https://doi.org/10.15202/25254146.2022v7n4p78

ELABORAÇÃO DE MODELO DE CLUSTERIZAÇÃO PARA RECOMENDAÇÃO DA CONTRATAÇÃO DE JOGADORES DE FUTEBOL

Daniel Pessoa Soeiro
Master Bunisses em Engenharia de Dados pelo ITLab/POLI da Universidade Federal do Rio de Janeiro (UFRJ), Rio de Janeiro, RJ, Brasil
dpsoeiro@ufrj.br

Claudio Miceli de Farias
Ph.D. em Ciência da Computação, ITLab/POLI da Universidade Federal do Rio de Janeiro (UFRJ), Rio de Janeiro, RJ, Brasil cmicelifarias@poli.ufrj.br

Ana Gabriella Amorim Abreu Pereira
D. Sc. em Ciência em Engenharia de Produção, ITLab/POLI da Universidade Federal do Rio de Janeiro(UFRJ), Rio de Janeiro, RJ, Brasil
aga.amorim@gmail.com

Manoel Villas Bôas Júnior
Mestre em Computação Aplicada, ITLab/POLI da Universidade Federal do Rio de Janeiro (UFRJ), Rio de Janeiro, RJ, Brasil
mvbjunior@poli.ufrj.br

Manuel Oliveira Lemos Alexandre
Mestre em Engenharia de Transportes, COPPE - Universidade Federal do Rio de Janeiro (UFRJ), Rio de Janeiro, RJ, Brasil
manuel.alexandre@pep.ufrj.br

RESUMO

A cultura data driven, ou seja, se tomar decisões baseadas em dados, têm se tornado cada vez mais forte dentro do mundo corporativo, pois se entende que os dados, transformados em informações, podem ser um balizador para decisões que façam a empresa crescer e economizar custos. Este conceito tem saído do mundo corporativo e sendo inserido no mundo do futebol, onde clubes usam modelos e dashboards para analisar o mercado em busca de jogadores que supram suas carências táticas e melhorar sua performance. Porém, este conceito ainda é embrionário no Brasil, seja pelo fato de que as decisões são tomadas única e exclusivamente pelos chamados “cartolas” (dirigentes de clubes) ou pela falta de recursos para acessar estas tecnologias, dado isto, este trabalho abordará uma nova proposta de mapeamento de mercado de jogadores no futebol brasileiro, usando modelo de clusterização de machine learning, alimentado por dados abertos da internet, para a aproximação de perfis de atletas e trazer a discussão de como pode ser benéfico para aumentar a competitividade de uma equipe.

Palavras-chave: *Scouting*, *Football Analytics*, *Ranking* Elo, redução de dimensionalidade, modelo de clusterização.

DEVELOPMENT OF A CLUSTERING MODEL FOR FOOTBALL PLAYER HIRING

RECOMMENDATION

ABSTRACT

The data-driven culture, which means making decisions based on data, has been becoming increasingly strong within the corporate world, as it is understood that data, when transformed into information, can serve as a guide for decisions that drive company growth and save costs. This concept has been extending beyond the corporate world and being incorporated into the realm of football, where clubs are using models and dashboards to analyze the market in search of players that fulfill their tactical needs and improve their performance. However, this concept is still in its early stages in Brazil, either due to decisions being solely and exclusively made by the so-called "cartolas" (club executives), or due to a lack of resources to access these technologies. Given this, this work will address a new proposal for mapping the player market in Brazilian football, using a machine learning clustering model fueled by open internet data. The aim is to approximate player profiles and open the discussion on how this can be beneficial in enhancing team competitiveness.

Keywords: Scouting, Football Analytics, Ranking Elo, dimensionality reduction, clustering model.

1 INTRODUÇÃO

O futebol nos últimos 20 anos vem sofrendo com uma verdadeira revolução financeira, onde muitos clubes pelo mundo receberam aportes financeiros volumosos e se tornaram times quase imbatíveis. Um exemplo disto é o do Manchester City (Inglaterra) que no ano de 2008 foi comprado pelo grupo árabe *Abu Dhabi United Group Investment and Development* (hoje *City Group*) por £ 210 milhões (TRIVELA, 2008) e que desde então criou uma verdadeira hegemonia em seu país, vencendo 15 títulos em 10 anos e, mais recentemente, sendo campeões da Champions League.

No Brasil, este mesmo movimento vem ocorrendo com certa força nos últimos anos liderado por clubes como Palmeiras e Flamengo que, segundo Zirpoli (2021), concentram quase 40% da arrecadação dos clubes da Série A e que tem se revezado na disputa dos principais titulos nacionais e continentais, ganhando juntos 4 Taças Libertadores, 2 Copas do Brasil, 4 Brasileiros e 5 Estaduais. Tal sucesso faz com que muitos clubes desejassem modernizar suas estruturas e receber mais investimentos para também se tornarem competitivos, e tal desejo foi impulsionado a partir de 2021, quando o conceito de Sociedade Anônima de Futebol (SAF) foi implementado, permitindo que clubes possam migrar de uma associação civil sem fins lucrativos para uma associação empresarial, adotando o formato de “clube-empresa” , vendendo suas ações de forma majoritária ou minoritária para alguma empresa ou acionista (CAPELO, 2022). Tal medida pode fazer com que muitos clubes se organizem financeiramente, modernizem suas gestões e recebam altos investimentos para montar times fortes que disputam títulos.

