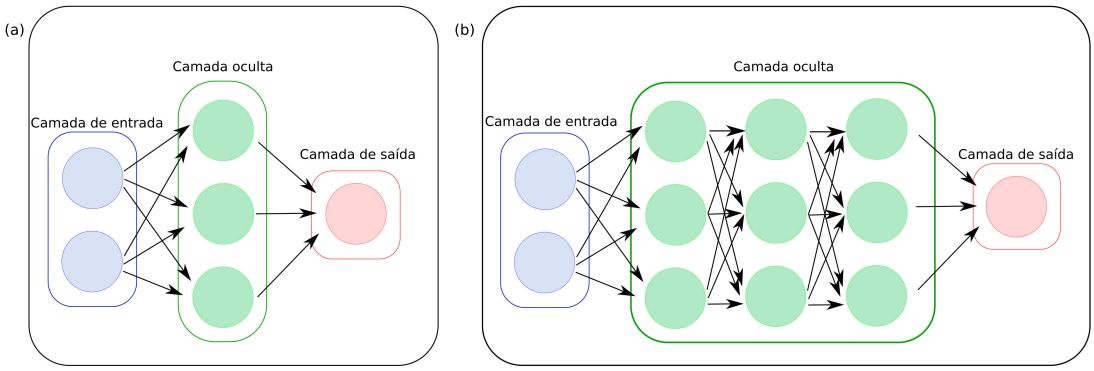


Figura 2.3: (a) Apresenta a arquitetura de uma rede neural com somente uma camda oculta, (b) apresenta a arquitetura de uma *DNN* com três camadas ocultas com neurônios totalmente conectados.

2.2.3 Redes Neurais Convolucionais

Redes neurais convolucionais (*CNN*, do inglês *Convolutional Neural Network*) são uma evolução das redes neurais artificiais, realizando o aprendizado com algumas camadas de convolução posicionadas normalmente no início da rede e responsáveis por extrair características das amostras fornecidas.

Essa extração ocorre através da aplicação de mapas de características, de modo a treinar os pesos dos neurônios de cada camada, dessa maneira, conseguem propagar a informação por essas camadas de forma bipiramidal, onde a quantidade de filtros aplicados aumenta, enquanto a quantidade de neurônios diminui, o que diminuindo o número de parâmetros da rede (HAYKIN, 2007).

As saídas das camadas convolucionais são matrizes bidimensionais chamadas de mapas de características, sendo geralmente seguidas por camadas de agrupamento (do inglês, *pooling*), que condensam a informação do mapa de características (GÉRON, 2019). Dessa maneira, a dimensão dos dados diminui, como ilustrado na Figura 2.4, em que a largura e a altura das matrizes bidimensionais decrescem nas camadas de *pooling*.

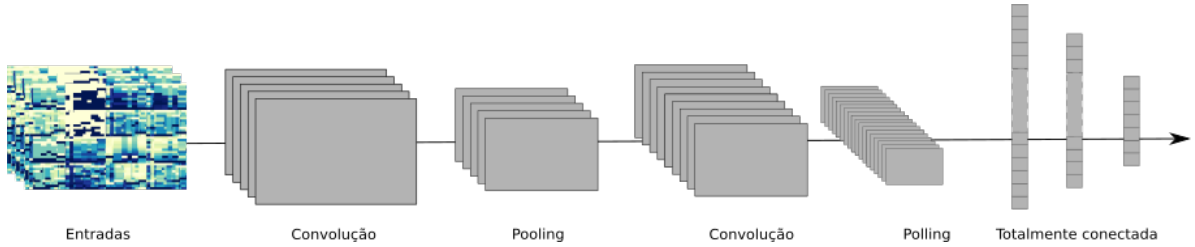


Figura 2.4: Arquitetura padrão de uma CNN, recebendo como entrada algumas imagens e possuindo diversas camadas de convolução e *pooling*, que diminuem a dimensão da informação passada entre as camadas até culminar em camadas totalmente conectadas no final.

2.2.4 Transferência de Aprendizado

Com uma CNN profunda, é possível reutilizar algumas de suas camadas de extração de características para um novo problema, principalmente quando a nova rede será utilizada para tarefas semelhantes. Esse processo é conhecido como transferência de aprendizado (do inglês, *transfer learning*) (PAN; YANG, 2010).

Essa técnica consegue acelerar a execução de um novo treinamento (dado não ser necessário treinar todas as camadas novamente) e diminui a quantidade de dados necessários. Por exemplo, com uma DNN treinada para identificar animais em imagens, algumas de suas camadas podem ser aproveitadas (extratoras de características da imagem) para gerar uma nova rede que cataloga pássaros.

A Figura 2.5(a) representa uma rede já treinada e a Figura 2.5(b) representa uma nova rede que aproveitou o treinamento das camadas iniciais e modificou somente o final para o seu problema específico. Nesse caso, as camadas "Oculta 1" e "Oculta 2" não tiveram seus pesos alterados durante o novo treinamento. A arquitetura da camada "Oculta 3" foi recebida da rede anterior, mas os pesos foram alterados durante o treinamento da nova rede, sendo esse processo chamando ajuste fino (do inglês, *fine-tuning*). A camada "Oculta 4" foi substituída por uma nova camada. Por fim, note que a nova rede tem um número de camadas menor que a rede original, dado que "Oculta 5" não foi aproveitada nem substituída.

Vale ressaltar que a modificação ao final da rede não precisa se resumir a simplesmente adicionar novas camadas totalmente conectadas, apesar de essa ser uma prática comum. É possível até mesmo adicionar classificadores nada relacionados a redes neurais artificiais. Assim, a camada chamada de "Nova Camada 4", na Figura 2.5, pode muito bem representar um classificador baseado em técnicas diversas como *Support Vector Machine*, *Random Forest* ou outros.

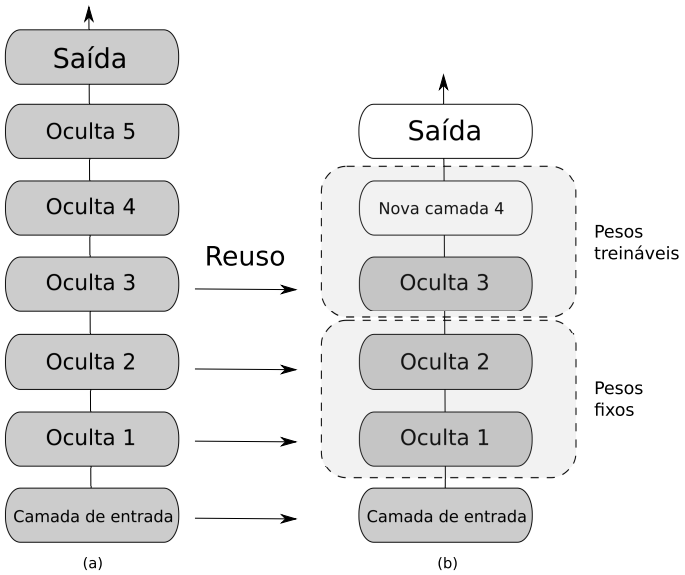


Figura 2.5: Processo de transferência de aprendizado. Em (a) uma DNN treinada com cinco camadas ocultas e uma camada para a saída. Em (b) uma nova DNN reutiliza a camada de entrada e os pesos treinados das duas primeiras camadas ocultas, deixando as seguintes (três e quatro) livres para aprender com os novos dados de entrada, gerando assim uma nova saída.

2.3 Validação dos Modelos

Treinados os classificadores de aprendizado de máquina para executar uma tarefa específica, uma validação dos resultados em instâncias não utilizadas no treinamento é necessária. A metodologia de separação dos dados e as métricas para descrever o desempenho na execução da tarefa precisam ser definidas antes de realmente iniciar o treinamento. Por fim, um teste de significância estatística verifica se as variações apresentadas pelas métricas são realmente importantes, ou se podem ser variações facilmente explicadas pela aleatoriedade presente nos experimentos. Esta seção introduzirá tais questões.

2.3.1 Validação Cruzada

A validação cruzada (do inglês, *cross-validation*) é uma maneira organizada de separar os dados em treinamento, validação e teste. Essa técnica permite medir como um classificador consegue generalizar o que foi aprendido com os dados de treinamento, computando o erro durante o treinamento com os dados de validação. Durante a fase de treinamento e validação, é possível fazer ajustes manuais nos hiperparâmetros de modo a gerar classificadores melhores. O melhor classificador é, então, selecionado e medido com os dados de teste.

De maneira mais precisa, a validação cruzada separa os dados inicialmente em dois conjuntos: treinamento e teste, sendo o segundo passo dividir o de treinamento em K subconjuntos $\{S_1, S_2, \ldots, S_k\}$ de aproximadamente o mesmo tamanho. Feito isso, para cada subconjunto S_i, será realizado o treinamento de um modelo com os $K - 1$ subconjuntos diferentes de S_i, deixando S_i para medir o erro do modelo treinado, sendo então chamado de conjunto de validação.

Completada a etapa de validação e tendo todos os hiperparâmetros configurados adequadamente para o classificador, o conjunto de testes é utilizado para medir a qualidade do melhor classificador, aquele que apresentou menor erro quando aplicado nos dados de validação (BENGIO; GRANDVALET, 2004). A Figura 2.6 ilustra as etapas.

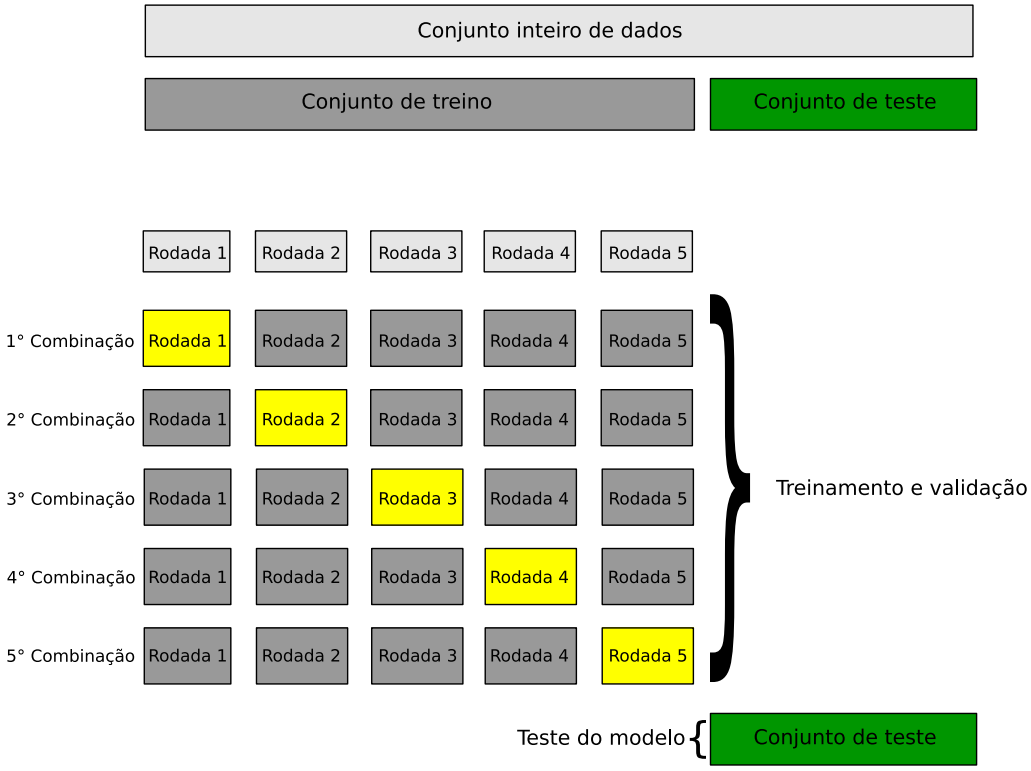


Figura 2.6: Validação cruzada com $K = 5$ subconjuntos. Primeiro divide-se os dados em dois conjuntos: treino e teste. Posteriormente, o conjunto de treino é subdividido em 5 subconjuntos e são realizadas cinco rodadas de treino e validação, alternando o subconjunto de validação. Ao final do processo, o melhor modelo (entre as cinco rodadas) é analisado com os dados do conjunto de teste.

São obtidos dois tipos de medidas dos modelos: o erro de validação e o erro de teste, onde o erro de validação permite identificar como o modelo se adaptou aos dados de treinamento, além de permitir configurar os melhores hiperparâmetros, e o erro de teste mostra a generalização com dados novos finalizada toda a etapa de aprendizado de máquina (BERRAR, 2019). Treinar

classificadores utilizando essa técnica permite diminuir o viés dos resultados e aumentar a confiabilidade.

