UNIVERSIDADE FEDERAL DE CAMPINA GRANDE
CENTRO DE ENGENHARIA ELÉTRICA E INFORMÁTICA
CURSO DE BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

DANILO DE MENEZES FREITAS

A MODERNIZAÇÃO DO JOGO: ANÁLISE E CARACTERIZAÇÃO DE JOGADORES DAS 5 PRINCIPAIS LIGAS DE FUTEBOL EUROPEIAS

CAMPINA GRANDE - PB

2022

DANILO DE MENEZES FREITAS

A MODERNIZAÇÃO DO JOGO: ANÁLISE E CARACTERIZAÇÃO DE JOGADORES DAS 5 PRINCIPAIS LIGAS DE FUTEBOL EUROPEIAS

Trabalho de Conclusão Curso apresentado ao Curso Bacharelado em Ciência da Computação do Centro de Engenharia Elétrica e Informática da Universidade Federal de Campina Grande, como requisito parcial para obtenção do título de Bacharel em Ciência da Computação.

Orientador : Fábio Jorge Almeida Morais

CAMPINA GRANDE - PB

2022

DANILO DE MENEZES FREITAS

A MODERNIZAÇÃO DO JOGO: ANÁLISE E CARACTERIZAÇÃO DE JOGADORES DAS 5 PRINCIPAIS LIGAS DE FUTEBOL EUROPEIAS

Trabalho de Conclusão Curso apresentado ao Curso Bacharelado em Ciência da Computação do Centro de Engenharia Elétrica e Informática da Universidade Federal de Campina Grande, como requisito parcial para obtenção do título de Bacharel em Ciência da Computação.

BANCA EXAMINADORA:

Fábio Jorge Almeida Morais
Orientador – UASC/CEEI/UFCG

Everton Leandro Galdino Alves
Examinador – UASC/CEEI/UFCG

Francisco Vilar Brasileiro
Professor da Disciplina TCC – UASC/CEEI/UFCG

Trabalho aprovado em: 02 de Setembro de 2022.

CAMPINA GRANDE - PB

RESUMO

O futebol, assim como diversas outras atividades do nosso cotidiano, recebeu e vem recebendo grandes mudanças com o advento da tecnologia. A grande evolução desse esporte no século XXI foi guiada, em certa parte, pelo avanço tecnológico implementado. A análise estatística é um grande expoente desses avanços, a maioria dos grandes clubes de futebol de hoje possuem um setor que alia essa análise a parte tática do jogo, essas atividades norteiam dirigentes e treinadores em como fazer as contratações corretas para o time e quais jogadores estão com bom desempenho em campo. Diante do exposto, utilizando dados extraídos de uma base chamada *StatsBomb*, disponível no site fbref.com, este trabalho busca fazer um estudo do impacto dos jogadores a partir da análise e visualização de dados. A partir da extração desses dados, o objetivo é observar, a partir do cálculo de uma métrica, os fundamentos e estilos de jogo presentes e observados hoje no esporte moderno, como construção de jogo, criação ofensiva, finalização, capacidade defensiva, etc. A partir dessa métrica, é possível ainda fazer uma análise, levando em consideração a posição dos jogadores, o que mostra onde cada uma das características está mais concentrada em campo. A partir disso, é possível observar, por exemplo, quantos defensores possuem tal característica ofensiva ou atacantes que realizam ações defensivas. Essa análise pode ser feita para várias posições e funções do futebol e ajuda a entender e demonstrar o impacto que cada um dos jogadores tem dentro de campo.

A modernização do jogo: Análise e caracterização de jogadores das 5 principais ligas de futebol europeias

Danilo de Menezes Freitas
Universidade Federal de Campina Grande
Campina Grande, Paraíba, Brasil
danilo.menezes.freitas@ccc.ufcg.edu.br

Fábio Jorge Almeida Morais
Universidade Federal de Campina Grande
Campina Grande, Paraíba, Brasil
fabio@computação.ufcg.edu.br

ABSTRACT

Football, like many other activities in our daily lives, has received and is receiving major changes with the advent of technology. The significant evolution of this sport in the 21st century has been driven, to some extent, by the technological advances implemented. Statistical analysis is a key exponent of these changes, most of the big football clubs today have a staff unit that relates this analysis to the tactical part of the game, these activities guide managers and coaches on how to make the right signings for the club and which players are performing well on the pitch. Given the above, using data extracted from a database called StatsBomb, available on the fbref.com website, this work aims to study the impact of players based on the analysis and visualization of data. By extracting this data, the aim is to observe, by calculating a metric, the fundamentals and styles of play present and observed today in modern sport, such as build-up play, offensive creation, finishing, defensive ability, etc. From this metric, it is also possible to make an analysis, taking into account the players' position, which shows where each of the characteristics is more focused on the pitch. From that, it is possible to observe, for example, how many defenders have offensive characteristics or attackers who perform defensive actions. This analysis can be done for many football positions and roles and helps to understand and demonstrate the impact each player has on the pitch.

RESUMO

O futebol, assim como diversas outras atividades do nosso cotidiano, recebeu e vem recebendo grandes mudanças com o advento da tecnologia. A grande evolução desse esporte no século XXI foi guiada, em certa parte, pelo avanço tecnológico implementado. A análise estatística é um grande expoente desses avanços, a maioria dos grandes clubes de futebol de hoje possuem um setor que alia essa análise a parte tática do jogo, essas atividades norteiam dirigentes e treinadores em como fazer as contratações corretas para o time e quais jogadores estão com bom desempenho em campo. Diante do exposto, utilizando dados extraídos de uma base chamada *StatsBomb*, disponível no site *fbref.com*, este trabalho busca fazer um estudo do impacto dos jogadores a partir da análise e visualização de dados. A partir da extração desses dados, o objetivo é observar, a partir do cálculo de uma métrica, os fundamentos e estilos de jogo presentes e observados hoje no esporte moderno, como construção de jogo, criação ofensiva, finalização, capacidade defensiva, etc. A partir dessa métrica, é possível ainda fazer uma análise, levando em consideração a posição dos jogadores, o que mostra onde cada uma das características está mais concentrada em campo. A partir disso, é possível observar, por exemplo, quantos defensores possuem tal característica ofensiva ou atacantes que realizam ações defensivas. Essa análise pode ser feita para várias posições e funções do futebol e ajuda a entender e demonstrar o impacto que cada um dos jogadores tem dentro de campo.