Porém, enquanto alguns afimam que a lei da SAF pode ser boa para aumentar a competitividade do campeonato nacional (FONTOURA, 2022), isto pode causar o efeito oposto

ao esperado e a diminuir signicativamente. Algo que demonstra este efeito é um estudo feito pelo Stadiumetric (2022), onde se faz uma análise da competitividade das principais ligas do mundo, comparando o período de 1979 à 1998 com o dos anos de 2000 à 2019, onde foi analisado o número de campeões distintos que se teve em cada temporada contida nestas décadas, e também quantos times diferentes conseguiram lutar pelo título em cada uma delas. Foram analisadas as 5 principais ligas do mundo, sendo elas a Bundesliga (Alemanha), Ligue 1 (França) , Premier League (Inglaterra), Calcio (Itália) e La Liga (Espanha).

A conclusão que se chegou foi que, se comparando as décadas, se houve uma grande diminuição da competitividade nestas ligas e um aumento na concentração dos times que disputam os títulos, principalmente na Bundesliga, Calcio e Premier League. O que explica isto é a correlação entre desempenho esportivo e investimento realizado, sendo que há um desequilíbrio na forma como os diversos clubes arrecadam dinheiro (STADIUMETRIC, 2022).

Então, como os clubes com menos recursos podem montar times competitivos e lutar contra os “gigantes”?

O Brentford Football Club (Brentford) é um modesto clube de Londres, que surpreendeu na temporada 2020/2021 da segunda divisão inglesa, apesar de ter o quarto menor orçamento. Isto foi possível graças à metodologia de contratação elaborada pelo proprietário do clube, Matthew Benham, um programador e multimilionário inglês, que emprega um sistema de inteligência de mercado, chamado scouting. Esse sistema utiliza dados e modelos avançados para analisar jogadores de ligas menos conhecidas, com custos baixos e alto potencial de crescimento, resultando em contratações assertivas e lucrativas, como o exemplo do atacante Scott Hogan, onde pagaram £ 750 mil para o modesto Rochdale em 2014 e, depois de marcar 21 gols em 36 partidas pelo Brentford, foi vendido ao Aston Villa Football Club por £ 12 milhões em 2017. (LAW, 2020)

No fim da temporada 2021, o time subiu à primeira divisão pela primeira vez na sua história e desde então faz campanhas consistentes, brigando “de peito aberto” com os gigantes. O ponto máximo mais recente do projeto foi na temporada 2022/2023, onde venceu o poderoso e endinheirado “City” por 2 x 0.

Quando um clube não possuí uma alta arrecadação ou está endividado, o que limita seus investimentos, o processo de montagem de seu elenco necessita ser mais eficiente, pois contratar o jogador “errado” pode prejudicar o clube dentro de campo e financeiramente. Hoje, existe uma larga massa de dados disponíveis que podem ser utilizados pelos clubes na hora de decidir qual jogador contratar, mas, principalmente no Brasil, ainda não se há a cultura data driven

implementada, ou por que os “cartolas” ainda tomam a decisão final ou por que não sabem o que fazer com estes dados. Sendo assim, se faz ainda mais necessário se demonstrar a importância e a aplicabilidade dos dados para se montar um time capaz de brigar com as equipes mais fortes.

O exemplo do Brentford pode e precisa ser copiado para todas as ligas profissionais do mundo, para que a competitividade possa se manter equilibrada e em alto nível e demonstra o fato de que os dados nas mãos certas, nos modelos certos e com as decisões certas podem trazer resultados consistentes, tanto dentro quanto fora de campo.

Este trabalho, inspirado por clubes que aplicam a tecnologia a seu favor para melhorarem seus desempenhos e serem mais assertivos no mercado de transferências, irá propor um modelo de clusterização que, utilizando dados abertos da internet que contém informações de estatísticas de atletas de todo o mundo, irá recomendar ao final os jogadores do Campeonato Brasileiro que mais têm perfil parecido aos melhores atletas do mundo que, em uma situação hipotética, seriam recomendados à clubes para uma possível contratação.

2 DESENVOLVIMENTO

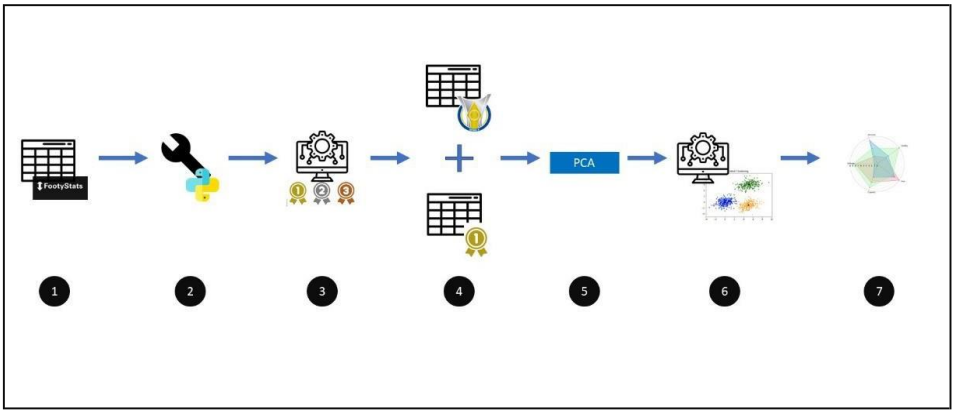
Este trabalho propõe uma solução acessível e de baixo custo para análise de mercado de jogadores, utilizando informações disponíveis na internet, um software open-source e conceitos de modelagem de dados, aprendizado de máquina e visualização de dados. A sistemática envolve a criação de um fluxo de trabalho estruturado, desde a coleta até a recomendação da contratação dos jogadores, com dados contendo estatísticas facilmente acessíveis e confiáveis. O processo é conduzido por meio da linguagem de programação Python, abrangendo limpeza, feature engineering e a recomendação de atletas utilizando o modelo de clusterização K-Means.