2.3.2 Métricas de Desempenho

Um classificador pode ter o seu resultado medido de diversas formas, uma das mais conhecidas é a matriz de confusão (do inglês, *confusion matrix*) (POWERS, 2008) representada na Figura 2.7. A matriz de confusão quando aplicada para um problema de duas classes permite classificar as respostas do classificador em quatro quadrantes. Para facilitar a explanação, suponha que um determinado classificador detecta se uma pessoa tem uma doença, fornecendo as respostas "Sim" para doente e "Não" para saudável. A pessoa, por sua vez, pode estar doente (fato rotulado como "Sim" no banco de dados) ou saudável (fato rotulado como "Não" no banco de dados). Dessa forma, a interpretação dos quadrantes seriam:

Verdadeiro Positivo (VP): o classificador predisse que a pessoa estava doente ("Sim"), sendo que a pessoa estava realmente doente ("Sim").

Verdadeiro Negativo (VN): o classificador predisse que a pessoa estava saudável ("Não"), sendo que a pessoa estava realmente saudável ("Não").

Falso Positivo (FP): o classificador predisse que a pessoa estava doente ("Sim"). No entanto, a pessoa estava saudável ("Não").

Falso Negativo (FN): o classificador predisse que a pessoa estava saudável ("Não"). No entanto, a pessoa estava doente ("Sim").

Com base nessa classificação, é possível obter métricas utilizadas posteriormente para computar a acurácia balanceada: revocação e especificidade.

- **Revocação**

  Revocação ou *recall* computa das instâncias rotuladas como positivas no banco de dados quantas foram corretamente classificadas como positivas.

$$recall = \frac{VP}{VP + FN} \tag{2.11}$$

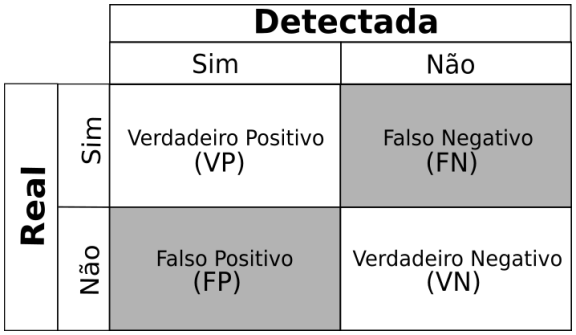


Figura 2.7: Matriz de confusão representando a distribuição das predições (**Detectada**) em comparação com o rótulo presente no banco de dados (**Real**).

- **Especificidade**

  A especificidade (do inglês, *specificity*) é a medida dos verdadeiros negativos, ou seja, computa das instâncias rotuladas como negativas no banco quantas foram corretamente classificadas como negativas.

$$specificity = \frac{VN}{VN + FP} \tag{2.12}$$

A acurácia balanceada (*BA*) é uma maneira de se analisar o resultado das predições de classificadores binários (BRODERSEN et al., 2010), sendo geralmente aplicada sobre um conjunto de dados não balanceados, ou seja, quando a distribuição das classes não está proporcional, existindo muito mais amostras de uma classe do que de outra, o seu funcionamento é descrito na Formula 2.13.

$$BA = \frac{recall + specificity}{2} \tag{2.13}$$

O denominador representa a quantidade de classes (nesse exemplo, dois representa uma classificação binária). A equação permite identificar qual a relação entre acertos em ambas as classes (positivas e negativas).

2.3.3 Teste de Hipótese

Um teste de significância, ou teste de hipótese, permite rejeitar uma hipótese estatística com base nos resultado das amostras (GRAYBILL; IYER; BURDICK, 1998). Por exemplo, verificar se um novo medicamento consegue diminuir a massa corporal. O teste acontece ao comparar os

resultados de dois grupos de pessoas (as que tomaram o medicamento e as que não tomaram o medicamento), os resultados poderiam ser algo como apresenta a Tabela 2.4.

A hipótese em questão a ser testada é definida como H_0 ou hipótese nula: "Ao ministrar a droga, é apresentada uma perda de massa comparada a quem não fez o seu uso".

Perda de massa dos voluntários de cada grupo

Voluntário de cada estudo	Com o novo medicamento	Sem o novo medicamento
Voluntário 1	3	2
Voluntário 2	5	7
Voluntário 3	3	4
Voluntário 4	8	3
Voluntário 5	7	12
Massa total perdida	26	28

Tabela 2.4: Exemplo da demonstração de perda de massa entre os dois grupos, os que tomaram o medicamento ou não. Cada linha representa um voluntário distinto que teve a sua perda de massa acompanhada, sendo no total de 10 voluntários distintos (5 para cada grupo)

Com os valores de cada grupo, é possível inferir se a diferença da perda de massa entre os dois grupos é estatisticamente relevante. Dependendo do resultado, aceitamos ou recusamos a hipótese nula. A relevância é medida por meio do valor de p, sendo que quanto menor o seu valor, menor a probabilidade da diferença entre os grupos de amostras ser um resultado ao acaso.

Existem diversos tipos de testes que podem ser utilizados para responder a essa questão, sendo eles divididos em paramétricos e não paramétricos. Um teste não paramétrico é recomendado quando a distribuição não segue uma distribuição normal ou quando a quantidade de amostras é insuficiente para realizar um teste paramétrico, uma vez que testes paramétricos necessitam de mais dados (MUNDRY; FISCHER, 1998).

Neste trabalho, foi selecionado o teste de *Wilcoxon*, utilizado para demonstrar se um conjunto de K amostras pertencem a uma mesma população ou populações com a mesma média. O teste é construído utilizando como entrada uma matriz onde r linhas representam as variáveis e K colunas os conjuntos que serão comparados. Outro ponto importante é que devem existir ao menos dois grupos de amostras $K >= 2$ (CORDER; FOREMAN, 2011).

Aplicando o teste de *Wilcoxon* ao conjunto de dados da Tabela 2.4, obtemos um valor de p de 0.81, assim não rejeitamos a hipótese nula H_0.

2.4 Explicabilidade de Modelos

É desejável que os modelos de aprendizado de máquina, além de apresentarem predições corretas, sejam interpretáveis. Dessa forma, seria possível entender como as variáveis fornecidas ao modelo influenciam em seu resultado.

Cada arquitetura de classificador pode ser interpretada de uma forma. Por exemplo: em modelos lineares é possível identificar o impacto de cada atributo através dos valores de seus coeficientes, bem como as árvores de decisão permitem uma interpretação por meio da ramificação dos atributos.

Entretanto, a interpretabilidade de determinadas arquiteturas de classificadores nem sempre é intuitiva, sendo este o caso das redes neurais e das combinações de modelos. Assim, é necessária a utilização de outras técnicas para compreender o resultado obtido.

Uma das possíveis técnicas para explicar como o classificador realizou a predição é a utilização da teoria dos jogos, sendo esta uma área de estudo usualmente empregada para caracterizar o comportamento de indivíduos, tendo como foco melhorar o retorno obtido da ação coletiva deles (MYERSON, 2013).

Conforme mencionado no parágrafo anterior, a aplicação da teoria dos jogos pode ser representada da seguinte forma: os indivíduos (aqueles que buscam por um melhor retorno) são os atributos fornecidos aos classificadores e, como resultado, é considerada a 'certeza' do classificador na predição da classe correta da amostra.

Uma técnica que aplica o conceito da teoria dos jogos na explicabilidade de classificadores são os valores de Shapley (do inglês, *Shapley Values*), obtidos identificando a alteração do resultado (nesse contexto, a alteração da certeza do classificador) removendo o indivíduo do coletivo (nesse contexto, não informando um atributo para o classificador) (SHAPLEY, 1951).

A 'remoção' do atributo ocorre treinando dois novos classificadores lineares, que simulam o classificador originalmente treinando contendo todos os atributos, porém, um com o atributo que está sendo analisado e outro sem. O resultado entre os classificadores são comparados e o *Shapley Value* do atributo é definido como positivo caso a inclusão dele melhore a 'certeza' da predição, e como negativo se o valor diminuir.

2.5 Aplicação dos conceitos

Os conceitos definidos neste capítulo foram aplicados na metodologia deste trabalho. Os atributos extraídos das redes complexas foram computados para todos os grafos construídos, assim possibilitando a obtenção de informações de ambas as equipes na partida.

Aprendizado de máquina foi o método escolhido para identificar quais seriam os padrões responsáveis para uma equipe chegar ao quarto final do campo. Esse processo envolve outros pontos apresentados neste capítulo, como a descida do gradiente responsável pela otimização da rede neural, que teve as duas possibilidades de camadas utilizadas: densas e convolucionais.

A transferência de aprendizado foi empregada para aproveitar redes já treinadas com outros conjuntos de dados, dessa forma poupando esforços para otimizar os pesos da nossa rede. A validação do treinamento ocorreu com a técnica de validação cruzada visando evitar o sobre ajuste aos dados, junto da acurácia balanceada para medir o erro na classificação das posses de bola.

Após o treinamento foi utilizado o teste estatístico de Wilcoxon para identificar se os resultados dos classificadores diferenciavam do acaso. Também foi selecionada a melhor rede para a análise dos atributos usando biblioteca SHAP.

Capítulo 3

Levantamento Bibliográfico

A literatura apresenta diversos trabalhos na área de análise de partidas de futebol, sendo necessário agrupar os trabalhos em conjuntos menores para entender o propósito de cada um. A área foi dividida nas seguintes categorias.

- **Campo**: métricas extraídas com base nas posições dos jogadores no campo e na movimentação.
- **Equipe**: métricas que analisam a posse de bola das equipes e situações que levam a equipe até o gol.
- **Grafos**: métricas de redes baseadas em estruturas de grafos, como: centralidade, eficiência e vulnerabilidade da rede.

Na Seção 3.1, abordaremos as métricas que tratam das posições dos jogadores em campo e sua movimentação, sendo o foco trabalhos que demostram a importância e maneiras de analisar as zonas do campo. Na Seção 3.2, o foco passa a ser as ações dos jogadores como equipe, aquelas que necessitam de interação direta entre os jogadores. A Seção 3.3 contextualiza métricas de grafos, demonstrando trabalhos da área de futebol e seus resultados obtidos utilizando análises sobre grafos específicos.

3.1 Campo

Esta seção apresenta métricas baseadas nas posições dos jogadores nas zonas do campo e na movimentação.

3.1.1 Zonas do Campo

Ao analisar as regiões de campo e quanto interferem nas chances e qualidade das finalizações, geralmente os pesquisadores particionam o campo em sub-áreas para medir a influência de cada uma no resultado da jogada. O estudo de Wright et al. (2011) particiona a zona de ataque em oito divisões como mostra a Figura 3.1 e mede a taxa de conversão de gols de cada uma, concluindo que, de dentro da grande área, ocorrem a maioria das tentativas de finalização. Nesse mesmo estudo, os autores apresentam que a taxa de retomada de bolas que geraram contra-ataques com finalização ocorrem no meio de campo (44% das ocorrências).

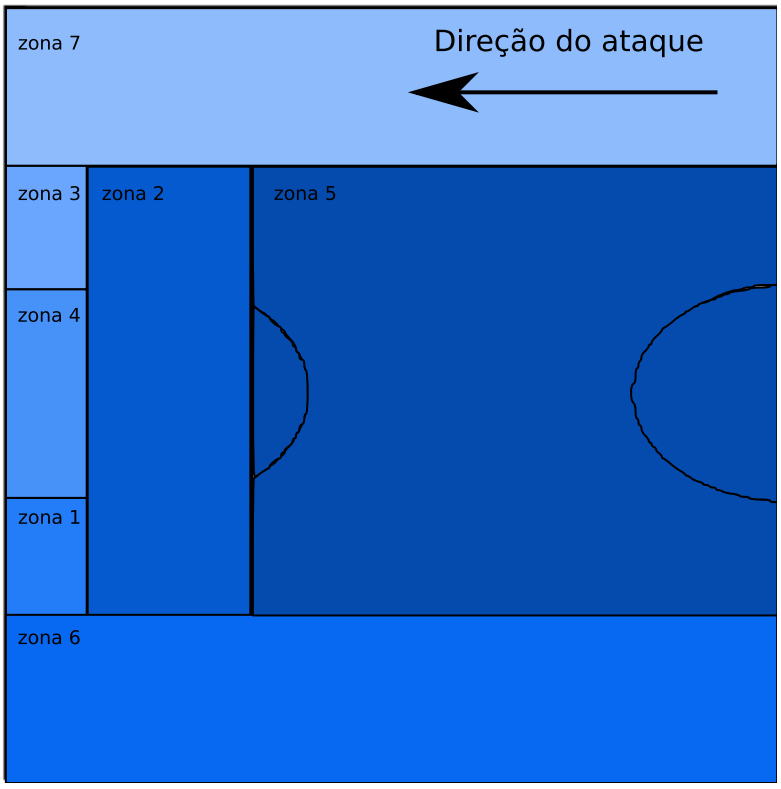


Figura 3.1: Zonas do campo definidas no trabalho de Wright et al. (2011). As zonas 1, 2, 3 e 4 representam 87% das finalizações.

