Instituto de MATEMÁTICA E ESTATÍSTICA UFRGS

DEPARTAMENTO DE ESTATÍSTICA
E-mail: dest@mat.ufrgs.br

Trabalho de Conclusão de Curso

Abordagem Multivariada para Comparação de Atletas de Futebol

Guilherme Teixeira Sartori

3 de junho de 2021

Guilherme Teixeira Sartori

Abordagem Multivariada para Comparação de Atletas de Futebol

Trabalho de Conclusão apresentado à comissão de Graduação do Departamento de Estatística da Universidade Federal do Rio Grande do Sul, como parte dos requisitos para obtenção do título de Bacharel em Estatística.

Orientadora: Profa. Dra. Márcia Helena Barbian

Porto Alegre
Maio de 2021

Guilherme Teixeira Sartori

Abordagem Multivariada para Comparação de Atletas de Futebol

Este Trabalho foi julgado adequado para obtenção dos créditos da disciplina Trabalho de Conclusão de Curso em Estatística e aprovado em sua forma final pela Orientadora e pela Banca Examinadora.

Orientadora:_____
Profa. Dra. Márcia Helena Barbian, UFMG
Doutora pela Universidade Federal de Minas Gerais, Belo Horizonte, MG

Banca Examinadora:

Prof. Dr. Rodrigo Citton Padilha dos Reis, UFMG
Doutor pela Universidade Federal de Minas Gerais, Belo Horizonte, MG

Porto Alegre
Maio de 2021

"É um Deus entre os homens aquele que vive o presente sem arrependimentos do passado e sem expectativas sobre o futuro." (Séneca)

Dedico este trabalho a minha família e meus amigos.

Agradecimentos

Primeiramente quero agradecer a minha orientadora Profa. Dra. Márcia Helena Barbian, por toda orientação e ajuda nesse projeto, ao Prof Dr. Rodrigo Citton Padilha dos Reis por aceitar fazer parte da minha banca. Ao meus pais, Daniel e Cinthia, o dia que decidimos que eu iria sair de casa, foi a minha primeira grande decisão da vida, senti muitas saudades, senti falta das nossas jantas, do fogo na lareira e do enorme amor que sempre recebi dentro de casa. Mas sem dúvidas sou muito grato por ter me possibilitado isso, aprendi muito. Vocês foram e são pais tão presentes que quase tudo que aconteceu durante a minha vida tenha relação com esse amor recebido por vocês. Mãe, tu é a maior fonte de amor que existe, nunca esqueça disso. Pai, tu é meu herói, esse posto é inegociável. Amo muito vocês. Quero agradecer muitos aos meus Avôs. Vó Lenira e Vó Sonia agradeço muito todo suporte que me proporcionaram durante esse tempo e principalmente no meu ano em Santa Maria, saibam que esse neto aqui só guarda coisas boas de vocês, vocês tem uma participação muito importante nessa minha etapa. Vô Moizes, és uma das pessoas com maior bondade que já conheci, agradeço por toda atenção e carinho recebi e toda vez que o Grêmio jogar, saibas que estão juntos na torcida. Vô Sartori, onde for que tu estejas, pode ter certeza que o teu neto, o teu amigão, o Gui. Só guarda lembranças boas de ti, queria muito ter convivido mais contigo, mas eu tenho na memória todas cartas que tu me escrevestes e sei que torce muito por mim. Aos meus Dindos, Di, tenho uma carinho muito grande por ti e pela família que formastes, sempre fostes uma das pessoas que mais me incentivou a estudar, com a tua criativa, seja nos desenhos ou nos versos, eu realmente desejo com toda minha sinceridade que vocês sejam muito felizes e tenho muito o que te agradecer. Digo, além de meu dindo, és um grande amigo, em muitos momentos me enxerguei em ti e te agradeço muito pelos anos em Porto Alegre, tu é meu parceiro. Muito mais que um familiar, sabes que tu pode contar sempre comigo, que nossa relação só cresce. Lolô, meu afilhado e meu amigão. Que tu seja sempre essa criança muito carinhosa e educada. O dindo vai ta sempre aqui disponível pra te ajudar no quer preciso, tenho certeza que com todo esse amor que tu tem dentro de ti, tu vai ser uma pessoa muito feliz. O dindo te ama. Dudinha, minha maninha. Vai demorar pra tu entender o que é um "TCC", segue assistindo "Masha e Urso" que é mais divertido, quando escrevo isso estamos em meio a uma pandemia, há pouquíssimas coisas positivas em uma pandemia, mas se fosse elencar uma, seria a possibilidade de ter passado muito tempo junto contigo e ter participado bastante da tua infância até o momento. Quando tiver idade pra entender tudo isso, só saiba que o mano te ama muito e tu é nossa princesinha. Ao Fábio e a Dani, eu e o Rafa já conversamos muito sobre isso, e sempre a conclusão é a mesma, tivemos muita sorte em vocês acompanharem nossa