Palavras chave

Ciência de Dados; Análise Descritiva; Clustering; Dados de Futebol

1. INTRODUÇÃO

Os esportes de alto rendimento demandam de seus atletas sempre que eles estejam no seu limite; uma pequena mudança no corpo ou técnica, já causa um grande impacto no resultado. Com essa necessidade de sempre melhorar e sempre evoluir, a estatística e a análise de dados trazem avanços positivos para o esporte. O futebol, assim como outros esportes, é uma atividade que gera muitos dados, e cada vez mais é possível observar e analisar esses dados seja para o planejamento técnico ou para o decisões do mercado de jogadores.

Existem ferramentas que monitoram cada passo de cada jogador em uma partida de futebol, assim como cada chute, a posição de cada chute, a perna com que a bola foi chutada, a velocidade, etc. Esse volume de dados, que vem em uma grande crescente, impacta cada vez mais no futebol que está tão presente na vida das pessoas; impacto esse se deve em grande parte ao uso de ferramentas de análise de dados, que usam cada uma das ações dos jogadores em campo para traduzir a maneira em que foi jogado o jogo e de que forma os jogadores se comportaram, trazendo informações sobre um atleta que não são diretamente conhecidas da comissão técnica do clube.

No entanto, a análise no futebol não é uma ideia tão nova quanto se pensa. Logo depois da segunda guerra mundial, um contador inglês chamado Charles Reep começou a observar o futebol de outra forma e começou a coletar dados, os anotando com caneta e papel. Depois de observar centenas de partidas, ele começou a observar, nos dados que coletou, padrões de certa estabilidade, e começou a desenvolver uma teoria [1]. O que ele percebeu foi que a cada passe concluído, a probabilidade de acertar o próximo diminuía, e concluiu que apenas 8,5% das ações do jogo continham mais de 3 passes e que apenas 2 de cada 9 gols

observados foram feitos com jogadas com mais de 3 passes. Essa teoria ficou conhecida como *long ball* na Inglaterra, e influenciou consideravelmente a forma como o futebol era jogado no país.

O objetivo de Reep era achar a fórmula secreta para vencer uma partida de futebol, utilizando os dados que ele observava. Essa fórmula consistia em um estilo de jogo direto, com passes mais objetivos, e esteve muito presente nos campos ingleses por muitos anos, especialmente nos anos 80. No entanto, depois de certo tempo, a teoria dele começou a ser criticada por ter sido uma análise rasa e começaram a surgir novas ideias e novos estilos de enxergar o futebol, o que tornou o esporte no que é hoje, uma atividade em constante evolução.

Todavia, tendo em vista essa diversidade de pensamento dos dias atuais, nota-se que a análise dos dados não deve ter como seu principal objetivo desvendar uma fórmula secreta, mas sim, potencializar os atletas e comissão técnica. Ademais, os dados gerados em grande volume trazem uma complicação para que sejam de fato utilizados, considerando isso, faz se necessário a aplicação de técnicas de análise para que se transformem em informações úteis. Levando isso em consideração, a motivação deste trabalho é, principalmente, fazer um estudo sobre o jogo de futebol e a partir disso realizar uma análise da forma como cada jogador impacta o jogo, buscando entender as multifacetas desse esporte, que Charles Reep e tantos outros analistas ajudaram a influenciar.

Desta forma, considerando que a quantidade de dados gerados por cada jogador em uma partida de futebol vem crescendo cada vez mais, o objetivo desse trabalho é gerar análises sobre o desempenho dos atletas a fim de entender as principais características que definem a qualidade de um jogador.

2. TRABALHOS RELACIONADOS

Atualmente uma enorme diversidade e volume de dados é gerada no esporte e servem de base para diferentes análises que envolvem desempenho, características da modalidade esportiva, dentre outras. Nesse sentido, BIALKOWSKI, Alina *et al*. [2] desenvolveram um trabalho que apresenta uma maneira diferente de enxergar o jogo de futebol. O trabalho analisou cerca de 400 milhões de *data points* baseados numa temporada de uma certa liga de futebol a partir do uso imprescindível de análise de dados para gerar informações úteis. Levando isso em consideração, o trabalho gera uma discussão sobre o posicionamento dos jogadores em campo. Por ser um esporte muito dinâmico e contínuo, as posições fixadas em campo podem ser um problema. Para isso, o trabalho introduz uma representação baseada em funções, que podem ser dinamicamente atualizadas a cada momento, trazendo uma facilidade de ajuste ao dinamismo e velocidade do esporte.

O trabalho de Memmert, D., Lemmink, K.A.P.M. & Sampaio, J. [3] atesta o avanço ocasionado no futebol gerado a partir das novas tecnologias. Essas inovações trouxeram novas possibilidades na captura de informações espaço-temporais, fazendo com que a análise de posicionamento dos jogadores seja mais eficiente, trazendo um melhor entendimento sobre o dinamismo e a complexidade de uma partida de futebol. Baseado em dados capturados em uma partida de alto nível entre Barcelona e Bayern de Munique, o trabalho foca, através de sistemas dinâmicos e redes neurais, em 3 diferentes abordagens: Análise da coordenação inter-jogadores, coordenação inter-time e coordenação inter-linhas, assim como interações entre os dois times e um coeficiente de compactação dos jogadores dentro de campo.