2.1 Workflow do Projeto

Um workflow bem estruturado é essencial para criar um modelo confiável de agrupamento de perfis de jogadores através do uso de machine learning. Isso requer habilidades analíticas e experiência para formular as perguntas adequadas, identificar fontes de dados relevantes, explorar e compreender os dados, e aplicar o feature engineering. Além disso, é crucial definir as variáveis finais que serão utilizadas para avaliar o desempenho do modelo. Um fluxo de trabalho organizado proporciona eficiência, qualidade e possibilita replicar o projeto com outros conjuntos de dados. (HOSNI, 2021; SALTZ, 2022).

A implementação deste trabalho segue as etapas (workflow) detalhadas na Figura 1, onde:

Figura 1 - Detalhe do workflow de implementação do modelo



FONTE:(O AUTOR,2023)

1. Captura de dados: Essencial para criar um modelo de classificação, o maior desafio é encontrar fontes relevantes que forneçam estatísticas e informações sobre os jogadores e seu desempenho no campeonato.
2. Tratamento dos dados: Nesta etapa, é necessário lidar com inconsistências, dados nulos, valores outliers e informações irrelevantes, garantindo a integridade do processo de modelagem.
3. Criação de ranking de atletas: No modelo de clusterização, a recomendação precisará ser baseada por meio de uma comparação com perfis dos melhores atletas do mundo em cada posição, onde os perfis mais a estes semelhantes serão recomendados. Os melhores atletas do mundo serão escolhidos usando um método quantitativo, usando o algoritmo Rating Elo
4. Combinação dos dados: Os dados dos top players serão combinados com os dados dos atletas do Campeonato Brasileiro para preparar a análise no modelo de classificação.
5. Aplicação de redução de dimensionalidade: Com o objetivo de evitar interferências de variáveis redundantes ou altamente correlacionadas, será aplicada redução de dimensionalidade.
6. Execução do algoritmo K-Means: Será definido o número de clusters a serem utilizados e o algoritmo K-Means será aplicado para agrupar os jogadores.
7. Avaliação de performance: Serão definidas métricas para avaliar o desempenho do modelo e a similaridade dos clusters criados, assim como para avaliar os jogadores recomendados pelo modelo em relação a seus desempenhos. Os jogadores com a melhor similaridade serão os jogadores recomendados.

Criar um fluxo de trabalho permite entender e planejar cada passo a ser executado,

entender os caminhos pelos quais os dados passarão, agir mais rapidamente caso algum step apresente gargalos e a garantia de que o projeto possa ser replicado no futuro com dados novos.

O detalhamento das propostas de cada uma das etapas e seus respectivos resultados serão explanados nas próximas seções.

2.2 Captura de Dados

A fonte de dados escolhida para o modelo é o site FootyStats (footystats.org), destinado a fornecer informações de jogos para apostadores de casas de betting. A seleção ocorreu devido às suas vantagens: ser gratuita, oferecer estatísticas abrangentes de todos os atletas nas competições, ter confiabilidade e atualização frequente dos dados, além de disponibilizar uma API para linguagem Python. A base, que abrange mais de 150 ligas em todo o mundo, foi extraída em formato CSV e carregada usando a biblioteca Pandas, resultando em um conjunto de dados com 3228 linhas e 277 colunas.

Essa escolha se destaca também pela confiabilidade dos dados, validados por analistas independentes, é que essencial para garantir a precisão do modelo e evitar erros para os apostadores que utilizam essas informações para suas decisões. Com essa fonte de dados, será possível explorar uma ampla gama de estatísticas relevantes para a modelagem do desempenho dos jogadores no Campeonato Brasileiro. Alguns indicadores disponíveis na base são número de gols marcados, número de assistências, número de jogos e número de desarmes.

2.3 Análise Exploratória e Tratamento de Dados

Para a etapa de tratamento de dados, após escolha e coleta da fonte, é importante ressaltar que se tomou a decisão de não realizar a exclusão de dados considerados outliers, pelo fato de que dados muito acima de média podem ajudar a evidenciar jogadores que possuem um desempenho elevado em uma determinada liga. Um exemplo disto é o exibido na Figura 2, que mostra a distribuição dos gols marcados por atacantes nas ligas observadas.

Figura 2 - Boxplot de gols marcados por atacantes

Fonte:(O AUTOR,2023)

Os dados revelam uma distribuição desigual dos gols marcados pelos atacantes. A maioria deles registra um baixo número de gols, como indicado pela mediana próxima ao primeiro quartil. Porém, existe um pequeno grupo que se destaca com números expressivos de gols, conforme refletido pela grande diferença entre o terceiro quartil (representando os atacantes que marcam mais gols) e o número máximo de gols marcados. Isso sugere que alguns poucos atacantes tiveram um desempenho excepcional em termos de gols marcados, enquanto a maioria teve uma performance mais modesta. Para contextualizar os números, apenas 6,64% dos 695 atacantes na base alcançaram 10 ou mais gols na temporada, enquanto 38% marcaram entre 1 e 10 gols, e 41% não marcaram nenhum gol.

A única medida prévia de tratamento adotada foi a remoção de dados com alta correlação, desde que fosse coerente no contexto do problema. A base FootyStats apresenta três quebras de dados com base nos indicadores - Total/Em casa/Fora - uma vez que o desempenho de um atleta pode variar dependendo do local da partida. No entanto, esses dados mostraram-se altamente correlacionados, por essa razão, foram eliminadas informações repetidas ou com alta correlação, reduzindo o número de colunas de 277 para "apenas" 106, uma vez que foram considerados dados irrelevantes para o modelo.