3.1.2 Pressão em Campo

Vilar et al. (2013) utilizaram a pressão de uma equipe como ponto de partida para analisar seu comportamento e predizer valores futuros de posse de bola. A pressão é definida como o deslocamento de uma quantidade de jogadores para uma determinada zona do campo (um time aplica pressão ofensiva quando os jogadores se direcionam para a zona de ataque com a posse

de bola), sendo a pressão entendida como a existência de um desequilíbrio entre a quantidade de atacantes e a quantidade de defensores da equipe adversária.

Barreira et al. (2014) estudaram como o comportamento das equipes pode diferir dependendo da zona do campo em que a equipe obteve a posse. Por exemplo, retomadas na região defensiva tendem a possuir mais passes do que as retomadas nas zonas de ataque.

3.1.3 Posição dos Jogadores

Fernandez-Navarro et al. (2016) apresentaram dados demonstrando como a pressão aplicada por uma equipe sobre a outra influencia no esquema tático, informação obtida por meio da síntese de outros artigos e de observações do campeonato espanhol de 2006, considerando principalmente o Futebol Clube Barcelona (campeão naquele ano). A equipe apresentava dominância nas partidas, aplicando pressão no campo de defesa adversário, aliada de uma alta taxa de posse de bola e passes rápidos. A pesquisa deixou questões em aberto para realizar a predição do vencedor das partidas baseada nas métricas de posicionamento.

A importância do posicionamento dos jogadores também pode ser vista na tese de Moura et al. (2011) sobre partidas do campeonato brasileiro de futebol, em que foi possível observar que a área das equipes variou entre atacar e defender, demonstrando correlação entre a distribuição dos atletas e a ação que estavam executando (atacando ou defendendo). Outra conclusão relevante foi que defesas mais espaçadas tendiam a sofrer mais gols, conforme os dados analisados pelo cálculo da mediana da dissipação dos jogadores.

3.1.4 Tomada de Posse de Bola

Moura et al. (2007) utilizaram dados de partidas do campeonato brasileiro de futebol e registraram 86 jogadores diferentes para identificar em qual região do campo (ataque ou defesa) as equipes realizavam a maior taxa de retomada de posse de bola revertidas em chutes a gol (com ou sem sucesso).

No total, foram utilizadas 106 finalizações (chutes a gol) para a conclusão, sendo identificado que a maioria das tomadas de posse ocorria no campo de defesa, sendo essa informação uma possível métrica para analisar se uma jogada se tornará perigosa.

A tese de Merlin (2020) também estudou posses de bola em partidas de futebol, utilizando quatro partidas da primeira divisão do campeonato brasileiro de futebol, estruturando os dados em 527 intervalos de posses de bola. O objetivo foi classificar as posses de bola utilizando

métricas como: tempo de posse de bola, quantidade de passes realizados com sucesso, área de cobertura dos jogadores de cada equipe e centroide dessa área. Foi possível agrupar as posses de bola em três categorias: curta, média e longa, considerando o tempo de posse e a quantidade de passes bem sucedidos, como mostra a Figura 3.2.

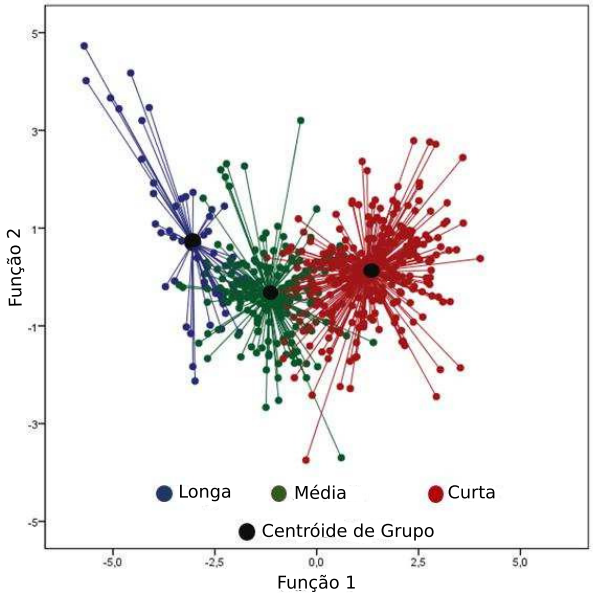


Figura 3.2: Agrupamento das posses de bola em curta, média e longa, considerando o tempo da posse e a quantidade de passes, imagem retirada da tese de Merlin (2020), onde as funções 1 e 2 foram obtidas através da discriminante linear de Fisher (BICKEL; LEVINA, 2004), que separa classes através da variância dos dados.

3.1.5 Jogadores por Zona

Como apresentado na pesquisa de Tenga et al. (2010), a quantidade de jogadores dentro de uma determinada zona do campo influencia na taxa de sucesso da equipe ofensiva. O experimento utilizou partidas do *Norwegian professional soccer 'Eliteserien'*. Foi possível concluir que equipes que possuíam mais pressão no ataque (existindo mais atacantes do que defensores) tendiam a ter um desempenho melhor contra defesas não balanceadas (sofriam pressão mais facilmente). Estes resultados foram obtidos por meio de técnicas de regressão linear entre a taxa de sucesso do ataque e a pressão de cada equipe.

Lucey et al. (2014) analisaram partidas profissionais de diversas ligas fornecidas pela Prozone Sports Ltd, empresa de análise de desempenho esportivo. Os dados continham 9.732 finalizações, considerando os dez segundos antes de cada finalização ocorrer e utilizando esses dados para predizer as chances de um gol acontecer baseado em: (i) quantidade de

defensores próximos à bola; (ii) existência de zagueiros entre a área de finalização e o gol; e (iii) posição do chute a gol. Foi possível identificar o comportamento dos atacantes em diferentes situações, por exemplo: quando não existem zagueiros entre o gol e a bola a determinada distância, a chance da finalização ser convertida em gol é maior.

Com os dados recolhidos, foi possível prever o gol com 70% de precisão em algumas jogadas. A quantidade de jogadores entre a bola e o gol também se mostrou um diferencial na decisão de realizar a finalização ou não, baseando-se no dados obtidos do estudo de Wright et al. (2011), em que a taxa de chutes ao gol foi identificada como 60% em situações em que havia entre 0 e 2 defensores à frente.

3.2 Equipe

Esta seção apresenta métricas que analisam a posse de bola das equipes e situações que conduzem ao gol adversário.

3.2.1 Quantidade de Passes

Diversos estudos mostram a correlação entre a quantidade de passes e a taxa de gols realizados como, por exemplo, o estudo de Wright et al. (2011), com objetivo de identificar quais aspectos tornam as finalizações mais assertivas utilizando regressão logística (usando como rótulo o fato de haver gol ou não). O estudo utilizou dados da *English FA Premier League*, com 1778 finalizações e 169 gols marcados, sendo verificado que 85% dos gols marcados foram com quatro passes ou menos. Outro ponto observado no mesmo trabalho foi que passes longos apresentam uma tava de conversão de gol menor que passes curtos.

A tese de Merlin (2020) também abordou o estudo de passes, utilizando o classificador *SVM* para rotular a dificuldade de 2.522 passes em três classes de dificuldade: baixa, média e alta, sendo os rótulos dos dados informados por especialistas na área. Como métricas para o classificador, foram utilizadas variáveis como: (i) proximidade de adversários do jogador que realizou o passe, (ii) velocidade do jogador passante, (iii) distância percorrida pela bola no passe e (iv) ângulo formado entre os jogadores. Como resultado, o classificador SVM apresentou uma acurácia balanceada média de 88% em classificar a dificuldade do passe de acordo com o rótulo dos especialistas.

3.2.2 Situações de Chance de Gol

Diversas métricas já foram utilizadas em modelos para identificar quais fatores geram maiores chances de marcar um gol. O estudo de Tenga et al. (2010) apresentou que contra-ataques geram oportunidades de gol, principalmente aqueles onde a bola foi roubada no meio-campo. Tenga et al. (2010) observaram que equipes que possuem uma maior pressão no ataque (mais atacantes do que defensores no campo de ataque) tendem a possuir melhor aproveitamento das finalizações.

O centroide das equipes foi utilizada para classificar posses de bola no trabalho de Frencken e Lemmink (2009). Essa métrica representa o centro do posicionamento dos jogadores da equipe, considerando a área de cobertura e os jogadores nas regiões mais extremas do jogo, como mostra a Figura 3.3, onde foram representados os jogadores mais extremos no posicionamento para definir a área total de cobertura da equipe.

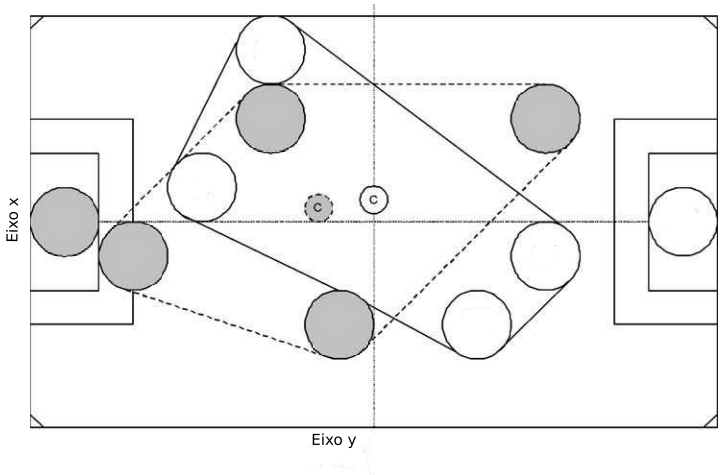


Figura 3.3: Demonstra a distribuição de duas equipes, onde o vértice *c* representa a centroide da equipe, imagem retirada do trabalho de Frencken e Lemmink (2009).

Observando a jogadas que originaram gols, foi possível concluir que a distância das centroides se aproximava de zero, e em 53% dos gols a centroide da equipe atacante ultrapassava a defensora em direção ao gol, demonstrando possuir uma maior pressão no campo de ataque.

3.3 Métricas de Grafos

A movimentação dos jogadores em campo, junto das ações que são executadas (por exemplo, passes, finalizações e tomadas de posse de bola), podem ser analisadas de diversas formas. O

trabalho de Rodrigues et al. (2019) apresentou como as partidas podem ser representadas em grafos e como métricas de redes complexas podem ser geradas sem ignorar os aspectos temporais que a partida possui. Diversas métricas foram extraídas como: centralidade, entropia, eficiência, entre outras.

Grafos também foram utilizados no trabalho de Grund (2012), que, utilizando 760 partidas da *English Premier League*, buscou identificar a relação entre a importância de um jogador nos passes da equipe, notando que quanto maior a centralidade, mais passes eram realizados ou recebidos. Para cada partida foram construídos dois grafos (um para cada equipe), sendo que os jogadores representavam os vértices e as arestas a quantidade de passes realizados entre os jogadores, dessa forma medindo a conectividade de um jogador durante o jogo. A conclusão foi que, conforme a importância de um indivíduo aumenta, pior a performance da equipe, geralmente devido à dependência do esquema tático em relação ao atleta.

O trabalho de Arriaza-Ardiles et al. (2018) extraiu, a partir de um grafo direcionado e ponderado, métricas como centralidade e agrupamento dos jogadores de uma mesma equipe, concluindo que existem padrões para jogadas ofensivas e defensivas. Também foi observado que jogadores com maiores valores de centralidade participavam de mais jogadas da equipe.

Grafos também foram utilizados na pesquisa de Dias et al. (2019), que desenvolveram um *framework* para facilitar a visualização das métricas de rede considerando toda a série temporal da partida. Como resultado, foi possível obter uma forma visual de observar as informações de cada jogador, com o passar do tempo, em uma única imagem.

A pesquisa de Clemente, Sarmento e Aquino (2020), com partidas da copa do mundo de futebol de 2018, também utilizou métricas de redes com objetivo de identificar padrões nas partidas, observando que jogadores com maiores centralidades das equipes vencedoras geralmente eram meio campistas ou laterais. Em comparação com um estudo similar da copa de 2014, notou-se uma diferença nesse padrão, onde em 2014 as maiores centralidades se encontravam nos zagueiros ou laterais.

3.4 Aplicações no Trabalho

Os trabalhos de Wright et al. (2011), Lucey et al. (2014) e Tenga et al. (2010) tiveram importância neste trabalho, visto que o objetivo dos classificadores treinados é caracterizar posses de bola que chegaram nas zonas próximas à grande área adversária. Outra pesquisa

utilizada como base para o problema tratado foi a de Fernandez-Navarro et al. (2016), que deixou para trabalhos futuros a questão da predição de eventos em partidas considerando métricas de posicionamento, sendo diretamente respondida por nossa análise, visto que grafos são uma representação do posicionamento dos jogadores.

