Relatório Técnico: Análise de Perfis de Jogadores (FIFA 26)

Autor: Enzo Araújo

Contexto: Disciplina de Aprendizado de Máquina Não-Supervisionado (IMD3003)

Introdução e Contextualização dos Dados

Este relatório apresenta o desenvolvimento e os resultados da análise de agrupamento aplicada ao dataset *"EAFC26 Mens Player Data Analysis Modeling"*. O objetivo central do estudo foi identificar padrões latentes e perfis táticos de jogadores de futebol através de técnicas de aprendizado de máquina não-supervisionado. A base de dados, composta por aproximadamente 16.000 instâncias, simula atributos projetados para o futuro jogo FIFA 26, combinando dados demográficos com métricas de desempenho técnico e físico. A complexidade do domínio exigiu uma abordagem capaz de distinguir nuances sutis entre posições e estilos de jogo que não são explicitamente rotulados nos dados brutos.

Pré-processamento e Engenharia de Atributos

A qualidade da clusterização dependeu diretamente de um *pipeline* rigoroso de tratamento de dados. Inicialmente, procedeu-se à limpeza do dataset, removendo colunas irrelevantes para a modelagem matemática, como identificadores únicos e URLs, e convertendo unidades físicas (altura e peso) para formatos numéricos padronizados. Um ponto crítico foi o tratamento de valores nulos nos atributos de goleiros para jogadores de linha, que foram preenchidos logicamente com zero.

Para enriquecer a capacidade discriminativa dos algoritmos, aplicou-se engenharia de atributos (*feature engineering*). Foram criadas *flags* binárias baseadas em razões estatísticas para capturar tendências de comportamento, como a variável `offensive` (para jogadores cujos atributos de ataque superam a defesa em margens significativas) e `all_around` (para atletas versáteis). Por fim, todas as variáveis numéricas foram submetidas ao `StandardScaler`, garantindo que a disparidade de escalas entre atributos como "salário" e "habilidade de drible" não enviesasse o cálculo de distâncias.

Metodologia de Modelagem e Visualização

A estratégia de modelagem seguiu uma abordagem comparativa. Inicialmente, utilizou-se o algoritmo **K-Means** como *baseline* para validar a existência de grupos coesos, utilizando o método do cotovelo na soma dos erros quadráticos (Inércia) para estimar o número ideal de clusters. No entanto, dada a natureza dos dados de jogadores, onde as fronteiras entre posições são frequentemente difusas, o modelo final adotado foi o **GMM (Gaussian Mixture Models)**. A escolha do GMM justificou-se pela sua flexibilidade em modelar clusters com diferentes variâncias e formas elípticas, ao contrário das esferas rígidas do K-Means. A seleção de modelos baseada no critério de informação BIC indicou uma estrutura ótima de 7 componentes (clusters).

Para a interpretação visual dos resultados em alta dimensionalidade, empregaram-se técnicas de projeção. Embora o PCA tenha sido usado para análise de variância global, a visualização final foi construída utilizando **UMAP (Uniform Manifold Approximation and Projection)**. Configurado com `n_neighbors=50` e distância mínima de `0.15`, o UMAP foi capaz de preservar a estrutura local dos dados, gerando uma projeção 3D que evidenciou separações nítidas entre grupos táticos.

Interpretação dos Padrões Encontrados

A análise dos centróides resultantes revelou que o algoritmo foi capaz de transcender as posições tradicionais, agrupando jogadores por função tática. Os sete perfis identificados demonstram comportamentos estatísticos distintos:

O **Cluster 4** isolou geometricamente os **Goleiros**, cujos atributos específicos (como reflexos e manejo) apresentaram valores cerca de 800% acima da média global. No setor ofensivo, o **Cluster 1** agrupou os **Finalizadores de Elite**, atacantes com foco exclusivo em finalização e voleios, negligenciando tarefas defensivas. Já o **Cluster 5** capturou os **Meias Criativos e Pontas**, definidos por alta agilidade e talento técnico (`skill_moves` e drible elevados).

No setor defensivo e de meio-campo, a análise revelou as nuances mais interessantes. O **Cluster 0** identificou os **Meio-Campistas Box-to-Box**, o "motor" do time, com altas taxas de interceptação e passe. A defesa foi segmentada em três grupos distintos: o **Cluster 2 (Zagueiro "Xerife")**, focado puramente em força física e destruição de jogadas; o **Cluster 6 (Defensor Técnico)**, composto por laterais e zagueiros com boa saída de bola; e, notavelmente, o **Cluster 3**, uma anomalia estatística que agrupou **Defensores Canhotos**. Este último grupo foi isolado não apenas por métricas defensivas, mas pela lateralidade, destacando a importância tática de jogadores canhotos para a geometria da saída de bola.

Conclusão

A aplicação conjunta de **GMM e UMAP** provou-se altamente eficaz para o problema proposto. O modelo não apenas replicou as divisões óbvias do futebol (goleiros vs. linha), mas conseguiu extrair subgrupos táticos sofisticados — como a distinção entre zagueiros técnicos e físicos — que seriam invisíveis em uma análise superficial. Os resultados confirmam que as estatísticas do jogo contêm padrões latentes que correspondem diretamente à intuição tática do futebol moderno.