2.3 Aplicação do Algoritmo Rating Elo

O processo inicial para a construção do modelo de recomendação envolve a identificação dos principais jogadores em cada posição nas principais ligas de futebol do mundo, utilizando critérios objetivos baseados em estatísticas de desempenho em campo. Esses jogadores são selecionados para servirem como referência na recomendação, conforme detalhado no Item 2.1, e é fundamental ressaltar que a escolha dos melhores atletas é feita de forma imparcial, evitando

critérios subjetivos, como votações de jornalistas para prêmios como a Bola de Ouro, que podem ser influenciadas por opiniões populares em vez de dados concretos.

Para selecionar os melhores atletas que servirão como modelo para a recomendação, se utiliza o algoritmo Rating Elo, desenvolvido pelo matemático húngaro Arpad Elo na década de 1950, com o objetivo de estabelecer um critério quantitativo para classificar os jogadores de xadrez de acordo com suas habilidades. Esse algoritmo confronta dois atletas individualmente, comparando-os com base em seus atributos individuais, além de considerar o histórico de desempenho ao longo do tempo. Dessa forma, o sistema não se limita apenas a vitórias e derrotas, mas proporciona uma análise detalhada e imparcial das habilidades de cada jogador, resultando em uma classificação mais precisa e significativa para o processo de recomendação. (CHESS.COM, 2023; MITTAL, 2020; VEISDAL, 2019). Ao confrontar os dois atletas, os comparando, o algoritmo retorna à pontuação do ranking atualizada e, consequentemente, suas respectivas classificações.

Da mesma forma que se utiliza o algoritmo para ranquear os melhores jogadores de xadrez, se utilizou também para se encontrar os melhores atletas das 5 principais ligas da Europa em cada uma das 4 posições (atacante, meio campista, defensor, goleiro), totalizando 20 atletas "modelos". A replicação do modelo utilizando em Python seguiu os seguintes passos:

1. Função 1: Como o ranking é novo, é necessário atribuir uma pontuação inicial neutra a todos os jogadores. Pelo método Elo, este valor deve ser de 1300 pontos para jogadores novatos sem cadastro no sistema (VEISDAL, 2019). Se criou uma base contendo esta pontuação inicial e os indicadores de cada atleta a serem usados na construção do ranking (número de gols etc.) conforme a Figura 3 mostra.

Figura 3-Criação de Campo com a Pontuação Inicial

	full_name	position	league	rating_elo	force_accurate_crosses_per_game_overall	force_aerial_duels_won_per_game_overall	force_appearances_overall	force
0	Aaron Ramsey	Midfielder	Ligue 1	1300	0.267867	0.178290	0.824138	
1	Abdallah Dipo Sima	Forward	Ligue 1	1300	0.159067	0.309455	0.882759	
2	Abdou Diallo	Defender	Ligue 1	1300	0.150000	0.150000	0.150000	
3	Abdoul Bamo Meité	Midfielder	Ligue 1	1300	0.150000	0.150000	0.443103	
4	Abdoulaye Bamba	Defender	Ligue 1	1300	0.188533	0.236157	0.501724	

Fonte:(O AUTOR,2023)

2. Função 2: Na abordagem proposta, as estatísticas de jogo selecionadas serão comparadas uma a uma para garantir que a comparação seja feita exclusivamente entre as mesmas propriedades de dois atletas distintos. Em um atributo Y (ex: Número de Gols), se tem um atleta A com um número Y_A e um atleta B com Y_B, esta primeira função retornará um valor

W para cada jogador dependendo do resultado da comparação.

- Se $Y_A > Y_B$, então $W_A = 1$ e $W_B = 0$
- Se$Y_A < Y_B$, então $W_A = 0$e $W_B = 1$
- Se $Y_A = Y_B$, então $W_A = W_B = 0{,}5$

3. Função 3: Após isto, é calculada a probabilidade de vitória esperada no "duelo" entre os dois jogadores que estão sendo comparados, levando em consideração suas pontuações atuais no *ranking Elo*. Essa medida é crucial para a posterior atualização do ranking, pois influencia diretamente os ajustes nas pontuações dos jogadores após a partida. Dado um atleta A com uma pontuação no *ranking* P_A e um atleta B, com P_B, a vitória esperada (EW) é calculada segundo as Equações 1 e 2 (GUERRERO, 2022).

$$EW_A = \frac{1}{\left(10^{\left(|P_A - P_B|/400\right)} + 1\right)} \qquad (1)$$

$$EW_B = 1 - EW_A \qquad (2)$$

4. Função 4: Nesta etapa, as pontuações do *ranking* são atualizadas após a comparação entre os atletas, considerando quem perdeu e quem ganhou a disputa. A variável Ké utilizada para ponderar a vitória ou a derrota na atualização do resultado. Os atletas A e B receberão, cada um, uma nova pontuação P'_A *e* P'_B, calculadas conforme mostram as Equações 3 e 4 (GUERRERO, 2022).

$$P'_B = P_B + K * (W_B - EW_B) \quad (3)$$

$$P'_A = P_A + K * (W_A - EW_A) \quad (4)$$

5. Função 5: Esta é a função principal, que executa um *looping* percorrendo a tabela por campeonato, posição e indicador, selecionando 2 jogadores distintos para fazer o cálculo e atualização da pontuação no *ranking*. O resultado de todo o processo deve ser uma base consolidada atualizada com a pontuação final de cada jogador.
6. Função 6: Com a base gerada na etapa anterior, serão selecionados os atletas com a maior pontuação final por posição por campeonato considerado, totalizando 20 jogadores. Esta base é unida à de dados dos jogadores do Campeonato Brasileiro, com o fim de preparação para a inserção no modelo de clusterização.

