Modelagem de Previsão e Valoração de Ações no Futebol baseada em Arquétipos

1 Introdução

Sport Analytics (SA) é uma área de estudo interdisciplinar em que técnicas computacionais de aprendizado de máquina, ciência de dados e estatística são aplicadas a problemas quantitativos no âmbito esportivo, o que vem gerando crescente interesse acadêmico e de mercado. Isso se nota no exterior a partir de, por exemplo, contribuições entre DeepMind e Livepool FC (16) e crescimento do SA Sloan no MIT. No Brasil, diversos clubes de futebol iniciaram departamentos de *analytics* recentemente, como Palmeiras, Atlético-MG, Red Bull Bragantino, entre outros. Além disso, foi fundado o Sports Analytics Lab (SALab) na UFMG, o qual desenvolve pesquisas na área de SA, além de organizar o evento Football Analytics Modeling and Experience (FAME), que terá sua terceira edição em 2024.

O uso de técnicas computacionais vem ganhando espaço em diversos setores de esportes, como prevenção de lesões(10), previsão de sucesso de jogadas[1] e agrupamento de atletas(6). Dentre os desafios recentes na área de SA, destaca-se o desenvolvimento de métodos de previsão e valoração de jogadas. A partir de dados históricos que representam jogos de forma contínua, busca-se aprender métricas para prever quais são as decisões mais prováveis de serem tomadas pelos jogadores em campo, além de atribuir valor às possíveis decisões. Este tipo de modelagem ajuda a metrificar criatividade e tomada de decisão, qualidades normalmente avaliadas apenas por meio de análise de vídeo. Atribuir valor a ações que não necessariamente geram gol é uma tarefa particularmente difícil no futebol por se tratar de um esporte de pontuação baixa, em que o volume de estatísticas simples é pequeno (ao contrário do basquete, por exemplo, em que cestas, assistências e rebotes são anotados em alto volume).

Métodos assim vêm sendo desenvolvidos com o advento de dados de *tracking* óptico de alta frequência, que fornecem a posição de todos os jogadores e da bola em todos os instantes de uma partida. Kovalchik destaca, em sua revisão(8), que, "nas últimas décadas, a proliferação de sistemas de dados de *tracking* nos esportes tem criado oportunidades empolgantes para pesquisa", uma vez que há uma crescente qualidade na descrição de padrões espaciais.

Apesar da alta qualidade de dados de *tracking*, a maioria dos modelos de *machine learning* em SA opta pela não inclusão de características individuais dos jogadores. Por isso, Kovalchik também destaca que a caracterização é um dos principais temas dentre os desafios metodológicos para novas pesquisas com dados de *tracking*. Em contrapartida, tal caracterização de atletas a nível individual é bem explorada no desenvolvimento de arquétipos, que são agrupamentos de jogadores de acordo com dados que caracterizem suas tendências de tomada de decisão e a qualidade de sua execução. Os arquétipos costumam ser utilizados quase exclusivamente no recrutamento de atletas, por serem facilmente interpretáveis. Todavia, raramente são incorporados a modelos mais complexos de *machine learning*.

Neste mestrado, portanto, serão unidas a criação de arquétipos de jogadores a modelos

[1] `https://arxiv.org/abs/2206.07212`

de previsão de jogadas e valoração de ações. A adição de caracterização individual contribui para o desempenho de analistas do esporte ao permitir novas aplicações com maior interpretabilidade. Um exemplo simplificado do impacto desta mudança em um modelo de previsão de jogadas é mostrado na Figura 1.

2 Referencial Teórico

Dados de *tracking* se mostraram particularmente úteis para atribuição de valor a jogadas e previsão de ações(8). Contudo, há diversas formas de fazê-lo: multi-modelos estocásticos(1), arquiteturas de aprendizado profundo(14; 5) e redes neurais de grafos (GNNs)(16; 11). O uso de GNNs é particularmente promissor na área de SA devido à facilidade de traduzir elementos do esporte em representações por grafos, além de permitir múltiplas tarefas a partir de uma mesma arquitetura: classificação de nós, grafos e arestas, além de regressões.