Por outro lado, as análises desenvolvidas neste trabalho de conclusão de curso buscam gerar informações sobre as características de jogo que diferenciam os jogadores a partir de métricas de suas atuações (ex: Gols, Assistências, Divididas). Essas análises podem ser úteis para identificar jogadores com diferentes características de jogo e habilidades diferenciadas.

3. METODOLOGIA

Esse trabalho fez uso de análises exploratórias e descritivas a fim de observar as diferentes características que os jogadores desempenham em uma partida de futebol, e de que forma eles causam impacto dentro de campo.

3.1 Coleta dos Dados

Os dados gerados pelos jogadores são capturados pela *StatsBomb*, e estão disponíveis no site fbref.com. A coleta dos dados foi feita com Python [4] através da ajuda de bibliotecas como *pandas* [5], *BeautifulSoup* [6], *Selenium* [7].

A amostra escolhida foi composta por jogadores concentrados nas 5 principais ligas europeias, que são:

- *Premier League* (Liga Inglesa)
- *La Liga* (Liga Espanhola)
- *Bundesliga* (Liga Alemã)
- *Serie A* (Liga Italiana)
- *Ligue 1* (Liga Francesa)

As competições citadas são as 5 melhores classificadas na UEFA, órgão que rege as competições europeias, e são comumente apelidadas de *Big 5*.

Primeiramente foi capturado para cada uma das ligas citadas uma tabela que contém dados de todos os jogadores que estão inseridos nela, com informações como nome do jogador, referência para a página com o perfil do mesmo, nacionalidade, etc. Essas tabelas foram extraídas utilizando uma ferramenta própria do site que faz com que as tabelas sejam convertidas para *csv.* Como eram apenas 5 tabelas, uma para cada liga, elas foram extraídas manualmente. A partir disso, esses dados foram armazenados com ajuda da biblioteca *pandas* de Python, a fim de serem utilizados na próxima etapa de coleta.

Com o *link* de referência de cada um dos atletas disponíveis na tabela, foi possível iniciar uma nova etapa de coleta. Nesta etapa, o objetivo era capturar dados estatísticos de cada um dos jogadores, esses dados estão expostos no site na forma de relatório, onde cada atleta tem o seu. Estes relatórios são compostos de várias estatísticas e são feitos comparando jogadores de mesma posição dentro das ligas do já citado *Big 5*, utilizando os últimos 365 dias jogados.

A extração dos dados desse relatório foi feita com o auxílio da mesma ferramenta do site que transforma as informações da tabela em *csv.* No entanto, para esta captura foi utilizada a biblioteca *Selenium,* que simula a interação humana com o *browser* de forma a coletar automaticamente os dados em formato

csv. Os dados coletados foram armazenados em arquivos para análises posteriores.

No total, considerando os jogadores que jogaram pelo menos 450 minutos nos últimos 365 dias, foram capturados relatórios de 2083 atletas espalhados por essas 5 grandes ligas. Da *Premier League*, foram capturados 422 relatórios, da *La Liga* foram capturados 431 relatórios, da *Bundesliga* foram 363 relatórios, da *Serie A* foram 444 relatórios e da *Ligue 1* foram recuperados 423 relatórios. No total, foi possível coletar informações detalhadas sobre 2921 jogadores do *Big 5* considerando a temporada 2021-2022. Todos os relatórios extraídos dos jogadores e o código-fonte utilizado para a coleta dos mesmos encontram-se publicamente disponíveis em https://github.com/danilomfreitas/football-impact.

3.2 Estatísticas Extraídas

A partir da coleta e estudo desses relatórios foi possível observar a presença de dezenas de estatísticas para mais de 2 mil jogadores. Os *reports* são formatados como uma tabela com três colunas, em uma delas nota-se o nome da estatística observada, em outra observa-se o valor bruto daquela estatística calculada para cada 90 minutos daquele jogador em campo, e por último uma coluna com o percentil, que demonstra para uma determinada estatística como o número bruto (por 90) se compara a uma amostra com todos os outros jogadores de mesma posição. Um exemplo que pode ser citado é do jogador Lionel Messi, que é um jogador que participa bastante dos gols. Na estatística *Goals* o Messi tem 0,36 a cada 90 minutos jogados, número que o coloca no percentil 75 entre os jogadores de sua posição.

As estatísticas estão agrupadas dentro da tabela por categoria. As categorias presentes nos dados são:

- *Standard Stats*: estatísticas gerais como *Goals* e *Assists*, que são as mais utilizadas pelos interessados no esporte. Essas aparecem novamente nas categorias específicas mencionadas abaixo;
- *Shooting*: estatísticas relacionadas a finalização como *Goals;*
- *Passing*: estatísticas de passe como a *Passes Completed*;
- *Pass Types*: estatísticas ligadas a tendências de passes dos jogadores como a *Crosses*;
- *Goal and Shot Creation*: estatísticas sobre criações de jogadas ofensivas;
- *Defense*: estatísticas de defesa como *Tackles*;
- *Possession*: estatísticas relacionadas a posse da bola, como *Touches*; e
- *Miscellaneous Stats*: outras estatísticas como *Yellow Cards*.

Ao término da coleta de dados foram observadas no total 134 diferentes métricas utilizadas para fazer a análise dos atletas. É interessante ressaltar que todos os jogadores de linha, todos excluindo os goleiros, possuem em seus relatórios as mesmas métricas, mesmo que joguem em posições diferentes. Isso ajuda a perceber o jogo de forma geral, sem o viés posicional, mesmo que de certa forma jogadores de defesa tenham números maiores para métricas defensivas e que jogadores de ataque tenham números maiores para métricas ofensivas.

Além dos dados coletados e do código-fonte, no repositório citado na seção anterior, há também uma descrição[1] completa de todas as estatísticas utilizadas nos relatórios e neste trabalho, com o objetivo de facilitar o entendimento das análises realizadas, uma vez que existem métricas consideradas não triviais.

3.3 Cálculo dos *Clusters*

Feita a coleta, cada jogador teria seu arquivo, com todas as métricas e os valores associadas a ela, bruto e percentil. Após isso, o próximo passo seria agrupar todos esses jogadores juntos.