Um estudo simples consegue evidenciar o comportamento diferenciado dos jogadores de elite selecionados pelo modelo Elo em relação ao restante dos atletas, pois conforme evidenciado na Tabela 1 seus indicadores são extremamente descolantes. Percebe-se que os jogadores de elite, inclusive, cometem menos faltas por jogo do que os demais atletas, evidenciando seu

comportamento *outlier*.

Tabela 1 – Comparativo *Top Players* x Demais Jogadores

Medida	***Top Players***	**Demais Jogadores**
Número de Jogadores	20	3207
Média de Assistências Por Jogo	0,14	0,03
Média de Faltas Cometidas	0,47	0,55
Média de Dribles Realizados Por Jogo	0,62	0,31
Média de Gols	6,33	1,06
Média de Passes Certos	599,14	290,68

Fonte: (O AUTOR, 2023)

Após esta etapa, a base com as informações dos jogadores top players foi unida a base que continha as informações dos atletas do Campeonato Brasileiro, a fim de preparar as informações para serem inseridas no modelo de classificação.

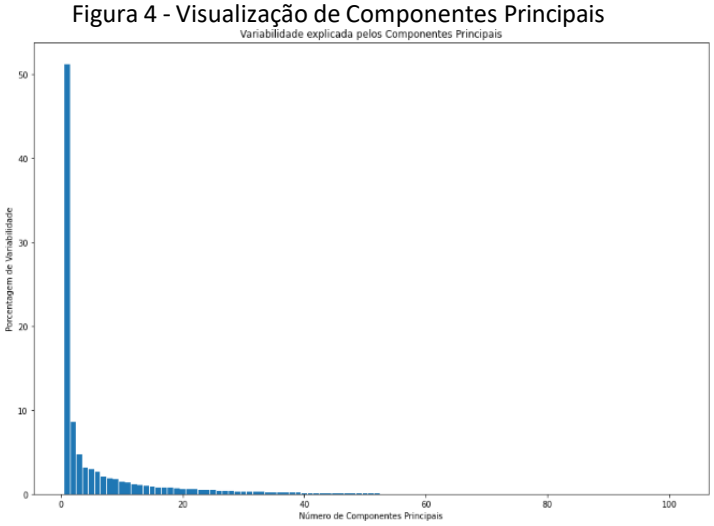
2.4 Redução de Dimensionalidade

Com a base definida, a última etapa antes do modelo consta em aplicar a redução de dimensionalidade, utilizando o método *Principal Component Analysis* (PCA), pois o *dataframe* ainda é muito extenso, mesmo com a remoção feita na etapa de análise exploratória, e aplicar este método garante que se tenha uma base mais condensada sem perder informações relevantes. Aqui se segue o que foi aplicado por Webster (2021), que ao criar um modelo de clusterização para recomendação de jogadores, definiu um número de componentes principais que abrangesse 60% da variância dos dados originais, obtendo resultados satisfatórios.

Como evidenciado pela Figura 4, os dois primeiros componentes principais definidos pelo PCA representam, aproximadamente, 65% da variabilidade dos dados, portanto este é o número escolhido para a aplicação do modelo.

É fundamental considerar que cada posição possui indicadores que podem exercer maior ou menor influência em sua performance global. Um exemplo claro disso é o número de gols marcados, que é um indicador relevante para um atacante, mas dispensável para um goleiro. Levando isso em conta, a base de dados foi subdividida em quatro partes, uma para cada posição, e o PCA foi aplicado a cada uma delas. Consequentemente, devido às variações nos valores de variância entre os indicadores de acordo com a posição, os componentes principais calculados também serão distintos.

Figura 4 - Visualização de Componentes Principais



Fonte: (O AUTOR,2023)

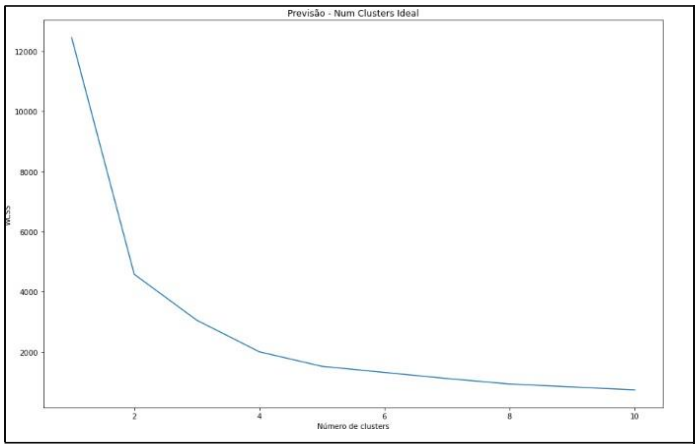
2.5 Aplicação do Modelo de Clusterização Para Recomendação

2.5.1 Definição do Número de Clusters

A definição do número de clusters no modelo será feita utilizando a metodologia da "Curva do Cotovelo". Esse procedimento envolve a execução do algoritmo K-means diversas vezes para diferentes valores de K (número de clusters) em um intervalo específico. Para cada valor de K, é calculada a distância ao centroide do cluster correspondente usando o método euclidiano, e em seguida, o quadrado dessa distância é obtido (TEMPORAL, 2019).