Apesar dos avanços, avaliar de modelos baseados em *tracking* é uma grande dor para SA, uma vez que há poucos problemas com *benchmarks* estabelecidos e diferentes autores podem avaliar modelos similares de forma diferente(5; 3), tornando difícil a comparação de desempenho. Uma forma de contornar esse problema é a validação por especialistas, como realizada no projeto TacticAI(16). Os autores apresentaram situações de jogo a analistas esportivos para que eles previssem a jogada seguinte; em seguida, compararam estatisticamente as respostas dos especialistas com os resultados do modelo, atestando sua capacidade de produzir resultados factíveis e úteis para um clube.

Outro ponto de questionamento em modelos de atribuição de valor a jogadas é a não inclusão da individualidade dos jogadores na forma de variáveis (como em (5)). O anonimato tem como objetivo criar um modelo generalista que descreve um jogador médio. Hoje, porém, é conhecido que há enviesamentos apesar do anonimato(2). Por exemplo, atacantes finalizam mais e melhor do que zagueiros, o que torna enviesado um jogador "médio" em um modelo de chutes. Ao adicionar individualidades ao modelo, abre-se mão do anonimato, mas ganha-se em interpretabilidade e aplicações.

Representações de jogadores em um novo espaço vetorial (*embeddings*) já foram usadas no auxílio de treinamento de redes neurais para avaliação de posses no basquete (14). O uso de *embeddings* no treinamento, embora adicione individualidades dos jogadores ao modelo, é uma técnica pouco interpretável. Em contrapartida, diversos autores pesquisaram formas de agrupar jogadores em arquétipos por meio de clusterização, seja definindo novas posições dadas as mudanças táticas do basquete (9), agrupando futebolistas semelhantes para auxiliar no recrutamento de jogadores jovens (6), entre outros. Na maioria das vezes, arquétipos são criados como um fim em si mesmo, como para *scouting*. Assim, nota-se que o desafio na criação dos arquétipos está no *trade-off* entre abstração e interpretabilidade, na busca por modelos precisos e explicáveis.

Conclui-se que a área de criação de arquétipos se desenvolveu distante das áreas de atribuição de valor e previsão de jogadas. Estas, contudo, não priorizaram a inclusão de individualidades de cada atleta em seus modelos. Este projeto, portanto, explorará este gap na literatura ao unir estas técnicas com maior interpretabilidade.

3 Metodologia

Os passos para execução da tarefa serão baseados na metodologia CRISP-DM, muito utilizada em tarefas de ciência de dados(12). As etapas são mostradas em alto nível na Figura

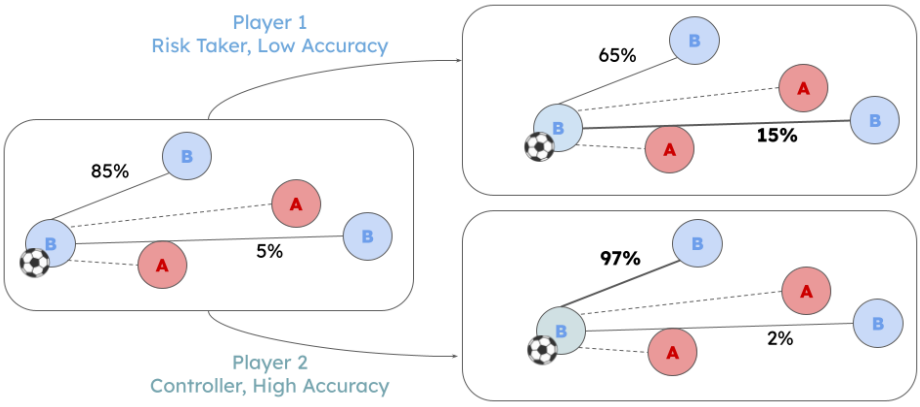

Figura 1: Neste exemplo simplificado, temos a probabilidade de passe completo do portador da bola para dois companheiros. A imagem da esquerda representa uma modelagem generalista, em que o indivíduo não é considerado. Na direita, é possível perceber que os arquétipos aos quais diferentes atletas pertencem altera as probabilidades de um modelo - no caso, um jogador mais avesso ao risco do que outro.

2, na qual se diferenciam tarefas individuais e tarefas que serão beneficiadas pela opinião de especialistas.