Diante disso, para obter todos os 2083 arquivos juntos foi utilizado a biblioteca *glob* [8], que agrupou todos os arquivos que possuíam a mesma extensão, nesse caso *.csv*, a partir do nome dos mesmos, formando uma grande lista de nomes dos jogadores. Esses nomes foram utilizados como identificador de cada um dos atletas durante todo o processo, e é basicamente composto pelo par "primeiro nome-último nome" (por exemplo: lionel-messi).

Considerando a grande diferenciação tanto do estilo de jogo quanto das estatísticas coletadas para os goleiros. Foi decidido não considerar dados desse tipo de jogador para as análises realizadas. Desta forma, os dados agrupados foram filtrados para remover dados de goleiros. Essa filtragem não tem impacto nos demais jogadores, uma vez que os jogadores das outras posições possuem as mesmas métricas em seus *reports*. Como resultado, o dataset usado nas análises posteriores possui um registro para cada jogador, com valores das métricas disponíveis no site e computadas para cada 90 minutos jogados, por exemplo um dado jogador jogou um total de 1354 minutos e fez 11 gols no período do relatório, então a cada 90 minutos em campo o jogador faz 0,73 gols. Essa medição é feita para que todos os jogadores sejam comparados a partir da mesma minutagem.

Com base no dataset gerado e filtrado foi realizada uma análise de agrupamento com base nas 134 métricas resultantes para cada jogador. Para isso, foi utilizado, primeiramente, uma análise visual a partir de gráficos de dendrogramas, que é basicamente uma representação hierárquica dos agrupamentos de dados, se assemelhando a uma árvore. Com o uso deste diagrama, a visualização de quantos grupos seriam necessários para observar as diferentes características dos jogadores se torna mais eficiente. A Figura 1 representa o gráfico citado.

[1] No site há algumas métricas descritas em inglês devido ao lugar de origem do site, para melhor compreensão foi escrita uma descrição para cada uma delas em português.

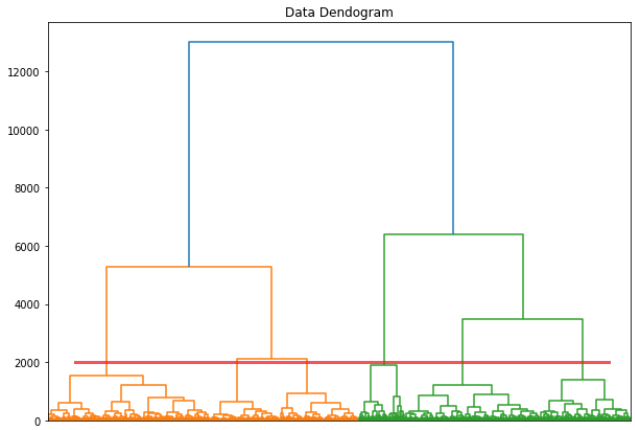

Figura 1: Dendograma da análise de agrupamento hierárquico para os dados analisados

O dendograma é uma representação visual de como os grupos estão separados. Esse método é usado para observar o número ideal de *clusters* que serão utilizados. Para que isso seja feito, basta observar as linhas verticais e horizontais traçadas no gráfico, principalmente a linha horizontal que corta as duas maiores linhas verticais, essa linha delimita a maior distância possível entre *clusters*. Então para definir a quantidade de grupos basta traçar uma linha horizontal em algum ponto abaixo dessa linha principal. Ao traçar no ponto demonstrado no gráfico, pode-se notar que há 6 traços verticais cruzando esta linha horizontal. Esse valor foi considerado como o número de *clusters* esperado na análise de agrupamento realizada.

O agrupamento foi realizado a partir da biblioteca *scikit-learn* presente no Python, a partir do cálculo dos *clusters* e análise dos centróides. Nessa biblioteca estão presentes módulos de *clustering*, incluindo o algoritmo *k-means*, modelo de agrupamento amplamente utilizado baseado no aprendizado não supervisionado. Com base no resultado da execução do *k-means* foram coletados os centróides para cada uma das 134 estatísticas em cada um dos 6 *clusters*. Os centróides foram utilizados para a construção de uma métrica chamada FT (*Football-Traits*).

3.4 Definição da métrica de seleção de características

Os dados dos centróides foram adicionados ao *dataset* base para identificar o grupo calculado para cada jogador analisado. Essa informação é necessária para identificar as principais métricas que diferenciam cada grupo. Essa seleção foi realizada a partir da proximidade da média com os valores dos 6 agrupamentos.

Para auxiliar com a análise das características, foi utilizado um cálculo de variação, que observa o quanto o valor de uma estatística difere da média global considerando todos os jogadores analisados. Para esse cálculo, a forma representada na imagem abaixo foi utilizada.

$$V\,\% = \frac{a - \overline{X}}{\overline{X}} \times 100$$

Onde o V% seria justamente essa variação, sendo o a o valor de determinada estatística para o centróide de um *cluster* e o $\bar{X}$ sendo o valor médio dessa mesma estatística no contexto geral.

Tendo isso em vista, foi feita uma análise para definir quais estatísticas seriam diferenciais de um *cluster* qualquer. Observadas as variações, ficou definido que tal estatística precisaria variar em pelo menos 50% da média aritmética calculada entre todos os *clusters*, seja para mais ou para menos, para que seja considerada um diferencial. Com o limite definido, foi possível observar após análise quais eram as diferentes métricas que caracterizavam cada um dos grupos, auxiliando o entendimento de como isso reflete no posicionamento dos jogadores e no jogo em si.

4. RESULTADOS

Essa seção demonstra os resultados obtidos a partir das análises realizadas a partir da metodologia anteriormente explicada e dos 1870 *data points* utilizados para este trabalho.