A transformação de métricas em imagens também pode ser observada no trabalho de Rodrigues et al. (2019), que representam grafos em imagens de ritmo visual, e em Dias et al. (2019), que codificaram séries temporais em imagens de sensoriamento remoto para identificar regiões com eucalipto em dados de satélite, mostrando que a abordagem é válida em escopos em que temos uma série temporal de alguma medida.

A maior diferença deste trabalho em relação aos demais estudados na literatura está na abordagem utilizada, onde a análise de partidas de futebol foi além da modelagem de grafos, realizando a união dessa representação com o processamento de imagens, gerando assim uma nova abordagem para o problema.

Capítulo 4

Metodologia

Este capítulo apresenta o tratamento feito nos dados e a divisão conceitual das partidas de modo a obter características para treinar a rede. Além disso, o capítulo aborda como interpretar as métricas de grafos no contexto do futebol.

O conjunto de dados é composto por dez partidas de futebol realizadas com diferentes equipes do campeonato brasileiro, dentre elas: Redbull, Athletico Paranaense, Corinthians, Palmeiras, Sport e Flamengo. Foram catalogados eventos destas partidas, como: passes, chutes, faltas e impedimentos, além da posições dos jogadores a cada instante de tempo. Esse conjunto de dados já foi tema de estudo sobre a dificuldade de passes em partidas de futebol (MERLIN, 2020)

A obtenção dos dados segue uma metodologia baseada no trabalho de especialistas de futebol. Os eventos e as posições dos jogadores foram extraídos por especialistas a partir de um vídeo das partidas gravado a 30Hz e utilizando o *software* de extração semi-automático Dvideo (LEITE DE BARROS et al., 2006), que extrai dos *frames* do vídeo a posição dos jogadores em relação ao campo de futebol, gerando coordenadas $(x_{p,t}, y_{p,t})$ para cada jogador p da partida em cada instante de tempo t, considerando o comprimento e a largura do campo.

4.1 Dados

Os dados extraídos das partidas foram organizados em intervalos de posse de bola (BPI - *Ball Possession Interval*), como ilustra a Figura 4.1. Cada BPI representa o intervalo de tempo desde o momento em que uma equipe conseguiu a posse de bola até o momento em que a perdeu.

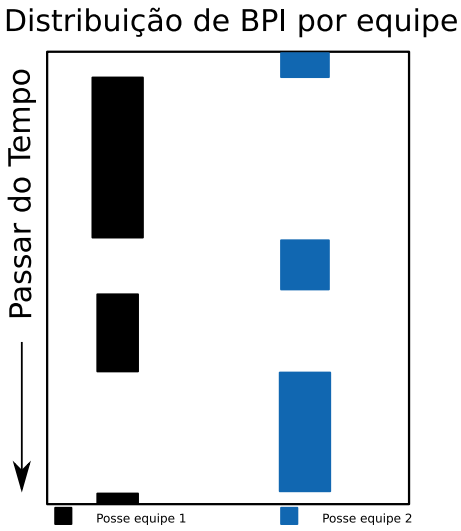


Figura 4.1: Os retângulos representam os intervalos de posse de bola (BPI) de cada equipe. No decorrer do tempo, a bola alterna de equipe, sendo algumas posses maiores que outra devido ao tempo que cada equipe permanece com o domínio da bola.

Cada BPI foi dividido em três partes, como ilustra a Figura 4.2. O significado de cada uma dessas partes é apresentado a seguir:

- **FETW** (*Feature Extraction Time Window*): representa os primeiros 167 *frames* (aproximadamente 5 segundos) do BPI. Nesse espaço de tempo, foram extraídas as características para treinamento dos classificadores.
- **Atraso** (*Lag*): intervalo variável de tempo entre o FETW e o final da posse de bola. O *lag* corresponde à janela central do BPI.
- **Target**: faixa de tempo final estipulada em 167 *frames* no BPI. Nessa faixa, é verificado se a equipe conseguiu chegar na zona de interesse, o que implica dizer que esta faixa gera o rótulo (sucesso ou fracasso) do BPI para compor o banco de dados, em conjunto com as características extraídas no FETW.

4.2 Seleção de BPIs

Neste trabalho, os intervalos de posse de bola são o centro das análises. Considerando as 10 partidas no banco de dados, observamos um número de 4.448 posses de bola, que variam em tamanho dependendo do tempo em que a equipe permaneceu com a bola, o que indica que uma análise da distribuição desses tempos é necessária.

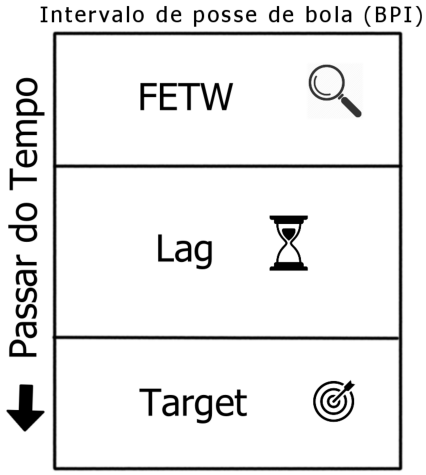


Figura 4.2: Divisão das janelas de posse de bola (BPI).

A distribuição do tamanho dos BPIs é apresentando na Figura 4.3, onde é possível observar que alguns BPIs possuem tamanho inferior a 500 *frames* (15 segundos). Um tamanho tão curto de BPI dificulta as análises, então foram filtrados BPIs com no mínimo 15 segundos de duração (500 *frames*), visto que as posses com intervalos menores não apresentaram uma movimentação expressiva dos jogadores em campo ou não possuíam nenhum passe realizado. Com esse critério de 15 segundos, foram selecionados 1.034 BPIs viáveis, sendo esta uma abordagem semelhante ao trabalho de Merlin (2020).

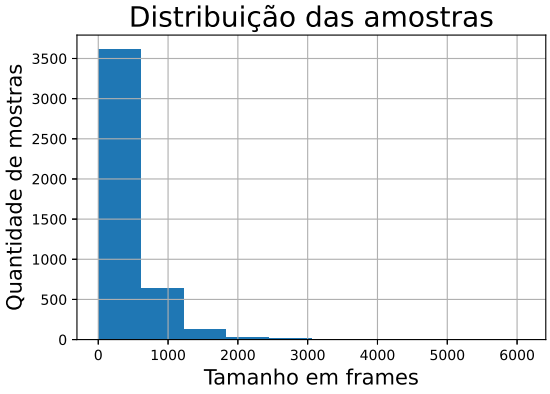
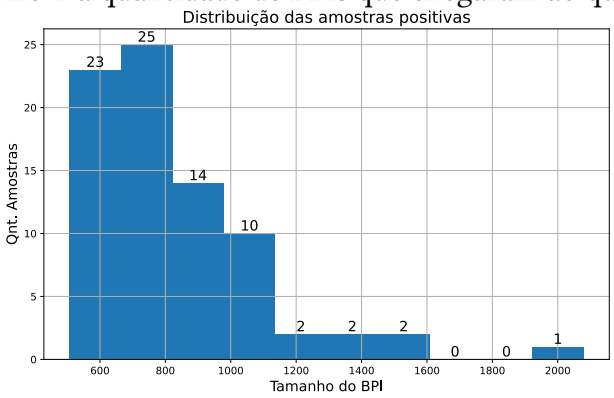


Figura 4.3: Distribuição dos intervalos de posse de bola por sua quantidade de *frames*, sendo o eixo X a quantidade de *frames* na posse e o eixo Y a quantidade de posses de bola.

Das 1.034 amostras selecionadas, somente 79 chegaram ao quarto final do campo, tornando assim esse conjunto de dados desbalanceado considerando que o objetivo é realização de uma classificação binaria. Outro ponto que precisa ser destacado é que os BPIs foram limitados somente com tamanho mínimo de *frames*. Dessa forma, as posses com intervalos maiores foram mantidas para análise. A distribuição das amostras positivas é apresentada na Figura 4.4.

Figura 4.4: Distribuição dos intervalos de posse de bola, onde o eixo X apresenta o tamanho do BPI em *frames* e o eixo Y a quantidade de BPIs que chegaram ao quarto final do campo.



4.3 Análise de Grafos

Para efetuar uma análise do posicionamento dos jogadores utilizando métricas de grafos, cada *frame* gerou um grafo para cada equipe (com e sem posse bola).

4.3.1 Construção do Grafo

Utilizamos a biblioteca `Network X` versão 2.6.2 implementada em Python (HAGBERG; SCHULT; SWART, 2008) para gerar, para cada frame da partida, um grafo $G = (V, A)$ em que o conjunto V são os jogadores e as arestas A são as possibilidades de passe entre os jogadores. As possibilidades de passe são criadas seguindo o conceito de triangulação de Delaunay. É dito que há uma possibilidade de passe entre dois jogadores se os critérios a seguir são satisfeitos:

- Não possuir nenhum jogador da mesma equipe entre os dois.
- Não possuir nenhum adversário com pelo menos 50 cm de distância dos dois jogadores.

Não foram descartadas possibilidades de passes entre jogadores quando um adversário estava entre eles, devido à representação dessa possibilidade (aresta do grafo) ser uma linha reta, o que não necessariamente seria a maneira que o passe seria realizado (ex. passe pode ocorrer com uma curvatura na trajetória da bola ou de forma aérea, em que o adversário não consegue interceptar a bola).

A partir dos grafos construídos na região FETW, foram extraídas as métricas citadas na Seção 2.1, permitindo caracterizar o comportamento das equipes em um momento da partida. Por exemplo, a Figura 4.5 exemplifica um grafo gerado a partir de um instante de tempo da

partida. O tamanho dos vértices representa a medida de centralidade de cada jogador, demonstrando a distribuição de centralidade no grafo em relação aos outros jogadores em campo.

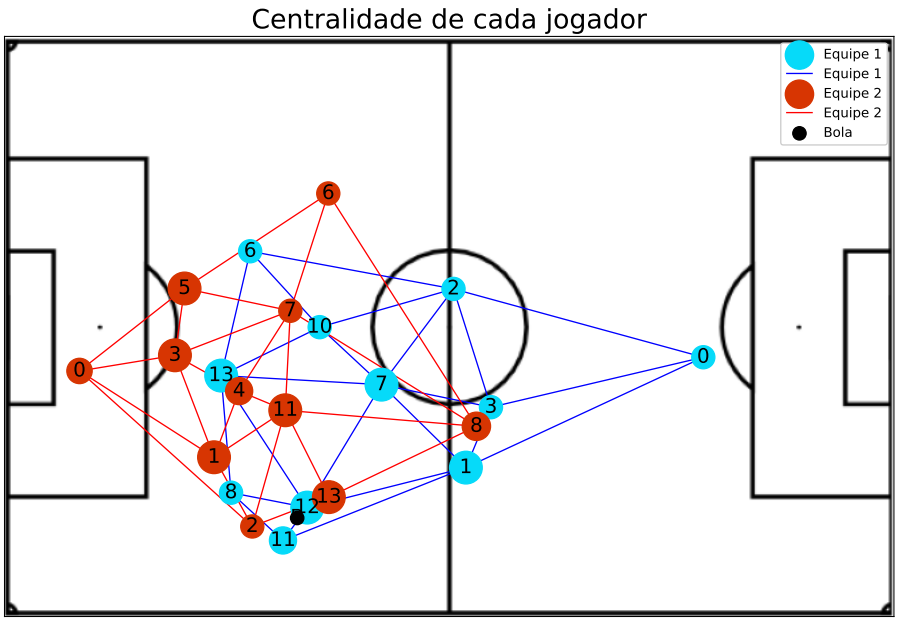


Figura 4.5: Demonstra como a centralidade dos jogadores pode ser observada em campo quando a partida é representada como um grafo construído conforme as especificações da seção 4.3.1. O tamanho dos vértices na imagem indica a proporção de sua centralidade em relação aos demais, onde quanto maior o vértice na imagem maior o seu valor de centralidade.

4.3.2 Métricas no Contexto do Futebol

Usando a definição de grafo apresentada, as métricas definidas na Seção 2.1 podem ser interpretadas da seguinte forma:

- **Centralidade de Intermediação**: demonstra a importância de um jogador na construção de jogadas, dado que quanto maior a sua centralidade, maior a sua capacidade para construção de jogadas e disponibilidade para receber passes.
- **Excentricidade**: representa a facilidade para que a bola chegue a um jogador considerando toda a rede.
- **Eficiência Global**: demonstra se os jogadores estão se posicionando de maneira que fiquem disponíveis para receber e realizar passes.