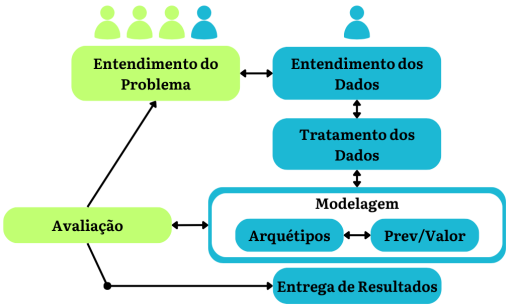

Figura 2: Metodologia adaptada da CRISP-DM.

Na fase de entendimento do problema, a revisão de literatura será aprofundada. Além disso, para melhor compreender as lacunas e as necessidades dos profissionais do esporte, serão consultados especialistas do mercado e da academia.

Em seguida, é preciso compreender os dados. O SALab, em parceria com a fornecedora de dados PFF, disponibiliza para seus integrantes grandes quantidades de dados de *tracking*, necessários para elaboração do projeto. Portanto, é preciso compreender a natureza dos *datasets*, suas limitações e desafios associados para garantir que não sejam estabelecidas metas pouco factíveis ou que não explorem todo o potencial.

Uma vez compreendidos, os dados precisam ser tratados de modo que sejam mais facilmente usados na modelagem. Esta, por sua vez, envolve a escolha de técnicas adequadas para cada uma das sub-tarefas: arquétipos e previsão/valoração de jogadas. Serão criadas arquiteturas capazes de unirem as duas técnicas, de modo a cumprir com o objetivo do projeto.

Os modelos serão avaliados de duas formas: por meio de métricas de perda convencionais e por validação com especialistas. Utilizando de parcerias e contatos do SALab com analistas de desempenho e outros profissionais em diferentes clubes de futebol do Brasil, pretende-se realizar validação estatística dos resultados dos modelos, bem como dos arquétipos em si.

Uma vez que bons resultados forem obtidos, será iniciada a escrita da dissertação. Além do texto, outras entregas incluem códigos estruturados para futuras pesquisas e aplicações, além de publicações em con-

ferências tanto na indústria quanto na academia, como KDD, KDMile e *workshops* de SA.

4 Cronograma

O diagrama na Figura 3 mostra o cronograma de tarefas a serem realizadas. Quanto às disciplinas, planejo expandir os conhecimentos em ciência de dados e aprendizado de máquina, para que sirvam de suporte no desenvolvimento da dissertação. Quanto ao estágio em docência, planejo ser monitor e co-lecionar a disciplina de Ciência de Dados Aplicada ao Futebol, dada organização vinculada ao SALab e sua proximidade com o tema proposto.

No primeiro semestre, serão realizadas as etapas de entendimento apresentadas na metodologia. Também é preciso definir conceitos a serem utilizados na criação de arquétipos baseados em comportamento dos jogadores. Dentre elas, se destacam: comportamento, decisão, influência, posse e ação. É preciso defini-las com rigor para a realização do projeto, embora pareçam triviais.

Uma vez estabelecidas, estas definições serão usadas para obtenção de informações que caracterizem jogadores, por meio do tratamento de dados. Em seguida, algoritmos de agrupamento como DBSCAN(4) e MiniBatchKMeans(13) serão testados para a criação dos arquétipos.

Concomitantemente, serão desenvolvidos os modelos de previsão e valoração de jogadas, utilizando redes neurais baseadas em grafos (GNNs). Seu desenvolvimento será um processo iterativo, buscando incorporar os arquétipos da melhor forma possível, de modo a minimizar funções de custo como Entropia Cruzada Binária ou Multi-classe(7). Dentre as arquiteturas possíveis, então as *Graph Attention Networks*(15), as redes convolucionais *CrystalConv* (11), entre outras.

Com a escrita da dissertação e a validação estatística das opiniões de especialistas, a metodologia será consolidada, garantindo o cumprimento das metas estabelecidas.

Referências

[1] D. Cervone, A. D'Amour, L. Bornn, and K. Goldsberry, *A multiresolution stochastic process model for predicting basketball possession outcomes*, Journal of the American Statistical Association, 111 (2016), p. 585–599.

[2] J. Davis and P. Robberechts, *Biases in expected goals models confound finishing ability*, 2024. URL:`https://arxiv.org/abs/2401.09940`.

[3] T. Decroos, L. Bransen, J. Van Haaren, and J. Davis, *Vaep: An objective approach to valuing on-the-ball actions in soccer (extended abstract)*, in Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-PRICAI-2020, IJCAI, July 2020.