4.1 Interpretação dos agrupamentos

Analisando a métrica FT construída e os resultados encontrados através dela é possível notar padrões que mostram as características incluídas em cada um dos agrupamentos encontrados. As imagens a seguir demonstram, a partir de gráficos, estatísticas que ajudam a diferenciar cada grupo.

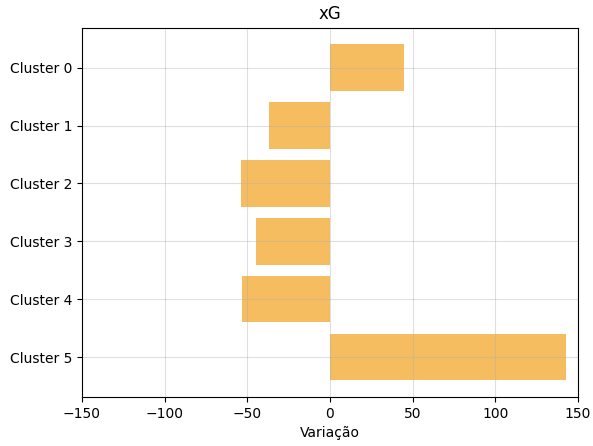

Figura 2: Gráfico da análise da métrica *xG* por *cluster* analisado

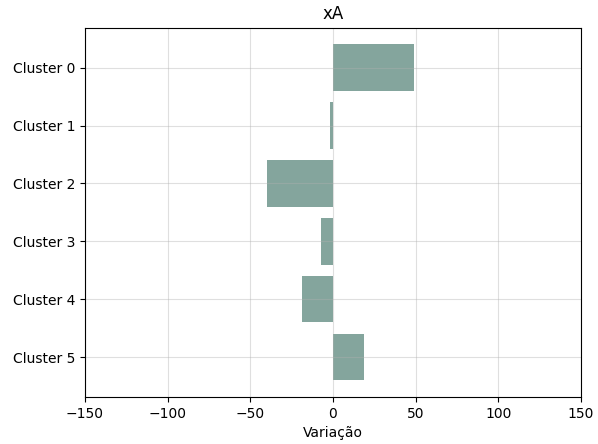

Figura 3: Gráfico da análise da métrica *xA* por *cluster* analisado

O fator a ser observado por esses dois gráficos é justamente a criação ofensiva. As métricas *xG* e *xA* são utilizadas para demonstrar de que forma as chances foram criadas em uma partida. O *xG*, explicado detalhadamente na descrição no repositório anteriormente citado, é uma métrica que calcula a qualidade das chances de um jogador ou time, utilizando a posição do chute, posição dos adversários, etc. Já o *xA* é um cálculo feito baseado nos passes, utiliza de dados referentes a posicionamento para calcular a chance de um passe se tornar uma assistência.

A partir dos gráficos é claramente perceptível que os *clusters* 0 e 5 concentram os jogadores com maior potencial de criação ofensiva, com o *cluster* 0 mais focado na criação de jogadas e o 5 mais focado nas finalizações. Os demais *clusters* possuem influência muito menor nas ações ofensivas, levando a crer que são compostos por jogadores mais recuados em campo.

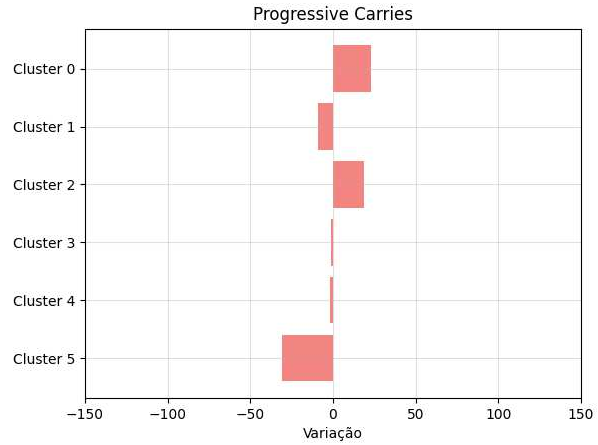

Figura 4: Gráfico da análise da métrica *Progressive Carries* por *cluster* analisado

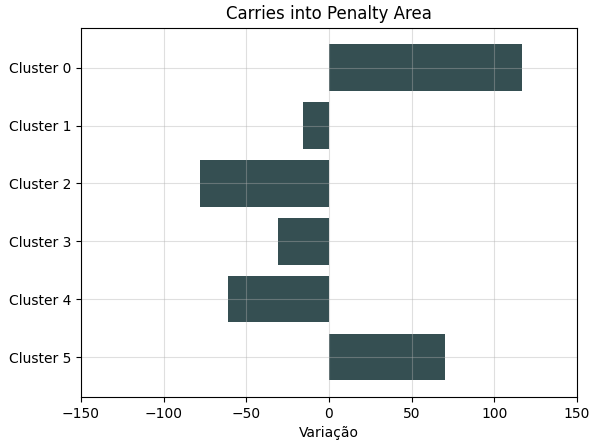

Figura 5: Gráfico da análise da métrica *Carries into Penalty Area* por *cluster* analisado

As Figuras 4 e 5 mostram a capacidade de construção de jogo através da condução de bola por parte dos jogadores. *Progressive Carries* é basicamente quantas vezes um jogador conduz a bola em direção ao gol adversário por pelo menos 5 jardas (ou 4,6 metros). Já *Carries into Penalty Area* é quando um jogador conduz a bola até a pequena área do adversário.

A partir disso, é possível observar que tanto o *cluster* 0 quanto o *cluster* 2 participam mais ativamente da construção de jogo, enquanto os atletas presentes no agrupamento 5 participam menos. Porém, quando se observa o segundo gráfico há uma diferença, essa métrica é mais associada a uma ação ofensiva do que de construção de jogo, por isso os *clusters* 0 e 5 tiveram altos números, e o *cluster* 2, mesmo com muitas conduções progressivas não chegou muito à área, levando a crer novamente que sejam jogadores com características mais recuadas, mesmo que participem muito da construção.