- **Eficiência Local**: assim como a eficiência global, demonstra a disponibilidade dos jogadores a participar de jogadas, porém considerando somente o seu impacto nos companheiros de equipe próximos.
- **Vulnerabilidade**: demonstra o impacto da ausência de um jogador para a equipe, onde valores negativos apresentam que a eficiência da equipe decai com a sua ausência.
- **Coeficiente de Clusterização**: demonstra o papel de um jogador para tornar possíveis triangulações de passes. Essas triangulações são importantes, pois permitem que a equipe se movimente com maior velocidade quando com a posse de bola.
- **Entropia**: representa a probabilidade de um jogador estar disponível para receber um passe dos companheiros de equipe próximos. Quanto maior o valor desta métrica, mais situações são possíveis em que esse jogador poderia participar da jogada.
- **PageRank**: demonstra o prestígio de um jogador para a construção de jogadas, ou seja, quanto maior o valor dessa métrica, maior a participação deste jogador em passes com outros jogadores que também possuem alta participação em jogadas.

4.3.3 Imagens de Ritmo Visual

A partir das medidas de grafo, geramos imagens no formato de ritmo visual, onde o eixo X representa cada *frame* do vídeo e o eixo Y cada jogador. A Figura 4.6 permite visualizar como as métricas extraídas do grafo podem ser representadas como ritmo visual, demonstrando a equivalência entre uma métrica de grafo e uma imagem de ritmo visual. Na Figura 4.6(a), a entropia dos jogadores 6 e 10 é mostrada em função do tempo, em segundos e, na Figura 4.6(b), esse mesmo trecho é representado na imagem de ritmo visual nas posições 6 e 10 do eixo Y. Os outros espaços do eixo Y são utilizados pelos outros jogadores da equipe em uma partida real, dado que a figura apresenta apenas os valores para os jogadores 6 e 10.

Computadas as métricas para todos os jogadores da equipe, foi criado um ritmo visual completo com 11 jogadores no eixo Y e 5 segundos no eixo X (Figura 4.7).

O processo de criação de imagens de ritmo visual descrito acima foi aplicado aos FETWs em BPIs com ao menos 15 segundos de duração, criando assim uma imagem para cada uma das 8 métricas descritas na Seção 2.1 para ambas equipes, totalizando 16 imagens de ritmo visual por BPI.

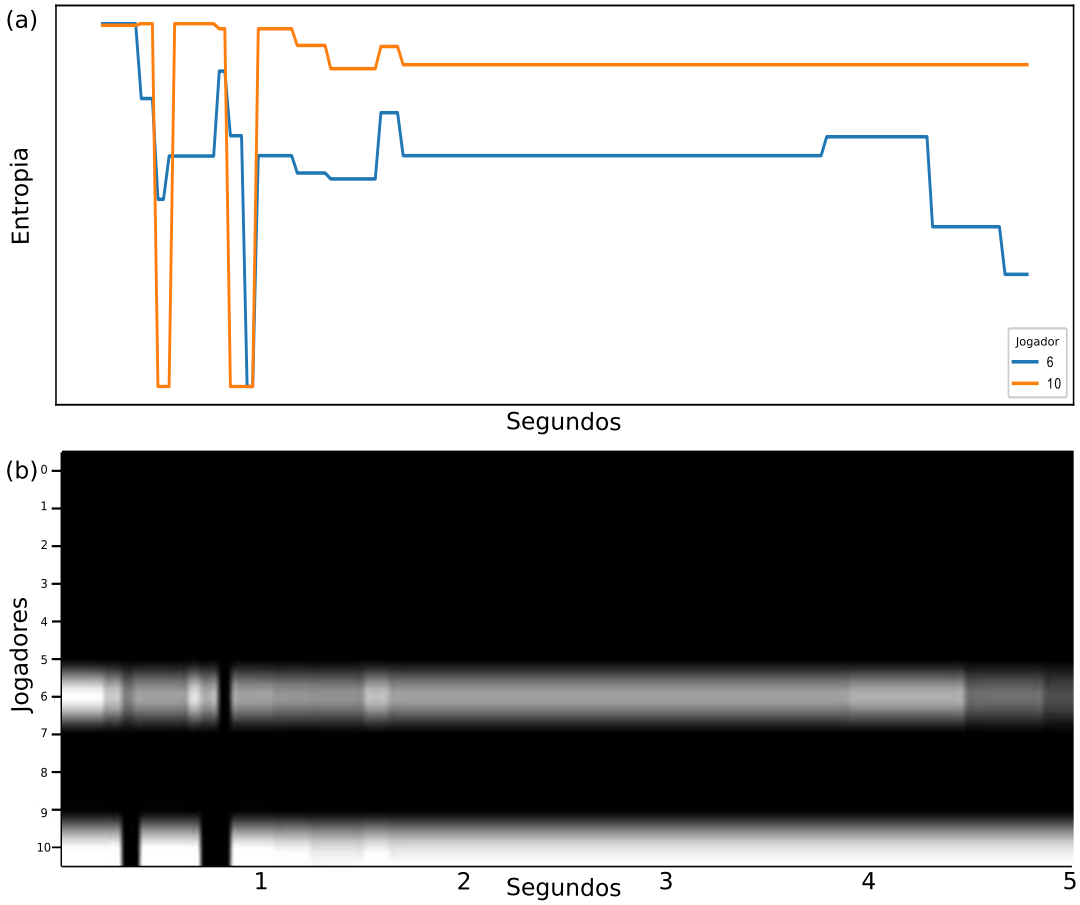


Figura 4.6: A figura demonstra como os valores de entropia para dois jogadores sofrem modificações com o passar do tempo, onde o eixo X representa os segundos da posse de bola e o Y o valor de entropia naquele momento, (b) apresenta os mesmos valores do item (a), porém em formato de ritmo visual, onde os tons claros representam valores mais elevados de entropia e os mais escuros valores menores.

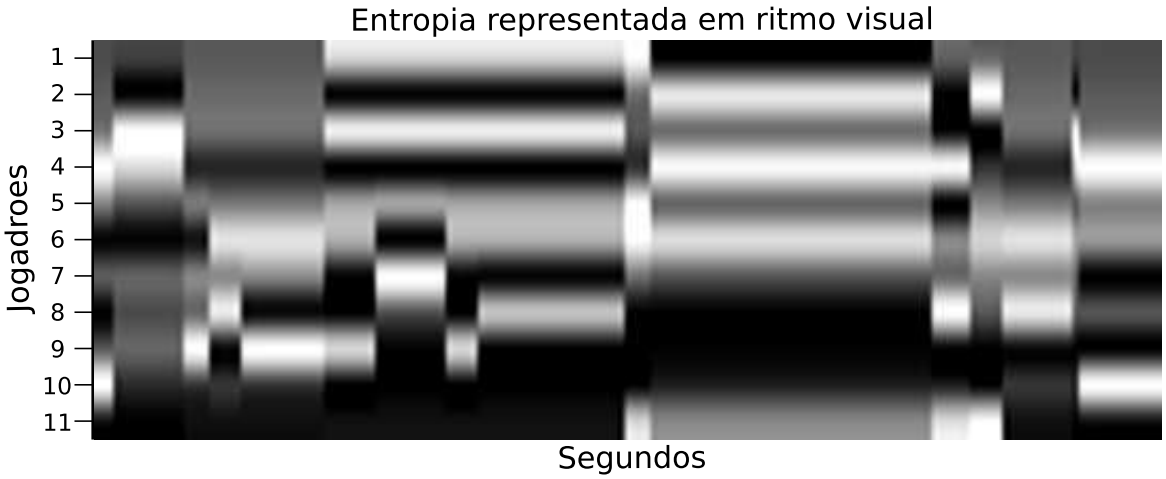


Figura 4.7: O eixo X representa o decorrer do tempo e o eixo Y os jogadores (valor da métrica de redes), os *pixels* com tons mais claros representam valores maiores para o jogador naquele instante de tempo, enquanto os tons mais escuros valores menores.

Inicialmente, foram gerados imagens de ritmo visual exclusivamente para a equipe com a posse de bola, porém os resultados obtidos com o treinamento da rede neural não foram satisfatórios. Dessa forma, foi decidido também considerar a equipe sem a posse de bola para a construção da entrada da rede.

4.3.4 Entrada da Rede Neural Convolucional

Para fornecer os ritmos visuais para a rede de extração de características (redes neurais convolucionais), as 16 imagens de cada BPI foram concatenadas em uma só, a escolha de considerar ambas as equipes foi utilizada devido existir estudos na literatura onde o comportamento da equipe atacante estar relacionado diretamente em como a equipe defensora está disposta em campo.

Como as imagens possuem somente uma camada, a sua resolução é de (11 x 167), onde os 11 pixels do eixo Y representam os jogadores e os 167 pixels do eixo X representam os segundos. A concatenação foi realizada em camadas mantendo a dimensão (11 x 167), porém agora com todas as métricas 'empilhadas' formando uma imagem multi camada de (11x167x16), como mostra a Figura 4.8.

4.4 Transferência de Aprendizado (*Transfer Learning*)

As imagens são fornecidas como entrada para uma rede profunda especializada em extrair caraterísticas. Em nossos experimentos, a computação utilizou a biblioteca Tensorflow (versão 2.3) na linguagem Python. A arquitetura de rede profunda utilizada foi a *EfficientNet B0* (TAN; LE, 2019) treinada com imagens do ImageNet (DENG et al., 2009) e sem as camadas totalmente conectadas colocadas no final da rede para classificação no problema original, sendo os pesos da ImageNet reajustados através do treinamento com os nossos dados através do ajuste fino.

A extração de características com essa arquitetura retorna um vetor de 1.280 números em ponto flutuante, posteriormente fornecidos como entrada para uma rede neural para realizar a classificação da amostra.

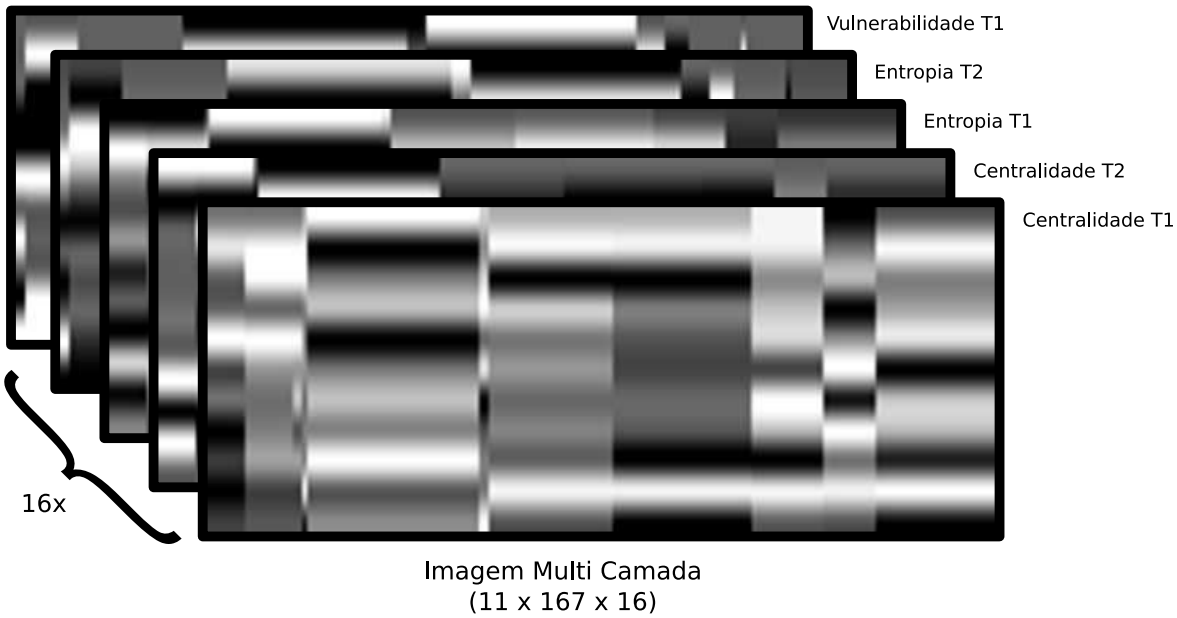


Figura 4.8: A figura apresenta uma amostra contendo as 8 métricas reunidas, sendo que cada métrica possui duas imagens de ritmo visual, uma representando a equipe com a posse de bola (atacante) e outra a equipe sem a posse de bola (defensora). Cada métrica foi representada por um canal na imagem, desta forma cada imagem tem 16 canais (um para cada uma das oito métricas de ambas as equipes). Assim, a imagem é representada com 11 pixels de altura (um para cada jogador), 167 pixels de largura (representando o tempo) e 16 canais (um para cada métrica). As métricas seguem a seguinte ordem nas camadas:(1) Centralidade, (2) Coeficiente de clusterização, (3) Excentricidade, (4) Entropia, (5) Eficiência global, (6) Eficiência local, (7) *PageRank* e (8) Vulnerabilidade.