É importante destacar que, após várias execuções, o método atinge um ponto de "efeito platô", onde o algoritmo não consegue mais melhorar o resultado para determinar o número ideal de clusters com a menor distância entre os centroides. Portanto, o número a ser adotado deve ser escolhido antes desse platô ser alcançado. Conforme evidenciado na Figura 5, o ponto mais baixo da curva ocorre quando o número de clusters é 4, tornando esse o valor a ser adotado (TEMPORAL, 2019).

Figura 5 - Curva do Cotovelo



Fonte: (O AUTOR,2023)

2.5.2 Avaliação do Modelo

Com a execução do modelo de clusterização para cada posição realizado, se parte para a análise dos clusters criados e da recomendação dos jogadores que mais se alinham os perfis dos melhores jogadores. O resultado da clusterização, de modo geral, pode ser visto no Quadro 1, com detalhe para a alta similaridade em cada um dos cluster, todas acima de 70%, garantindo uma similaridade percentual média de 76,75%, um resultado que pode ser considerado excepcional.

Quadro 1 – Resultado dos Clusters Obtidos

Posição	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Similaridade (%)
Atacante	105	44	68	16	72%
Defensor	98	117	16	67	75%
Meio Campo	138	199	49	3	86%
Goleiro	49	22	2	2	74%

Fonte: (O AUTOR, 2023)

Outras 3 medidas utilizadas para medir a qualidade dos agrupamentos são o coeficiente de silhueta, o índice Calinski-Harabasz e o índice Davies-Bouldin. O Quadro 2 detalha os resultados para cada um destes indicadores.

Quadro 2 – Análise Quantitativa Dos Agrupamentos Obtidos

Posição	Silhueta	Índice CH	índice Davies-Bouldin
Atacante	0,55	398	0,74
Defensor	0,52	449	0,88
Meio Campo	0,59	563	0,69
Goleiro	0,59	230	0,67

Fonte: (O AUTOR, 2023)

O coeficiente de silhueta é uma métrica que varia de -1 a 1, indicando a qualidade dos agrupamentos em análise de clusterização. Valores próximos de 1 sugerem agrupamentos bem definidos, onde os pontos estão mais próximos dos membros do mesmo cluster em relação aos outros clusters (FELCAM, 2020). No caso apresentado, todos os valores de silhueta acima de 0.5 indicam bons agrupamentos para todas as posições.

O Índice Davies-Bouldin mede a similaridade média entre cada cluster e aquele mais similar a ele, e quanto menor seu valor, melhor é a separação entre os agrupamentos (NOGUEIRA e DOS SANTOS, 2017). Nota-se que todas as posições apresentaram valores de Davies-Bouldin abaixo de 1, indicando que os agrupamentos são relativamente bons e os clusters estão bem separados.

O Índice CH (Calinski-Harabasz) avalia a relação entre a dispersão dentro e entre os clusters, sendo que valores mais altos indicam uma melhor separação entre os clusters (JUNIOR, 2019). Observa-se que a posição de meio campo apresentou o maior valor de Índice CH (563), sugerindo que os agrupamentos nessa posição possuem maior separação entre os grupos e menor dispersão dentro dos clusters em relação às outras posições, porém todos os demais clusters possuem resultados relativamente satisfatórios. O índice CH da posição de goleiro é extremamente baixo em relação aos demais, porém seu coeficiente de silhueta e seu índice Davis-Bouldin corroboram para se chegar à conclusão de que seus agrupamentos estão dentro do esperado.

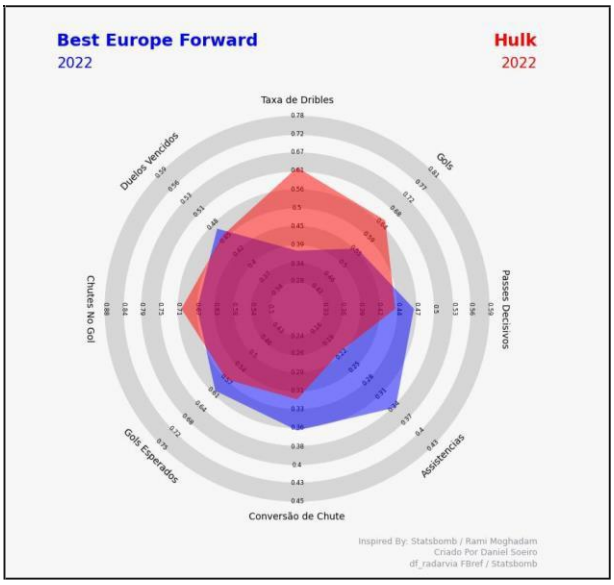
Os valores de silhueta, Índice CH e Índice Davies-Bouldin indicam que os agrupamentos possuem boa qualidade e apresentam separação satisfatória entre os jogadores pertencentes a diferentes posições.

2.5.3 Recomendação dos Jogadores Por Similaridade de Perfil

2.5.3.1 Atacantes

Nos atacantes, o atleta que possuí a melhor similaridade com os melhores atletas da Europa é o Hulk, do Atlético Mineiro. Na Figura 5, comparando o perfil dos *top players* com o de Hulk, é perceptível o desempenho semelhante em indicadores como “chutes no gol”, “duelos vencidos” e “gols esperados”, sendo que o atleta brasileiro chega a ser superior em alguns quesitos, corroborando para a escolha de Hulk nesta recomendação.