mãe e nosso pai, vocês fazem eles mais felizes e nos fazem felizes, somos muito gratos pelo que vocês fizeram/fazem por nos. A isa, por ter feito parte de grande parte da minha graduação e ter me incentivado muito na minha primeira experiência como empreendedor, um sincero muito obrigado. Ao meu amigo e sócio Pedro, Junto a ti encarrei até hj uma das maiores "loucuras" e maiores realizações da minha vida, fizemos mais de 300 churrascos, já atendemos mais de 10 mil pessoas, valeu muito a pena. Brigamos, se abraçamos, fizemos um projeto muito legal o quão me orgulho muito e vou carregar sempre no meu coração. Aos meus amigos de Santa Maria, tenho um carinho absurdo por vocês e por todo acolhimento que tiveram comigo nos anos que moramos na mesma cidade e cada vez que a nos encontramos só confirma meu sentimento, carrego vocês no peito. Aos meus amigos da Vida, aos "Los Pollos Hermanos", espero que vocês saibam o quão esse amigo aqui torce pelo sucesso e felicidades de vocês. E quando falo em sucesso é muito mais que financeiro, que a gente possa manter essa amizade por muito tempo e espero que no futuro a única coisa que mude é que vamos combinar nossos churrascos com os filhos juntos, por que tenho certeza que é uma amizade pro resto da vida, o "sarta veio" está sempre disponível pra ajudar no que for preciso e disponível pra qualquer função junto dos amigos. Muito churrasco, viagens e amizades pra todos nós. Amo vocês. E por último, e mais importante, meu irmão, meu primeiro grande amigo. Rafa, o sentimento de irmão é uma das coisas que mais valorizo nessa vida, tu é a eterna ligação que tenho da nossa família, das lembranças dos nossos momentos juntos, te agradeço por tudo e que tu seja muito feliz, da maneira que tu quiser ser, eu vou estar contigo.

Resumo

Este trabalho tem como objetivo identificar quais jogadores dos campeonatos de futebol se destacam tecnicamente, além de distinguir grupos de atletas com características semelhantes, o que poderia auxiliar em possíveis contratações de clubes que buscam jogadores com habilidades complementares em uma equipe. Para isso, foram utilizadas técnicas de estatística multivariada em duas das cinco principais ligas de futebol do mundo, a francesa e a espanhola, os dados foram obtidos através da tecnica de *web scapping*. A primeira etapa da abordagem proposta consiste na separação de atletas conforme suas posições e trasnformação de variáveis do banco de dados, a etapa seguinte é a redução de dimensionalidade das variáveis transformadas, por meio da técnica de componentes principais. Na terceira etapa foram utilizadas técnicas de agrupamento não-hierárquicas, por fim, na etapa quatro, procurou-se jogadores semelhantes adotando duas métricas: distância euclidiana e similaridade por cosseno. Os jogadores de maior *performance* dentro dos grupos foram considerados como referência e comparados com jogadores semelhantes através da análise do gráfico de radar, que indica visualmente os pontos fortes de cada atleta.