4.5 Rede Neural Profunda

As imagens concatenadas com as 8 métricas são utilizadas como entrada para a rede, porém, antes de chegarem às camadas da *Efficientnet B0*, a entrada precisa passar por duas camadas convolucionais, que utilizam a função de ativação *ReLU* e possuem 16 e 32 filtros respectivamente.

Após essas duas camadas, foi acrescentada uma camada remodeladora (do inglês, *reshape*), alterando a posição dos valores da matriz de entrada para um formato 32x495x3. Após esse processamento, a imagem é recebida pela *EfficientNet B0* para a extração de características. Esse processo é necessário porque a implementação da *Efficientnet B0* necessita que a entrada possua ao menos a resolução de (32x32) com o máximo de 3 canais. Como esse não é o formato original da nossa imagem, foi necessário aplicarmos esse tratamento pré-extração. A extração resulta em um vetor de 1.280 pontos flutuantes para cada amostra fornecida na entrada, sendo esses dados a entrada para a parte classificadora da rede.

A classificação ocorre em quatro camadas densas totalmente conectadas com 1.280, 640, 320 e 160 neurônios cada respectivamente e com *Dropuout* entre elas de 0.5, 0.2 e 0.2. E a saída é dada por uma camada *Softmax* que devolve a probabilidade da amostra pertencer a determinada classe (no nosso caso, de ser um BPI que teve sucesso ou não).

Essa arquitetura pode ser observada na Figura 4.9, onde cada caixa representa uma camada na rede, exceto a caixa *EfficientNet B0*, que representa várias camadas convolucionais.

4.6 Otimização de Hiperparâmetros - *Grid Search*

O Processo de *Grid Search* permite testar diversos hiperparâmetros para o modelo. Dessa forma, é possível selecionar aqueles que melhor classificam os dados. Para que o processo funcione, os dados precisam ser divididos em treino, validação e teste, sendo muito recomendado uma metodologia baseada em *cross-validation*.

Cada partida de futebol possui dois tempos (primeiro e segundo). Dessa maneira, temos uma divisão lógica de vinte grupos de dados, dado que temos dez partidas. Desses vinte grupos contendo intervalos de posse de bola, seis grupos foram separados como dados de teste, em uma proporção de aproximadamente 70% treinamento e 30% teste.

Os 14 grupos restantes foram utilizados na busca de hiperparâmetro com o *grid-search*, separando os dados em treinamento e validação seguindo a abordagem 5-*fold*

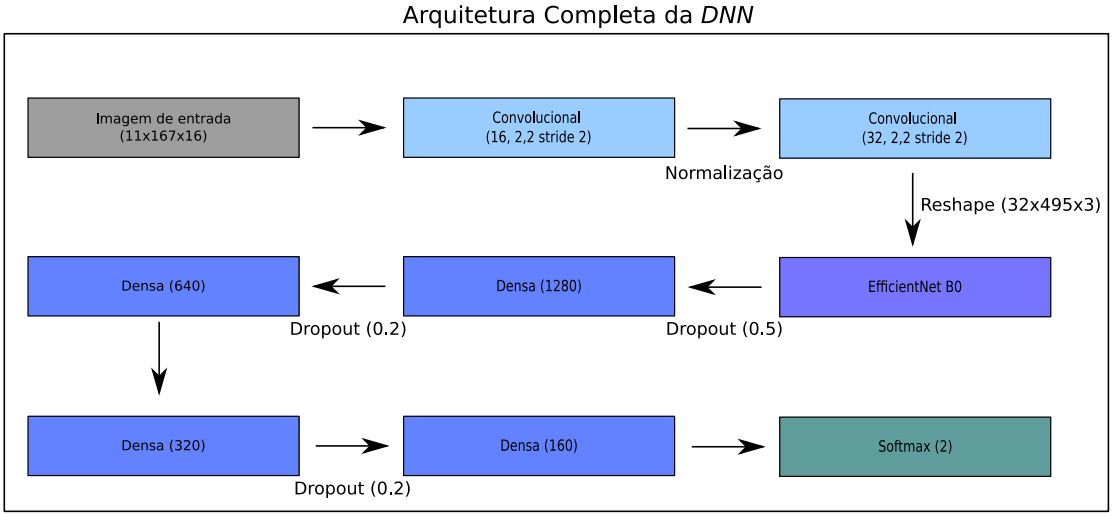


Figura 4.9: A figura apresenta a arquitetura completa da DNN, onde a entrada da rede é a imagem concatenada de todas as 8 métricas para às duas equipes (com e sem posse de bola), seguida de duas camadas convolucionais com filtros 2x2 e passo de 2 e uma camada de *reshape* para um formato aceito pela *EfficientNet B0* (32x495x3). Há também quatro camadas densas com *dropout* entre elas, e saída da rede é dada por uma camada *softmax* de dois neurônios.

cross-validation, onde para cada iteração os cinco grupos de dados foram revezados entre treinamento e validação. Como foi previamente feita uma separação de alguns dados de testes para serem utilizados após o *cross-validation*, assim reduzindo o ruído positivo introduzido pela otimização dos hiperparâmetros.

Toda a metodologia descrita acima foi dita uma rodada e, ao total, foram geradas 10 rodadas. Em outras palavras, foram executadas 10 rodadas com os 5 grupos de treinamento e validação para obtermos a média de acurácia nos testes. As 10 rodadas são importantes para, ao final do trabalho, computar testes estatísticos.

4.7 Análise da Contribuição das Características

Para analisar a contribuição das características da nossa rede, foi utilizada a biblioteca de visualização para Python (versão 3.8) SHAP (versão 0.40). Essa biblioteca foi desenvolvida para realizar a classificação dos atributos utilizando *Shapley values*, um conceito da teoria dos jogos que busca encontrar a importância de cada participante para o resultado (LUNDBERG; LEE, 2017). Nesse caso, a importância de cada métrica obtida no FETW para a classificação do BPI.

Neste trabalho, foi utilizado o método *Explainer* que implementa os *shapley values* como descrito na Seção 2, que de posse da entrada do modelo identifica como cada valor influencia

na saída do classificador. Foi selecionado esse método porque pode ser aplicado a qualquer tipo de classificador (rede neural, regressão, árvore de decisão, ...), sendo o único requisito que a saída do modelo seja "probabilista", ou seja, o resultado precisa ser a 'confiança' que o classificador possui de que a amostra pertence a determinada classe.

Como esse método permite somente explicar a decisão (gerando os SHAP values) de uma única amostra, para identificar quais as principais métricas de todo o conjunto de amostras positivas (aquelas em que a equipe atacante conseguiu chegar ao quarto final do campo), foram analisadas pela biblioteca cada uma delas individualmente e os resultados somados, assim demonstrando as métricas de uma forma que abrangesse todas as situações a que os classificadores foram expostos.

4.8 Resumo da Metodologia

A metodologia completa é composta pela extração das informações do campo para grafos, seguida da computação das métricas dos grafos gerados junto da sua representação em imagens de ritmo visual, sendo os resultados das equipes concatenados em uma única imagem por posse de bola (BPI).

Com as amostras geradas (ritmo visual das métricas de cada BPI concatenado), o próximo passo foi fornecê-las como entrada da *EfficientNet B0*, utilizada no processo de transferência de aprendizado e ajuste fino realizado junto ao treinamento de uma camada totalmente conectada para a classificação, sendo esse processo dividido em dez rodadas de treino e validação, realizando assim a otimização dos pesos da rede.

Com a rede treinada, os resultados foram analisados pela biblioteca SHAP, apresentando o impacto de cada *pixel* da imagem para a predição das classes.

Capítulo 5

Resultados

Com a DNN treinada no protocolo citado na metodologia, onde os dados foram divididos em 10 rodadas diferentes e em cada rodada as partidas de treino, validação e teste foram alternadas, permitindo assim comparar os resultados obtidos no contexto de partidas diferentes. Os resultados de acurácia balanceada obtidos podem ser observados na Tabela 5.1.

Tabela 5.1: Acurácia Balanceada da DNN em cada rodada de teste.

	Acurácia Balanceada (%) para cada Rodada									
Rodada	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Acurácia	74%	79%	52%	75%	77%	82%	67%	75%	89%	74%

Para confirmar que os resultados apresentados nas rodadas de teste não são apresentados devido ao acaso, foi criado um classificador aleatório como baseline, onde seus resultados apresentam 50% de chance de selecionar a classe correta para a amostra. Seu resultado é apresentado na Tabela 5.2 com um valor médio de 52%.

Tabela 5.2: Acurácia Balanceada do classificador aleatório em cada rodada de teste.

	Acurácia Balanceada (%) para cada Rodada									
Rodada	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Acurácia	50%	63%	58%	51%	50%	40%	51%	47%	44%	63%

Utilizando o teste não paramétrico de Wilcoxon, foi criada a hipótese nula *H0* que verifica se o resultado apresentado pela rede diferem de uma distribuição aleatória. O teste apresentou um valor de $p < 0.01$ assim sendo rejeitada a hipótese *H0*. Esse valor demonstra que os resultados da rede diferem para cada conjunto de treino, validação e teste não por acaso, mas sim devido

a diferenças do comportamento dos jogadores em cada partida. De forma geral, os resultados demonstram que *framework* consegue aprender os padrões de uma partida e generalizar para outras, visto que a acurácia balanceada média foi de 74%. A matriz de confusão apresentada na Tabela 5.3 apresenta a distribuição de acertos e erros da rede.

Matrizes de confusão para cada rodada de teste									
Rodada 1		Rodada 2		Rodada 3		Rodada 4		Rodada 5	
15	12	18	9	1	25	12	11	13	9
19	285	25	342	0	292	6	288	11	284
Rodada 6		Rodada 7		Rodada 8		Rodada 9		Rodada 10	
19	7	11	19	13	7	21	5	8	7
23	275	3	301	37	246	5	278	8	177

Tabela 5.3: Matriz de confusão para cada rodada de teste, onde para cada rodada é apresentado os valores de verdadeiro positivo e falso negativos na primeira linha. Na segunda linha são apresentados os dados de falso negativos e verdadeiros negativos.

Os resultados apresentados na Tabela 5.3 demonstram como foi o comportamento da rede nas dez rodadas de teste. É possível observar que as rodadas com uma acurácia menor (3 e 7) têm a sua matriz de confusão diferente das demais, possuindo uma tendência a classificar amostras positivas (que chegaram ao quarto final do campo) como negativas. Esse efeito provavelmente ocorreu devido à diferença entre as partidas de treino e teste, onde comportamentos singulares podem ter atrapalhado na generalização da rede.

Um ponto a se destacar é a quantidade de amostras para teste entre as rodadas diferirem. Essa particularidade ocorre devido aos conjuntos serem separados por partidas, onde algumas partidas possuem mais chegadas ao quarto final do campo do que outras, sendo esse outro ponto que reforçou a necessidade de repetir o experimento para confirmar os seus resultados.

5.1 Estudo dos Atributos

Estudamos a importância de cada atributo utilizando a biblioteca SHAP (LUNDBERG; LEE, 2017). As cores da Figura 5.1 definem a influência do pixel da imagem na classificação. A influência dos atributos é dada como positiva quando a sua utilização faz com que os resultados da rede se aproximem dos reais, e negativa quando o oposto disso ocorre, ou seja, a presença do atributo induz o classificador a classificar incorretamente a amostra.

Os tons de rosa representam que o pixel influenciou negativamente para a classificar a real classe da amostra, e os tons em azul representam que o pixel influenciou positivamente o classificador para prever o rótulo correto.

Foram analisados os principais atributos somente para as amostras de sucesso (BPIs que chegaram ao quarto final do campo), dado que o intuito deste trabalho é identificar quais métricas conseguem indicar o sucesso da posse de bola em chegar ao quarto final do campo. A Figura 5.1 demonstra os valores de cada pixel para uma amostra de sucesso e outra que não chegou ao quarto final do campo.

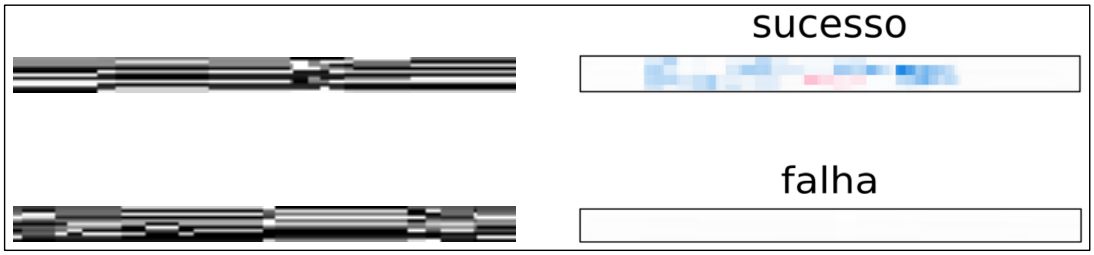


Figura 5.1: Representação do SHAP values sobre os pixels de uma amostra positiva (sucesso, chegou ao quarto final do campo) e uma amostra negativa (falha, não chegou ao quarto final do campo).