Figura 5 - Comparativo – Hulk

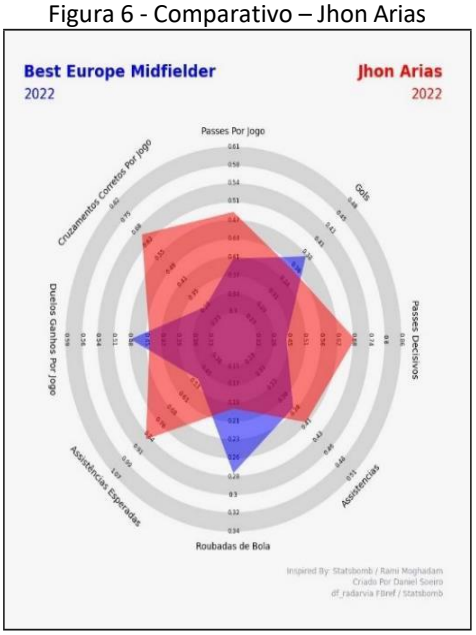


Fonte: (O AUTOR,2023)

2.5.3.2 Meio Campo

O atleta que possuí a melhor similaridade com perfil desejado é Jhon Arias, do Fluminense. Na Figura 6, observando o comparativo dos perfis, se percebe uma total dominância de Arias em todos os atributos definidos, sendo melhor que o perfil top player na maioria deles, com mais passes por jogo, mais cruzamentos corretos e mais assistências. Tendo isto, se recomenda Jhon Arias como meio campista com o perfil desejado.

Figura 6 - Comparativo – Jhon Arias

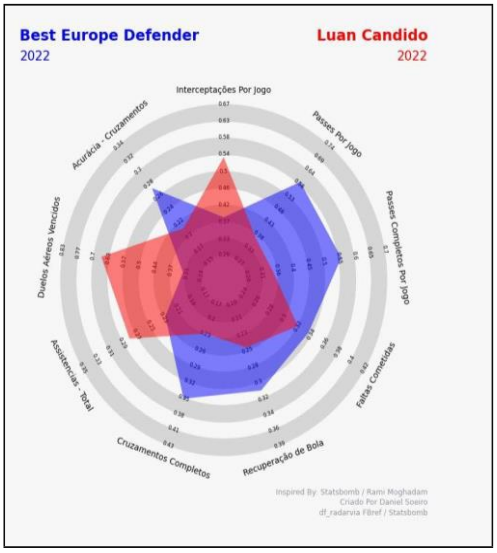


Fonte:(O AUTOR,2023)

2.5.3.3 Defensor

O zagueiro com a melhor similaridade é Lucas Cândido, do Red Bull Bragantino. Com a Figura 7 se tem a comparação dos perfis, onde se percebe que Lucas é melhor nos duelos aéreos, tem mais interceptações por jogo, mais assistências e números semelhantes em acurácia nos cruzamentos, corroborando para a recomendação dele.

Figura 7 - Comparativo – Luan Cândido

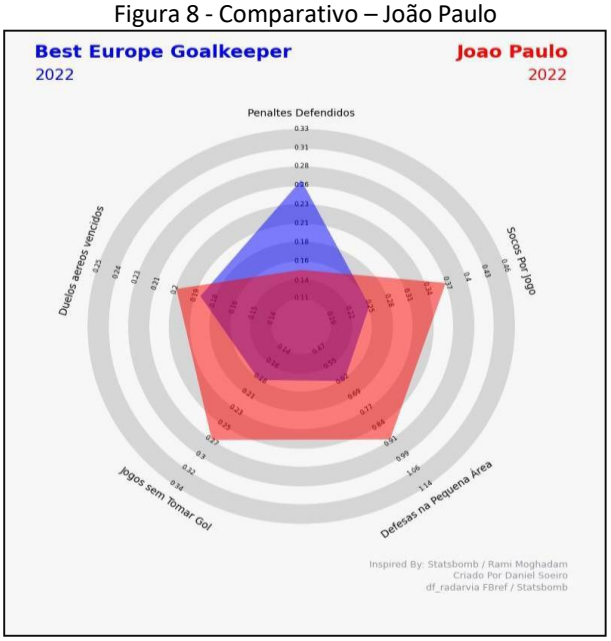


Fonte: (O AUTOR,2023)

2.5.3.4 Goleiro

Aplicando a similaridade nestes jogadores, o perfil mais semelhante é o de João Paulo, goleiro do Santos. No comparativo entre os perfis, na Figura 8, se percebe que João Paulo é dominante em quase todos os requisitos, como mais jogos sem tomar gol, mais defesas dentro da pequena área e mais interceptações. Ele não é um exímio pegador de pênaltis, mas seu perfil para recomendação se mostra acima das expectativas.

Figura 8 - Comparativo – João Paulo



Fonte:(O AUTOR,2023)

3 CONSIDERAÇÕES FINAIS

Ao final deste projeto, se pode concluir que a aplicação de um workflow de machine learning com modelo de clusterização obteve resultados expressivos, com base na comparação de perfis, dos indicadores individuais e análise de similaridade, e desde que se apliquem técnicas corretas e extração, análise e tratamento dos dados, de definição do número de clusters utilizando métodos empíricos e análises visuais corretas, podem beneficiar não só um aumento na competitividade do futebol, mas no esporte com um todo, levando informação que ajude clubes e atletas a tomarem decisões mais sábias.

Aplicar a metodologia foi necessário, pois o grande diferencial deste projeto não foi a utilização do modelo em si, mas a construção de uma sistemética bem estruturada de análise dos resultados, garantindo que se teria a recomendação do melhor perfil possivel. Uma grande prova disto são as comparações individuais dos indicadores de cada atleta recomendado e sua similaridade com o perfil alvo, mostrando quantitativamente o porquê determinados atletas possuem maior destaque.