Palavras-Chave: Estatística Esportiva, Futebol, Análise Multivariada, Componentes Principais, *Cluster*, Medidas de Similaridade, Gráfico de Radar.

Abstract

This work aims to identify which players of the championships of technically stand out, in addition to distinguishing groups of athletes with similar characteristics, which could assist in possible club signings who look for players with complementary skills in a team. For this, multivariate statistical techniques were used in two of the five main football leagues in the world, the French and the Spanish, the data were obtained through of the web scrapping technique. The first stage of the proposed approach consists of the separation of athletes according to their positions and the transformation of variables from the data, the next step is to reduce the dimensionality of the transformed variables, through the principal component technique. In the third stage, non-hierarchical clustering techniques, finally, in step four, we sought to similar players adopting two metrics: Euclidean distance and similarity by cosine. The highest performing players within the groups were considered as a reference and compared to similar players through analysis of the radar graph, which visually indicates the strengths of each athlete.

Keywords: Sports Statistics, Football, Multivariate Analysis, Core Components, Cluster, Similarity Measures, Radar Chart.

Sumário

Lista de Figuras

Lista de Tabelas

1 Introdução

O futebol é uma paixão mundial, sendo um dos esportes mais conhecidos e com maior poder financeiro do mundo (Brooks et al., 2016). Uma das características que faz o esporte ser mais emocionante e com maior quantidade de adeptos é sua competitividade e sua imprevisibilidade, como uma expulsão, um gol nos últimos segundos, um chute de fora da área, são todas ações que podem ocorrer a qualquer momento e mudar totalmente o resultado de uma partida. Um exemplo clássico é a final da copa do mundo de 2006, em que a discussão entre o jogador francês Zinedine Zidane e o zagueiro italiano Marco Materazzi acarretou em uma expulsão e mudou de forma decisiva o resultado do jogo e o campeão da copa do mundo. Essa imprevisibilidade representa a complexidade de uma partida, da relação entre jogadas, da combinação dos jogadores, de interação entre os adversários. Isto possibilita análises estatísticas mais complexas, que levem em considerção uma estrutura de dependência entre os jogadores e as variáveis que mensuram características associadas a esses atletas, como a quantidade de chutes a gol ou números de faltas. Nesse trabalho, serão consideradas análises multivariadas, que podem abordar diferentes medidas de desempenho envolvendo o esporte.

Existem fatores não relacionados a habilidade, mas que influenciam o resultado de uma partida. Um exemplo é a vantagem da equipe mandante, que apesar de possuir causas conhecidas, são questões de difícil mensuração. Questões como, por exemplo, efeito da torcida, efeitos de viagem, familiaridade, viés do árbitro, territoriedade e esquema tático (Lago-Peñas et al., 2016). Estatísticas brutas de resumo como os gols, chutes e assistências ainda são a maneira mais comum para comparar o desempenho do jogador analiticamente (Brooks et al., 2016). Com isso, este trabalha visa contribuir de maneira prática, por meio da aplicação de técnicas des estatística multivariada, visando os seguintes objetivos:

i) identificar jogadores com características semelhantes;

ii) quando houver perda de um atleta, propor uma possível substituição por um atleta do elenco que exerça uma função similar dentro da equipe;

iii) auxiliar em possíveis contratações para equipe na busca de habilidades complementares.

De maneira teórica, com a utilização de técnicas de redução de dimensionalidade e de agrupamento, as quais consideram diferentes medidas de distâncias de similaridade entre os jogadores, tais técnicas estatísticas que conjuntamente, podem vir a ser aplicadas na área do esporte.

Nesse trabalho serão analisadas 380 partidas do campeonato francês e 380 partidas do campeonato espanhol, essas competiçoes foram escolhidas, pois estão entre as 5 maiores ligas do cenário do futebol (Transfermarkt, 2021). Devido a grande quantidade de jogos, os dados foram coletados atráves de *web scrapping*, tal técnica utiliza "robos"para obter informações de sites na web. Com base nestes dados, uma análise quantitativa está prevista. Primeiro é realizada uma transformação nas variáveis, também são utilizada o método de redução de dimensionalidade, os jogadores são agrupados e comparados, através das distâncias euclidianas e cosseno, a fim de identificar as características individuais predomintantes dos atletas.