Com cada amostra positiva fornecida para o SHAP predizer a influência de cada região da imagem através dos shapleys values, esses resultados foram somados de forma absoluta (ignorando se eram positivos ou negativos), criando o total de shapley values das amostras positivas. Assim, para o valor total foi obtido o percentual que cada métrica possui do todo, assim ranqueando as métricas que possuem uma maior participação (percentual) como mais importantes. Esses valores podem ser observados em dois conjuntos, o total para a equipe com a posse de bola na Figura 5.2 e o total para a equipe sem a posse na Figura 5.3.

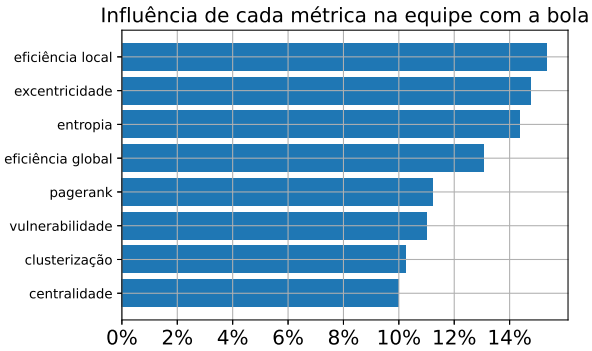


Figura 5.2: A imagem demonstra que as três principais métricas são: eficiência local, excentricidade e entropia.

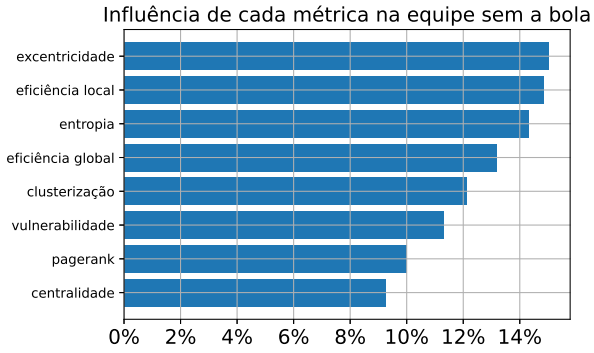


Figura 5.3: A imagem demonstra que as três principais métricas são: excentricidade, eficiência local, e entropia.

5.2 Discussões dos Resultados

Observando os resultados obtido pela biblioteca SHAP de ambas as equipes, é possível identificar que elas apresentam semelhanças nas porcentagens de participação de cada uma das métricas no total absoluto de SHAP values, sendo que todas elas apresentam ao menos 9% de participação.

As melhores métricas (top-3) para ambas as equipes foram: eficiência local, excentricidade e entropia, alternando somente o top-1 entre as equipes, demonstrando que essas métricas apresentam bons resultados para a análise de partidas de futebol para equipes com e sem a posse de bola.

A excentricidade indica o quão fácil é para a bola chegar até um jogador na rede, a escolha dessa métrica para ambas as equipes possuem diferentes interpretabilidades. Para a equipe com a posse de bola altos valores de excentricidade mostram a necessidade de o time estar separado, assim aumentando as opções para passes e jogadas ofensivas, como mostra a Figura 5.4. Já para o time que está defendendo, os mesmo valores altos de excentricidade ajudam a equipe que está atacando, dado que criam brechas entre os jogadores defensivos que podem ser aproveitadas pelos atacantes para chegar próximo a sua meta. Essas conclusões encontram embasamento nos trabalhos de Grund (2012), Arriaza-Ardiles et al. (2018) e Clemente, Sarmento e Aquino (2020).

Eficiência local apresentou grande influência para ambas as equipes. Essa métrica indica como a remoção de um vértice da rede impactaria na eficiência dos seus vizinhos. Valores elevados para essa métrica indicam que removendo esse jogador (por exemplo, quando ele está sobre marcação) os jogadores próximos (seus vizinhos na rede) perdem uma importante forma de receber a bola. Pela perspectiva da equipe com a posse de bola, essa informação

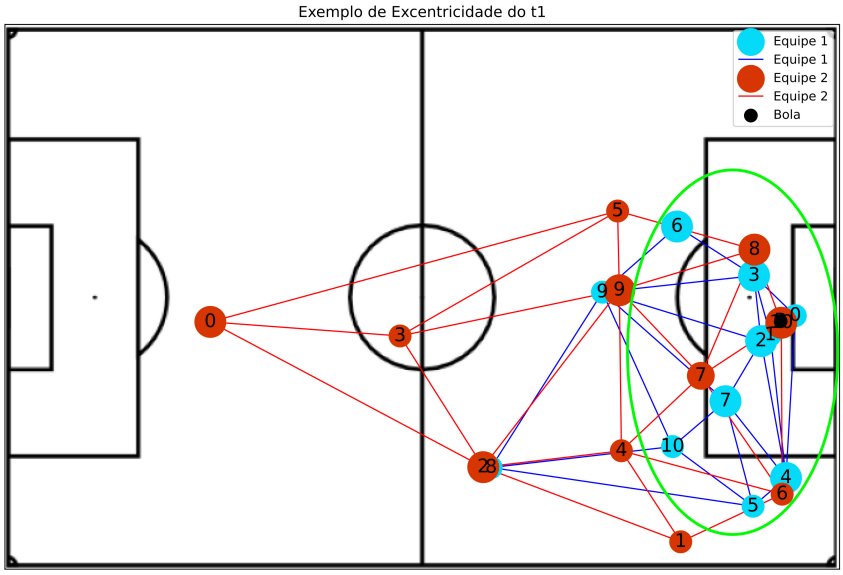


Figura 5.4: A imagem demonstra uma situação de alta excentricidade para a equipe t1 (com a posse de bola), onde é possível observar que os jogadores se encontram separados (porém com bastante opções de passes) na área em verde destacada na figura.

demonstra a necessidade de um "jogador central" responsável por construir as jogadas, como mostra a Figura 5.5. Quando considerada a equipe defensora, essa métrica nos mostra que quando a defesa se concentra em um único defensor, isso auxilia para que a equipe atacante seja bem sucedida em chegar na região de interesse.

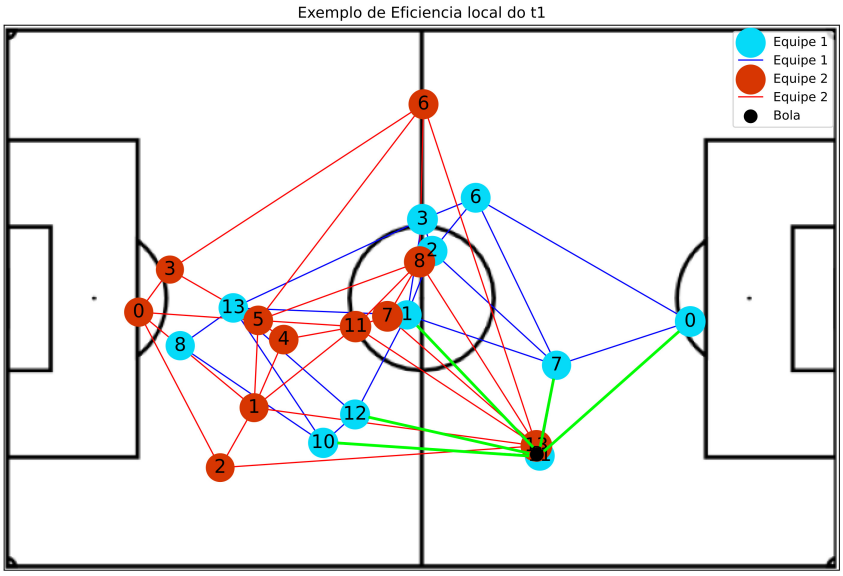


Figura 5.5: A imagem demonstra uma situação de alta eficiência local para o jogador 11, onde no momento que ele recebeu a bola possuía cinco opções de passes (destacadas em verde). Essa situação demonstra que, caso ele estivesse sobre marcação (impossibilitado de receber a bola), tal jogada não seria possível.

Nesse trabalho, a métrica de entropia representa a probabilidade de um jogador receber um passe a partir dos companheiros de equipe que estão próximos. Valores altos dessa métrica nos mostram que provavelmente o jogador estará livre de marcação, assim podendo receber passes e permitindo a criação de várias jogadas ofensivas, como mostra a Figura 5.6. Entretanto, quando analisada pelo ponto de vista da equipe defensora esses valores altos na métrica indicam que o defensor está isolado em campo (sem aplicar marcação a nenhum jogador ofensivo), assim criando aberturas por onde a equipe com a posse de bola pode chegar ao seu objetivo com maior facilidade, um exemplo dessa situação é apresentado na Figura 5.7.

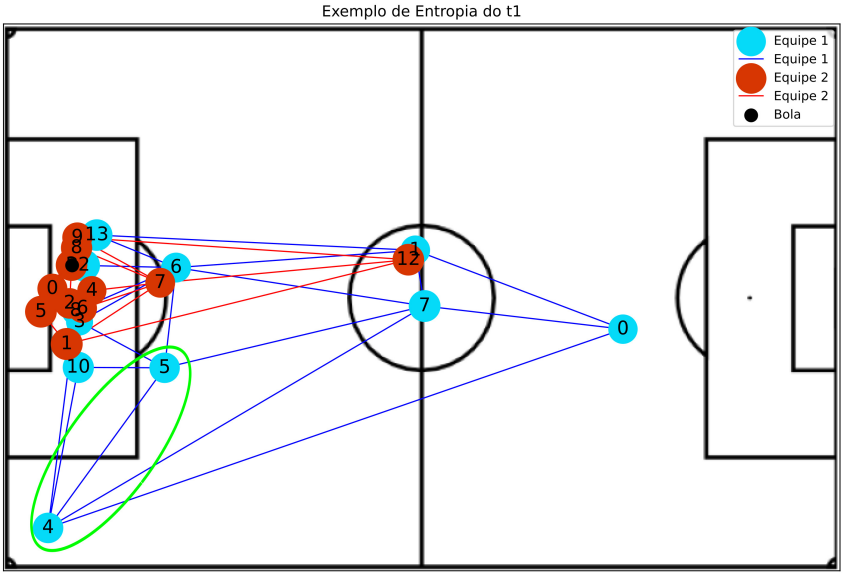


Figura 5.6: A imagem demonstra uma situação de alta entropia para a equipe t1 (com a posse de bola), onde é possível observar que os jogadores 5 e 4 (região em verde destacada na figura) se encontram livre de marcação e com várias opções de passes (arestas incidindo no vértice), assim possuindo uma maior chance de receber passes.

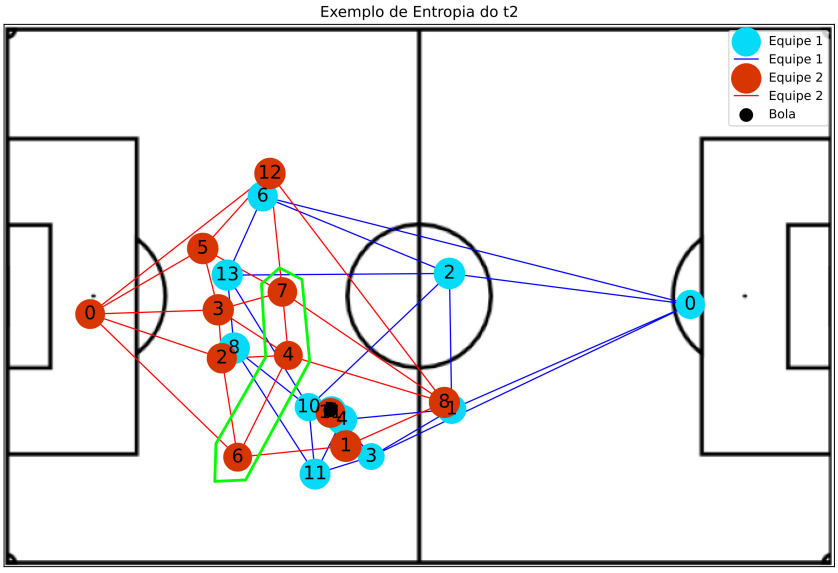


Figura 5.7: A imagem demonstra uma situação de alta entropia para a equipe t2 (sem a posse de bola), onde é possível observar que os jogadores 7, 4 e 5 (região em verde destacada na figura) se encontram distantes de jogadores da equipe com a posse, assim não realizando a marcação de nenhum adversário, dessa forma facilitando a realização de passes entre os jogadores adversários.

Capítulo 6

Conclusões

Esta dissertação no contexto de futebol analisou se um modelo baseado nos primeiros cinco segundos de posse de bola é capaz de prever se, ao final dessa posse, a bola estará no quarto final do campo. Os resultados permitem responder positivamente a esse objetivo, visto que, em média, o classificador apresentou uma acurácia balanceada de 74% para diferentes partidas.