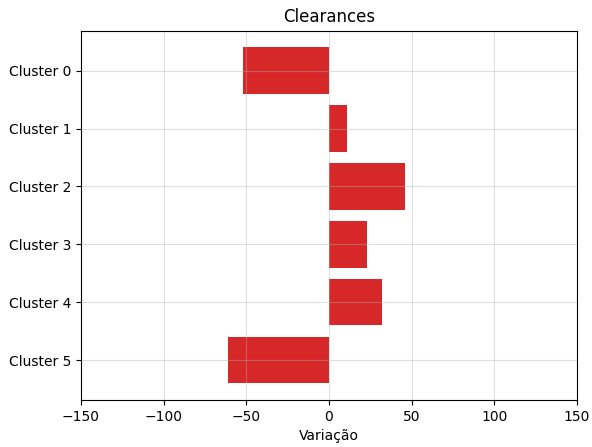

Figura 6: Gráfico da análise da métrica *Clearances* por *cluster* analisado

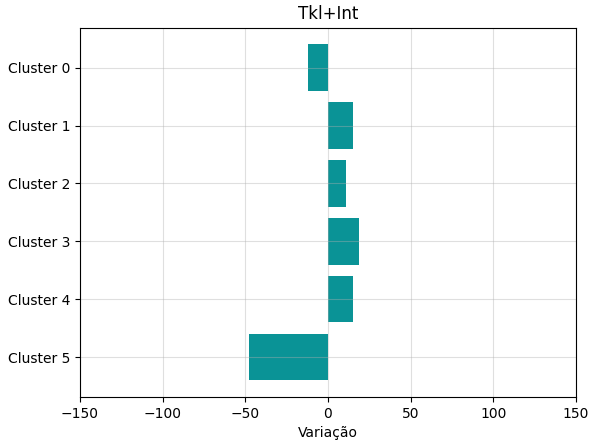

Figura 7: Gráfico da análise da métrica *Tkl+Int* por *cluster* analisado

Os gráficos demonstrados nas Figuras 6 e 7 mostram métricas relacionadas à ações defensivas. *Clearences* é uma métrica que observa as ações de corte dos jogadores, que é quando a bola está em jogo em situação de perigo e é chutada para longe tirando a chance de gol do adversário. Jogadores que se encontram nos *clusters* 1, 2, 3 e 4 foram os mais ativos nesse aspecto, o que reforça a discussão que neles se encontravam jogadores mais recuados, que são os que participam mais dessas ações, já os atletas nos *clusters* 0 e 5, participaram bem menos, reforçando também a discussão de que são compostos por jogadores mais avançados no campo.

Em relação a métrica *Tkl + Int,* que é composta por valores de divididas e interceptações, descreve ações defensivas que também são mais presentes em jogadores mais recuados em campo. Isto pode ser observado na Figura 7, visto que os agrupamentos 1, 2, 3 e 4 são os mais ativos nesse quesito, mostrando também, novamente, que os agrupamentos 0 e 5 possuem jogadores menos participativos na defesa.

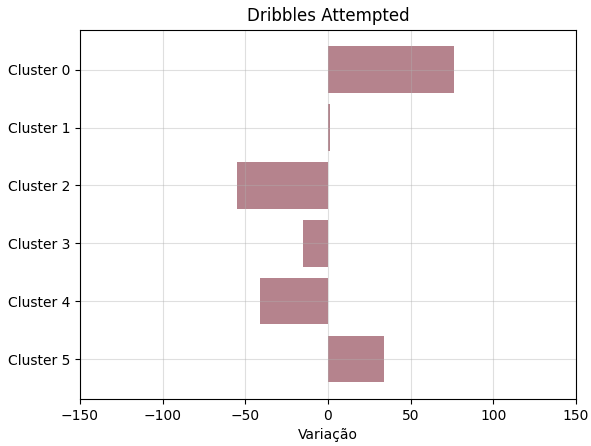

Figura 8: Gráfico da análise da métrica *Dribbles Attempted* por *cluster* analisado

Como já observado, acredita-se que os jogadores pertencentes aos grupos 0 e 5 possuem mais características ofensivas. No entanto, um aspecto que diferencia os grupos é a métrica *Dribbles Attempted* demonstrada na Figura 8. Esta métrica consiste *n*a quantidade de dribles tentados por um jogador, e isso fala muito sobre as características dos jogadores do grupo 0. São eles que driblam mais, que conduzem mais a bola até a área adversária e que criam mais chances de gol. São a mente criativa do time, que oferece as oportunidades para os finalizadores.

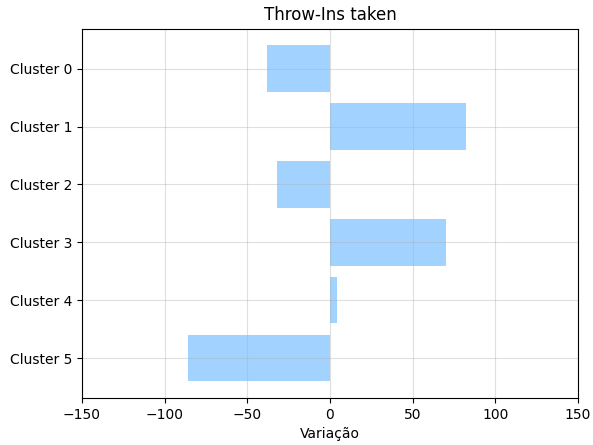

Figura 9: Gráfico da análise da métrica *Throw-Ins taken* por *cluster* analisado

A Figura 9 mostra um gráfico com a métrica *Throw-Ins taken*. Essa estatística não é uma característica tão importante quanto as outras, porém mostra uma tendência relevante para a análise. Ela fala sobre a quantidade de arremessos laterais cobrados por um jogador. Geralmente, esses arremessos são cobrados por jogadores mais próximos as alas ou jogadores mais recuados, pois os jogadores mais ofensivos estão mais próximos da área, buscando o gol. O que leva a crer que atletas presentes nos *clusters* 1 e 3 possuem mais tendência a estarem nas alas, e os dos *clusters* 0, 2 e 5 estarem mais presentes no meio.