A tecnologia no futebol ainda tem muito espaço para crescer e se desenvolver, principalmente no futebol brasileiro, porém, para que a competitividade se mantenha, é necessário que todos, de todas as divisões tenham acesso democrático a ela. O legado que fica deste trabalho é a prova de que dados podem ajudar clubes a serem mais eficientes em suas contratações, contratando atletas de melhor performance e podendo aumentar o nivel de força de suas equipes, ajudando o futebol a entrar na era da modernidade e, cada vez mais, abandonando métodos ultrapassados que só prejudicam mais as grandes instituições esportivas do nosso país.

REFERÊNCIAS

CAPELO, Rodrigo. O que é SAF? Entenda formato que mudou o futebol brasileiro. **GLOBOESPORTE.COM**, 2022. Disponível em: https://ge.globo.com/negocios-do-esporte/noticia/2022/09/02/o-que-e-saf-entenda-o-formato-de-clube-empresa-que-mudou-o- futebol-brasileiro.ghtml. Acesso em: 7 junho 2023.

CHESS.COM. Chess Terms. **Chess.com**, 2023. Disponível em: https://www.chess.com/terms/elo- rating-chess. Acesso em: 23 junho 2023.

FELCAM, Igor. Entendendo Clusters e K-Means. **CWI Software**, 2020. Disponível em: https://medium.com/cwi-software/entendendo-clusters-e-k-means-56b79352b452. Acesso em: 14 jun. 2023.

FONTOURA, Luâ. SAF amplia a competitividade de clubes brasileiros. **Medium**, 2022. Disponível em: https://medium.com/betaredacao/saf-pode-ampliar-a-competitividade-dos-clubes- brasileiros-4d71299a5298. Acesso em: 7 junho 2023.

GUERRERO, Daniel. How to calculate ELO score for international teams using python. **MLearning.ai**, 2022. Disponível em: https://medium.com/mlearning-ai/how-to-calculate-elo- score-for-international-teams-using-python-66c136f01048. Acesso em: 26 junho 2023.

HOSNI, Youssef. End-to-End Machine Learning Workflow. **Medium**, 2021. Disponível em: https://medium.com/mlearning-ai/end-to-end-machine-learning-workflow-part-1-b5aa2e3d30e2. Acesso em: 01 julho 2023.

JUNIOR, Gilney. Qualidade de Agrupamentos (Ciência de Dados). **Medium**, 2019. Disponível em: https://medium.com/@gilneyjnr/qualidade-de-agrupamentos-ci%C3%AAncia-de-dados-4b1176bef5e5. Acesso em: 14 jun. 2023.

LAW, Joshua. Cultura das estatísticas aproxima time pequeno de Londres da Premier League. **UOL**, 2020. Disponível em: https://www.uol.com.br/esporte/futebol/ultimas-noticias/2020/06/11/estatistica-empurra-time-modesto-de-londres-em-sonho-de-ir-a-premier-league.htm. Acesso em: 9 junho 2023.

MITTAL, Raghav. What is an ELO Rating? **Medium**, 2020. Disponível em:

https://medium.com/purple-theory/what-is-elo-rating-c4eb7a9061e0. Acesso em: 23 junho 2023.

NASCIMENTO, Caio. Sobre Abelhas e Lobos - A curiosa união entre Brentford e Midtjylland. **Footure**, 2020. Disponível em: https://footure.com.br/brentford-midtjylland-benham-ankersen/. Acesso em: 9 junho 2023.

NOGUEIRA, Murilo C.; DOS SANTOS, Raul T. **Estudo comparativo de algoritmos de agrupamento para a definição de Zonas de Manejo**. Universidade Federal de Mato Grosso - Instituto de Computação. Cuiabá, p. 6. 2017.

SALTZ, Jeff. What is a Data Science Workflow? **Data Science Process Alliance**, 2022. Disponível em: https://www.datascience-pm.com/data-science-workflow/. Acesso em: 4 julho 2023.

STADIUMETRIC. Índice de Competitividade – Brasileirão vs. Ligas Europeias. **SportsInsider**, 2022. Disponível em: https://sportinsider.com.br/brasileirao-ligas-europeias/. Acesso em: 8 junho 2023.

TEMPORAL, Jessica. Como definir o número de clusters para o seu KMeans. **Pizza de Dados**, 2019. Disponível em: https://medium.com/pizzadedados/kmeans-e-metodo-do-cotovelo-94ded9fdf3a9. Acesso em: 12 jul. 2023.

TRIVELA. Empresa árabe afirma ter comprado Manchester City. **Trivela**, 2008. Disponível em: https://trivela.com.br/inglaterra/empresa-arabe-afirma-ter-comprado-manchester-city/. Acesso em: 20 jun. 2023.

VEISDAL, Jørgen. The Mathematics of Elo Ratings. **Cantor's Paradise**, 2019. Disponível em: https://www.cantorsparadise.com/the-mathematics-of-elo-ratings-b6bfc9ca1dba. Acesso em: 23 junho 2023.

WEBSTER, Edd. **Finding The Next Gerard Piqué**. Parma Calcio 1913. Parma, p. 30. 2021.

WYSCOUT. Pricing. **Wyscout**, c2023. Disponível em: https://wyscout.com/pt-pt/pricing/. Acesso em: 7 jul. 2023.

ZIRPOLI, Cassio. O ranking de receitas dos clubes do Brasil em 2021; 20 maiores geraram R$ 6,9 bilhões. **Cassio Zirpoli**, 2022. Disponível em: https://cassiozirpoli.com.br/o-ranking-de-receitas- dos-clubes-do-brasil-em-2021-20-maiores-geraram-r-69-bilhoes/. Acesso em: 7 junho 2023.