Esta dissertação também analisou as características mais importantes utilizadas pelos melhores classificadores. Através da análise do resultado da rede com a biblioteca SHAP, foi observado que a eficiência local, excentricidade e entropia foram as principais métricas utilizadas do ponto de vista de ambas as equipes, e que essas métricas são condizentes com a literatura de análise de partidas de futebol.

Este trabalho deixa como ponto de estudo futuro a utilização desse *framework* de análise de posses de bola com métricas de grafos e transformação em ritmo visual para outros pontos de pesquisa, como, a classificação de passes, previsão de resultados e predição da qualidade de jogadas.

Referências bibliográficas

ALBERT, R.; BARABÁSI, A.-L. Statistical mechanics of complex networks. **Reviews of Modern Physics**, American Physical Society (APS), v. 74, n. 1, p. 47–97, jan. 2002. ISSN 1539-0756. DOI: 10.1103/revmodphys.74.47. Disponível em: <http://dx.doi.org/10.1103/RevModPhys.74.47>.

ARRIAZA-ARDILES, E. et al. Applying graphs and complex networks to football metric interpretation. **Human movement science**, Elsevier, v. 57, p. 236–243, 2018.

BABOOTA, R.; KAUR, H. Predictive analysis and modelling football results using machine learning approach for English Premier League. **International Journal of Forecasting**, Elsevier, v. 35, n. 2, p. 741–755, 2019.

BARREIRA, D.; GARGANTA, J.; MACHADO, J.; ANGUERA, M. T. Effects of ball recovery on top-level soccer attacking patterns of play. **Revista Brasileira de Cineantropometria & Desempenho Humano**, SciELO Brasil, v. 16, n. 1, p. 36–46, 2014.

BENGIO, Y.; GRANDVALET, Y. No unbiased estimator of the variance of k-fold cross-validation. **Journal of machine learning research**, v. 5, Sep, p. 1089–1105, 2004.

BERRAR, D. Cross-validation. **Encyclopedia of bioinformatics and computational biology**, Academic, v. 1, p. 542–545, 2019.

BICKEL, P. J.; LEVINA, E. Some theory for Fisher's linear discriminant function,naive Bayes', and some alternatives when there are many more variables than observations. **Bernoulli**, Bernoulli Society for Mathematical Statistics e Probability, v. 10, n. 6, p. 989–1010, 2004.

BRANDES, U. A faster algorithm for betweenness centrality. **The Journal of Mathematical Sociology**, Routledge, v. 25, n. 2, p. 163–177, 2001. DOI: 10.1080/0022250X.2001.9990249. eprint: https://doi.org/10.1080/0022250X.2001.9990249. Disponível em: <https://doi.org/10.1080/0022250X.2001.9990249>.

BRODERSEN, K. H.; ONG, C. S.; STEPHAN, K. E.; BUHMANN, J. M. The balanced accuracy and its posterior distribution. In: IEEE. 2010 20th international conference on pattern recognition. Istanbul, Turkey: IEEE, 2010. p. 3121–3124.

CARLING, C.; BLOOMFIELD, J.; NELSON, L.; REILLY, T. The role of motion analysis in elite soccer: Contemporary performance measurement techniques and work rate data. **Sports Medicine**, Adis Online, v. 38, n. 10, p. 389, 2012.

CBF. **CBF apresenta relatório sobre papel do futebol na economia do Brasil - Confederação Brasileira de Futebol**. São Paulo: Brasil: CBF, 14 dez. 2019. https://www.cbf.com.br/a-cbf/informes/index/cbf-apresenta-

relatorio-sobre-papel-do-futebol-na-economia-do-brasil. (Accessed on 09/21/2020).

CLEMENTE, F. M.; SARMENTO, H.; AQUINO, R. Player position relationships with centrality in the passing network of world cup soccer teams: Win/loss match comparisons. **Chaos, Solitons & Fractals**, Elsevier, v. 133, p. 109625, 2020.

CORDER, G. W.; FOREMAN, D. I. Nonparametric statistics for non-statisticians. John Wiley & Sons, Inc., 2011.

DENG, J. et al. Imagenet: A large-scale hierarchical image database. In: IEEE. 2009 IEEE conference on computer vision and pattern recognition. Miami, FL, USA: IEEE, 2009. p. 248–255.

DIAS, D. et al. Image-Based Time Series Representations for Pixelwise Eucalyptus Region Classification: A Comparative Study. **IEEE Geoscience and Remote Sensing Letters**, IEEE, 2019.

FERNANDEZ-NAVARRO, J. et al. Attacking and defensive styles of play in soccer: analysis of Spanish and English elite teams. **Journal of sports sciences**, Routledge, v. 34, n. 24, p. 2195–2204, 2016.

FRENCKEN, W.; LEMMINK, K. Team kinematics of small-sided soccer games: A systematic approach. In: 1°. London: Routledge, jan. 2009. p. 161–166. ISBN 0-415-42909-9.

FRENCKEN, W.; LEMMINK, K.; DELLEMAN, N.; VISSCHER, C. Oscillations of centroid position and surface area of soccer teams in small-sided games. **European Journal of Sport Science**, Taylor & Francis, v. 11, n. 4, p. 215–223, 2011.

GÉRON, A. **Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems**. Sebastopol, Ca: O'Reilly Media, 2019.

GRAYBILL, F.; IYER, H.; BURDICK, R. **Applied Statistics: A First Course in Inference**. New Jersey, United States: Prentice Hall, 1998. (Data Warehousing Institute Series from). ISBN 9780136214670. Disponível em: <https://books.google.com.br/books?id=Hqqfn3u8-T8C>.

GRUND, T. U. Network structure and team performance: The case of English Premier League soccer teams. **Social Networks**, Elsevier, v. 34, n. 4, p. 682–690, 2012.

HAGBERG, A. A.; SCHULT, D. A.; SWART, P. J. Exploring Network Structure, Dynamics, and Function using NetworkX. In: VAROQUAUX, G.; VAUGHT, T.; MILLMAN, J. (Ed.). **Proceedings of the 7th Python in Science Conference**. Pasadena, CA USA: SciPy, 2008. p. 11–15.

HAYKIN, S. **Redes neurais: príncipios e prática**. São Paulo, Brasil: Bookman Editora, 2007.

LANGVILLE, A.; MEYER, C. A Survey of Eigenvector Methods of Web Information Retrieval. **SIAM Review**, v. 47, jan. 2004. DOI: 10.1137/S0036144503424786.

LATORA, V.; MARCHIORI, M. Efficient Behavior of Small-World Networks. **Phys. Rev. Lett.**, American Physical Society, v. 87, p. 198701, 19 out. 2001. DOI: 10.1103/PhysRevLett.87.198701. Disponível em: <https://link.aps.org/doi/10.1103/PhysRevLett.87.198701>.

LEITE DE BARROS, R. M.; GUEDES RUSSOMANNO, T.; BRENZIKOFER, R.; JOVINO FIGUEROA, P. A method to synchronise video cameras using the audio band. **Journal of Biomechanics**, v. 39, n. 4, p. 776–780, 2006. ISSN 0021-9290. DOI: https://doi.org/10.1016/j.jbiomech.2004.12.025. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0021929005000400>.

LUCEY, P. et al. quality vs quantity: Improved shot prediction in soccer using strategic features from spatiotemporal data. In: PROC. 8th annual mit sloan sports analytics conference. Pittsburgh, PA, USA: Disney Research, 2014. p. 1–9.

LUNDBERG, S. M.; LEE, S.-I. A Unified Approach to Interpreting Model Predictions. In: GUYON, I. et al. (Ed.). **Advances in Neural Information Processing Systems 30**. Long Beach, CA, USA: Curran Associates, Inc., 2017. p. 4765–4774. Disponível em: <http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf>.

MALTA, P.; TRAVASSOS, B. Caraterização da transição defesa-ataque de uma equipa de Futebol. **Motricidade**, Revista Motricidade/Edições Desafio Singular, v. 10, n. 1, p. 27–37, 2014.

MARCHIORI, M.; VECCHI, M. de. Secrets of soccer: Neural network flows and game performance. **Computers & Electrical Engineering**, Elsevier, v. 81, p. 106505, 2020.

MERLIN, M. **New Approach To Analyze The Passing In Soccer Matches Using Multivariate And Machine Learning Techniques**. Nov. 2020. Tese (Doutorado) – UNIVERSIDADE ESTADUAL DE CAMPINAS, Faculdade de Educação Física, Campinas, SP.

METZ, J. et al. Redes complexas: conceitos e aplicações. São Carlos, SP, Brasil., 2007.

MOURA, F. A. et al. Análise quantitativa da distribuição de jogadores de futebol em campo durante jogos oficiais, 2011.

MOURA, F. A. et al. Analysis of the shots to goal strategies of first division brazilian professional soccer teams. In: ISBS-CONFERENCE Proceedings Archive. Ouro Preto, Brasil: ISBS, 2007.

MUNDRY, R.; FISCHER, J. Use of statistical programs for nonparametric tests of small samples often leads to incorrect Pvalues: examples from animal behaviour. **Animal behaviour**, Elsevier Science, v. 56, n. 1, p. 256–259, 1998.

MYERSON, R. **Game Theory: Analysis of Conflict**. Cambridge, Massachusetts, United States: Harvard University Press, 2013. ISBN 9780674728622. Disponível em: <https://books.google.com.br/books?id=oGUET9JBytEC>.

NOH, J. D.; RIEGER, H. Random walks on complex networks. **Physical review letters**, APS, v. 92, n. 11, p. 118701, 2004.

PAN, S. J.; YANG, Q. A Survey on Transfer Learning. **IEEE Transactions on Knowledge and Data Engineering**, v. 22, n. 10, p. 1345–1359, 2010. DOI: 10.1109/TKDE.2009.191.

POWERS, D. Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation. **Mach. Learn. Technol.**, v. 2, jan. 2008.

RODRIGUES, D. C. U. M.; MOURA, F. A.; CUNHA, S. A.; TORRES, R. d. S. Graph visual rhythms in temporal network analyses. **Graphical Models**, Elsevier, v. 103, p. 101021, 2019.

RUDER, S. An overview of gradient descent optimization algorithms. **arXiv preprint arXiv:1609.04747**, 2016.

RUIZ-RUIZ, C.; FRADUA, L.; FERNÁNDEZ-GARCÍA, Á.; ZUBILLAGA, A. Analysis of entries into the penalty area as a performance indicator in soccer. **European Journal of Sport Science**, Taylor & Francis, v. 13, n. 3, p. 241–248, 2013.

SAMUEL, A. L. Some studies in machine learning using the game of checkers. **IBM Journal of research and development**, IBM, v. 3, n. 3, p. 210–229, 1959.

SARAMÄKI, J. et al. Generalizations of the clustering coefficient to weighted complex networks. **Phys. Rev. E**, American Physical Society, v. 75, p. 027105, 2 fev. 2007. DOI: 10.1103/PhysRevE.75.027105. Disponível em: <https://link.aps.org/doi/10.1103/PhysRevE.75.027105>.

SHANNON, C. E. A mathematical theory of communication. **The Bell system technical journal**, Nokia Bell Labs, v. 27, n. 3, p. 379–423, 1948.

SHAPLEY, L. S. **Notes on the N-Person Game — II: The Value of an N-Person Game**. Santa Monica, CA: RAND Corporation, 1951. DOI: 10.7249/RM0670.

TAN, M.; LE, Q. Efficientnet: Rethinking model scaling for convolutional neural networks. In: PMLR. INTERNATIONAL Conference on Machine Learning. Long Beach, California: Proceedings of Machine Learning Research, 2019. p. 6105–6114.

TENGA, A.; HOLME, I.; RONGLAN, L. T.; BAHR, R. Effect of playing tactics on goal scoring in Norwegian professional soccer. **Journal of Sports Sciences**, Routledge, v. 28, n. 3, p. 237–244, 2010.

TRAVENÇOLO, B.; DA F. COSTA, L. Accessibility in complex networks. **Physics Letters A**, v. 373, n. 1, p. 89–95, 2008. ISSN 0375-9601. DOI: https://doi.org/10.1016/j.physleta.2008.10.069. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0375960108015867>.

VILAR, L.; ARAÚJO, D.; DAVIDS, K.; BAR-YAM, Y. Science of winning soccer: Emergent pattern-forming dynamics in association football. **Journal of systems science and complexity**, Springer, v. 26, n. 1, p. 73–84, 2013.

WRIGHT, C. et al. Factors associated with goals and goal scoring opportunities in professional soccer. **International Journal of Performance Analysis in Sport**, Taylor & Francis, v. 11, n. 3, p. 438–449, 2011.