[4] M. Ester, H. P. Kriegel, J. Sander, and X. Xu, *A density-based algorithm for discovering clusters a density-based algorithm for discovering clusters in large spatial databases with noise*, Proceedings - 2nd International Conference on Knowledge Discovery and Data Mining, KDD 1996, (1996).

[5] J. Fernández, L. Bornn, and D. Cervone, *A framework for the fine-grained evaluation of the instantaneous expected value of soccer possessions*, Machine Learning, 110 (2021), p. 1389–1427.

[6] M. Imburgio and S. Goldberg, *Introducing davies: A framework for*

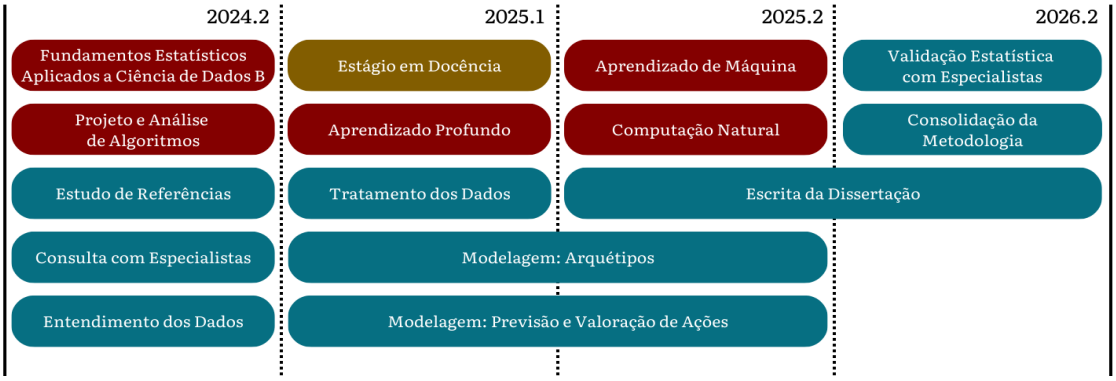

Figura 3: Cronograma com disciplinas e atividades da dissertação.

identifying talent across the globe, Sept. 2020. URL:https://www.americansocceranalysis.com/home/2020/9/16/davies-determining-added-value-of-individual-effectiveness-including-style.

[7] M. R. Izadi, Y. Fang, R. Stevenson, and L. Lin, *Optimization of graph neural networks with natural gradient descent*, 2020. URL:https://arxiv.org/abs/2008.09624.

[8] S. A. Kovalchik, *Player tracking data in sports*, Annual Review of Statistics and Its Application, 10 (2023), p. 677–697.

[9] A. B. Macedo, *Machine learning uncovers nine distinct player types in the nba*, Mar. 2021. URL:https://www.samford.edu/sports-analytics/fans/2023/Machine-Learning-Uncovers-Nine-Distinct-Player-Types-in-the-NBA.

[10] I. Ruiz-Pérez and et al., *A field-based approach to determine soft tissue injury risk in elite futsal using novel machine learning techniques*, Front. Psychol., 12 (2021), p. 610210.

[11] A. Sahasrabudhe and J. Bekkers, *A graph neural network deep-dive into successful counterattacks*, Mar. 2023. URL:https://ussf-ssac-23-soccer-gnn.s3.us-east-2.amazonaws.com/public/Sahasrabudhe_Bekkers_SSAC23.pdf.

[12] C. Schröer, F. Kruse, and J. M. Gómez, *A systematic literature review on applying crisp-dm process model*, Procedia Computer Science, 181 (2021), p. 526–534.

[13] D. Sculley, *Web-scale k-means clustering*, Proceedings of the 19th International Conference on World Wide Web, WWW '10, (2010).

[14] A. Sicilia, K. Pelechrinis, and K. Goldsberry, *Deephoops: Evaluating micro-actions in basketball using deep feature representations of spatio-temporal data*, 2019. URL:https://arxiv.org/abs/1902.08081.

[15] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, *Graph attention networks*, 2017. URL:https://arxiv.org/abs/1710.10903.

[16] Z. Wang, P. Veličković, and et al., *Tacticai: an ai assistant for football tactics*, Nature Communications, 15 (2024).