4.2 Distribuição dos dados na *FT*

A partir do estudo feito em relação a cada um dos *clusters*, é importante observar de que forma eles estão distribuídos em relação aos atletas analisados. A figura abaixo apresenta a quantidade de jogadores incluídos em cada um dos agrupamentos.

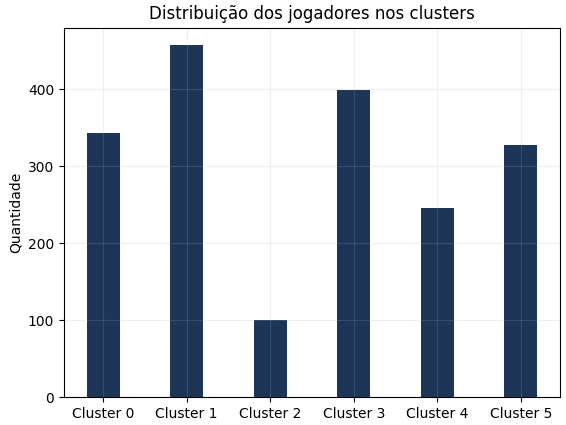

Figura 10: Gráfico da distribuição dos jogadores em cada um dos *clusters*

Observa-se na Figura 10 que a distribuição é, de certa forma, equilibrada, mesmo que seja perceptível que no *cluster* 2 há bem menos jogadores que os outros e que no *cluster* 1 há mais atletas que os outros.

O *cluster* 0 discutido anteriormente concentra 343 jogadores, sendo portanto 18,3% de todos os atletas analisados. O *cluster* 1, o maior observado no gráfico, concentra 457 jogadores, que é em torno de 24,4% dos atletas analisados. O *cluster* 2, claramente o menor observado no gráfico, concentra apenas 100 jogadores, que equivale apenas a 5,3% dos atletas analisados. O *cluster* 3 concentra 399 jogadores, número também expressivo que equivale a 21,3% de todos os atletas. O *cluster* 4 é o segundo menor, concentrando 246 jogadores, que equivale a 13,1% dos atletas analisados. O *cluster* 5 concentra 328 jogadores, o que equivale a 17,5% de todos os atletas analisados.

4.3 Relação Jogadores-Posição

Antes de analisar de que forma os jogadores se enquadram dentro da métrica FT, é necessário observar de que forma os jogadores com dados extraídos se comportam dentro de campo. O gráfico representado na Figura 11 mostra a quantidade de jogadores em cada um dos terços de campo, onde no primeiro terço se encontram os jogadores de defesa, no segundo terço se encontram os jogadores de meio-campo e no terceiro terço se encontram os jogadores de ataque.

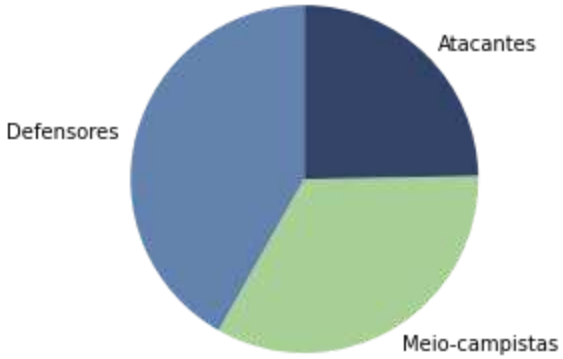

Figura 11: Gráfico da distribuição dos jogadores em relação às posições em campo

Segundo os dados encontrados no site fbref.com, os defensores estão em maior número, 782 dos jogadores são considerados defensores, que é em torno de 41,8% dos atletas analisados. Esse número alto de defensores pode ser explicado por alguns padrões táticos do futebol, geralmente as equipes jogam com esquemas táticos de 4 ou 5 jogadores na defesa, o que faz com que a demanda de jogadores de defesa seja maior.

O segundo maior grupo de jogadores é o de meio-campistas, de acordo com o site esse grupo concentra 626 jogadores, que é em torno de 33,5% dos atletas analisados. O que faz dos meio-campistas terem números expressivos também é relacionado ao padrão tático do futebol. Geralmente nos esquemas táticos empregados atualmente são utilizadas linhas de 3 ou 4 meias, que é significativo, porém menos que os defensores, geralmente utilizados com linhas de 4 ou 5. Sendo assim, a demanda para jogadores meio-campistas é geralmente menor que a de defensores.

O grupo com menos jogadores é o de atacantes. Segundo o site, 462 jogadores são considerados atacantes, que é em torno de 24,7% dos atletas analisados. Esse número menor também pode ser explicado por padrões táticos observados no futebol. Geralmente atacantes são empregados em linhas de 2 ou 3 atacantes, podendo às vezes ter apenas 1. Como são utilizados em menor quantidade, faz com que a demanda deles seja menor que a dos defensores e meio-campistas.

4.4 Rotulação de agrupamentos

Levando em consideração todas as discussões relacionadas aos resultados encontrados, é possível reconhecer padrões e características específicas dos grupos e o que diferenciam os jogadores incluídos em cada um deles. Tendo isso em vista, é possível definir rótulos que ajudam a compreender melhor os agrupamentos. Estes rótulos estão descritos abaixo:

0. Playmaker Ofensivo: os atletas incluídos nesse grupo são especialistas na criação de jogadas ofensivas, utilizando do drible e de passes agudos para criar oportunidades para os jogadores finalizadores.

1. Equilibrado médio: são os jogadores com características mais equilibradas, não são os que mais finalizam, nem os que mais desarmam, são atletas que trazem equilíbrio e fazem de tudo um pouco. Concentra a maior parte dos jogadores, sendo eles de grande maioria defensores e atacantes.