2 O Futebol

O futebol é a modalidade esportiva mais popular no Brasil e tem sua origem na Inglaterra, assim como outros esportes como rugby e cricket. O futebol surgiu nos colégios ingleses a partir de adaptações do jogo com bola. O que começou como um jogo informal com finalidade de diversão passou a ter regras oficiais a partir de 1863, pela criação da Foot-ball Association, dando um caráter mais voltado à "seriedade do esporte" do que a "ludicidade do jogo" (Guterman, 2013).

Assim, o futebol era praticado dentro desses centros educacionais europeus por jovens com a finalidade de passatempo e manutenção do corpo saudável. Após ser bastante difundido dentre jovens de classe alta inglesa, chega à capital paulista com Charles Muller, trazendo consigo duas bolas de couro e um manual de regras e, sem imaginar de sua ação, surge um fenômeno sociocultural que abrangeria todas as camadas sociais no Brasil (Guterman, 2013).

Segundo (Helal et al., 2001), inicialmente eletista, a segregação com negros e pobres era comum, porém depois de lutas e resistências houve uma democratização, ascensão e afirmação desses grupos marginalizados. A globalização trouxe a descaracterização do elemento nacional, com os principais jogadores brasileiros indo atuar nos campeonatos europeus (Guterman, 2013).

Junto com os Jogos Olímpicos, o esporte é líder mundial em termos de cobertura de mídia e quantidade de jogadores profissionais. Segundo o autor (Dietschy, 2013) vários fatores explicam o domínio do futebol como o hábil equilíbrio entre técnica e a força do jogo e a simplicidade de suas regras.

O desempenho do futebol depende de um conjunto de fatores, como áreas técnicas/biomecânicas, táticas, mentais e fisiológicas. Um dos motivos da popularidade do futebol mundialmente é que os jogadores podem não precisar ter uma capacidade extraordinária em nenhuma dessas áreas de atuação, mas possuir um nível razoável em todas elas. Existem tendências para um treinamento e seleção mais sistemáticos, influenciando os perfis antropométricos dos jogadores que competem no mais alto nível. Como acontece com outras atividades, o futebol não é uma ciência, mas a ciência pode ajudar a melhorar o desempenho. Os esforços para melhorar o desempenho do futebol geralmente se concentram na técnica e na tática em detrimento da aptidão física (Bangsbo, 1994).

2.1 Análise de Desempenho no futebol

Se antigamente o que contava era o instinto para a as tomadas de decisões no futebol, atualmente se tem recorrido aos analistas futebolísticos, esses profissionais

utilizam de dados para auxiliar na tomada de decisões, organizando os dados e tirando deles algum aprendizado (Sally e Anderson, 2013).

O autor (Freitas, L. F., 2017) cita que há relatos que Charles Reep foi um dos primeiros analistas futebolisticos, em 1950 elaborava relatórios por meio de um sistema de anotações de distância, direção, altura e resultados dos passes e finalizações. Com o passar dos anos e aumento da capacidade financeira dos clubes e novas técnicas de análises, aumentaram os investimentos em profissionais especializados, a fim de tornar as decisões menos subjetivas.

Dado o grande campo, os numerosos jogadores, a rotatividade limitada de jogadores e a pontuação escassa, o futebol é, sem dúvida, o mais desafiador de analisar de todos os principais esportes coletivos (Liu et al., 2020).

A habilidade de executar padrões de movimento com eficiência e eficácia é o aspecto mais importante do desempenho no futebol, nesse sentido os jogadores devem aplicar habilidades cognitivas, perceptivas e motoras a situações que mudam rapidamente. Tem havido tentativas de medir esses parâmetros para fins de identificação (ou desenvolvimento) de talentos e aquisição de habilidades e pesquisa de intervenção (Ali, 2011).