2. Construtor recuado: os atletas presentes nesse grupo participam muito da criação inicial de jogo, com passes de média e longa distância e conduções de bola, utilizadas com o objetivo de quebrar as primeiras linhas de marcação do adversário, além disso possuem grande participação defensiva. São o grupo com a menor quantidade de jogadores.

3. Equilibrado defensivo: composto por jogadores que possuem características semelhantes aos do grupo 1, com bastante equilíbrio, porém com uma composição maior de defensores.

4. Âncora defensiva: os atletas incluídos nesse grupo são jogadores com baixa participação ofensiva, e em todo o resto são equilibrados, possuem baixos números de criação e finalização.

5. Finalizador de jogadas: os jogadores presentes neste agrupamento são os com maior quantidade de chutes e consequentemente de gols, são os artilheiros de seus times, sua principal função é marcar gols. Esse grupo é constituído majoritariamente de jogadores atacantes.

4.5 Relação FT-Posição

Entendido a forma em que os jogadores estão distribuídos dentro de campo e observando também a métrica FT formulada, é possível relacionar os jogadores e seu posicionamento com os agrupamentos observados pela métrica.

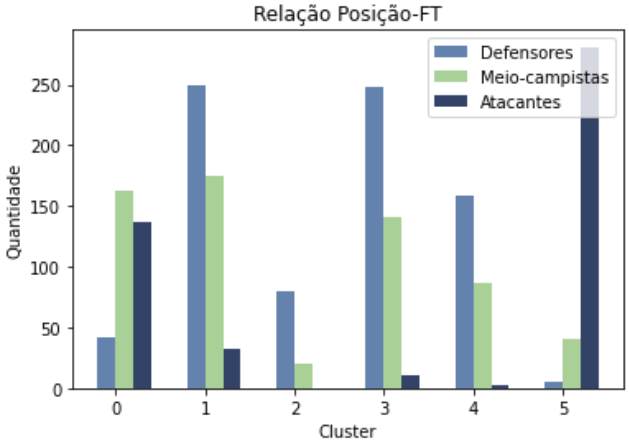

Figura 12: Gráfico relacionando o posicionamento dos jogadores e os *clusters* analisados

Considerando a Figura 12, é possível observar que os jogadores se enquadram de forma adequada em relação aos rótulos dados para cada um dos grupos. O *playmaker* ofensivo é composto majoritariamente por meias e atacantes e o grupo finalizador de jogadas concentra a maioria dos atacantes analisados, que é o esperado. O esperado para os outros grupos eram de jogadores mais recuados, tendo em vista as características demonstradas por esses grupos. Isso se concretiza ao analisar os rótulos definidos para cada um deles, os equilibrado médio e defensivos possuem números similares de jogadores recuados, porém o médio possui número maior de meio-campistas. O grupo de construtores recuados não possui atacantes, reforçando a característica defensiva do grupo. E por fim, o grupo dos âncoras defensivos é reforçado também pela falta de atacantes, visto que na análise são tidos como jogadores pouco participativos nas ações ofensivas.

5. CONCLUSÕES

O futebol é um esporte multifacetado, popular no mundo inteiro, mesmo que as regras sejam as mesmas, cada país tem seu jeito de jogar e cada jogador tem suas características. No passado, para observar o jeito e as distinções de cada um dos atletas só assistindo ao jogo em um estádio ou pela televisão. Hoje isso mudou, a tecnologia trouxe mudanças importantes para um esporte centenário como o futebol, tanto dentro como fora de campo.

Na atualidade, cada ação executada dentro de um jogo é capturada, criando uma imensa quantidade de dados. E é nisso que a ciência de dados trabalha, em como transformar grandes quantidades de dados em informações que tragam valor para atletas, comissão técnica e aficionados no futebol. Tendo isso em vista, a partir da análise desses tipos de dados, esse trabalho conseguiu capturar informações valiosas para calcular de que forma os jogadores impactam dentro do jogo e como ele utiliza de suas características e tendências para ajudar seu time a sair com a vitória em uma partida ou campeonato.

De acordo com a métrica construída, a FT, 6 grupos ficaram evidentes na análise, e todos eles possuem características que os distinguem. Esses grupos podem ser facilmente identificados ao assistir uma partida na televisão, tem o craque do time camisa 10 que cria as jogadas para o goleador camisa 9 e tem também o volante forte que marca o time adversário inteiro.

Considerando isso, tem-se que o trabalho cumpre os objetivos anteriormente almejados. Ele traz uma proximidade maior dos números crus com o espectador do esporte. Fazendo com que ele não apenas assista, mas entenda o que está se passando por dentro das 4 linhas do campo. Como trabalhos futuros é possível evoluir esse trabalho para analisar a relação dos grupos definidos e o valor de mercado dos jogadores, ou até mesmo a relação entre o perfil e posicionamento do jogador com o grupo ao qual ele foi enquadrado.

6. REFERÊNCIAS

[1] SALLY, David; ANDERSON, Chris. Os números do jogo: Por que tudo o que você sabe sobre futebol está errado?. [S. l.]: Companhia das Letras, 2013. 360 p. ISBN 9788565530392.

[2] A. Bialkowski, P. Lucey, P. Carr, Y. Yue, S. Sridharan and I. Matthews, "Large-Scale Analysis of Soccer Matches Using Spatiotemporal Tracking Data," 2014 IEEE International Conference on Data Mining, 2014, pp. 725-730, doi: 10.1109/ICDM.2014.133.

[3] Memmert, D., Lemmink, K.A.P.M. & Sampaio, J. Current Approaches to Tactical Performance Analyses in Soccer Using Position Data. Sports Med 47, 1–10 (2017).

[4] Python. https://www.python.org.

[5] Pandas. https://pandas.pydata.org

[6] BeautifulSoup. https://www.crummy.com/software/BeautifulSoup/

[7] Selenium. https://www.selenium.dev

[8] Glob. https://docs.python.org/3/library/glob.html#module-glob