Unindo a forte questão cultural com o tamanho populacional, possibilita ao Brasil ser um dos maiores exportadores de talentos do esporte no mundo. Como é citado na matéria (Perez, Rafael and Alves, Carlos Eduardo, 2020) o Santos futebol Clube, na última década, faturou mais de 1 bilhão de reais na venda de jogadores oriundos das suas categorias de base.

Com o avanço da tecnologia e métodos computacionais, as análises no futebol tiveram um significativo aumento, diversas empresas investem em análises cada vez mais avançadas, com intuito de fazer previsões sobre partidas, campeonatos e jogadores. Uma das empresas referências no assunto é a Opta Sports, a qual aborda medidas avançadas de desempenho, como a métrica *Expected goals*. O objetivo dessa varíavel é mostrar a probabilidade de um jogador marcar o gol, dado a localização, dificuldade, entre outras variáveis da finalização do atleta. Existem também análises mais avançadas e em constante desenvolvimento, como rastreamento de jogadores baseado em câmeras SportVU, essa tecnologia já é utilizada com frequência no basquete. Os dados gerados são fornecidos às equipes, mas não são informações disponíveis ao público. Outra grande empresa do setor, a (**?**) possui várias ferramentas baseadas em algoritmos estatísticos e inteligência artificial, tais como:

- o sistema de rastreamento que fornece dados por meio da extração e processamento de coordenadas dos jogadores e da bola;
- a tecnologia GPS em tempo real que permite a tomada de decisão ao vivo;
- uma auto codificação e edição de vídeo que permite aos usuários acesso à análises personalizadas e *feedback* de desempenho de equipes, enquanto a partida ainda está em andamento.

No estudo (Liu et al., 2020) os autores utilizam Deep Reinforcement Learning (DRL) com o intuito de aprender dinamicas complexas nas análises do futebol profissional. Para aprender uma função de valor de ação, onde se aplica valores da função aprendida, para medir o desempenho geral de um jogador pelos valores de impacto agregado de suas ações em todos os jogos em uma temporada. Os autores

concluiram que os métodos Deep RL, os quais conhecidamente possuem ótimos resultados em jogos de tabuleiro, como o xadrez, também são altamente promissores em esportes coletivos físicos.

3 Análise Multivariada

As análises no futebol são abordadas de forma mais geral, o livro *"The Numbers Game: Why everything you know about football is wrong"* que examinou estatísticas em larga escala respondendo a perguntas como "a posse é importante?"ou "Como o salário dos jogadores e dos treinadores impacta no sucesso da equipe?"(Sally e Anderson, 2013). McHale e Scarf desenvolveram um modelo de previsão de correspondência e, em (Mchale et al., 2012), um sistema de avaliação foi desenvolvido para o *Barclay's Premier League* usando vários sub-índices. A abordagem usa uma variedade de dados de desempenho de jogadores individuais (como gols, chutes, assistências, desarmes), a fim de determinar um 'index' o qual mede a contribuição individual para o sucesso da equipe.

O artigo (Schultze e Wellbrock, 2018) utiliza uma ideia mais simples: sempre que há gols da equipe, todos os jogadores em quadra no momento do evento recebem pontos positivos. Sempre que a equipe sofre gols, os jogadores que estão jogando recebem pontos negativos. Consequentemente, jogadores que têm pontuações positivas nessa métrica estavam em quadra quando a equipe teve um bom desempenho. A principal contribuição desse sistema é que pode ser uma maneira eficiente de avaliar rapidamente o jogador e o desempenho da equipe considerando mais uma informação, além da estatística descritiva tradicional (Schultze e Wellbrock, 2018).

Em muitos esportes, a técnica é considerada um dos principais fatores para o sucesso competitivo e uma das principais características que marcam os melhores atletas (Lees, 2002). Técnica neste contexto denota o padrão de um conjunto de movimentos empregado por atletas individuais em situações padrão do esporte praticado. A técnica individual de um atleta surge como um padrão coordenativo específico após prática extensiva.

O autor do estudo (Gløersen et al., 2018) investigou técnicas de atletas individuais, o pesquisador utilizou técnicas de componentes principais para analisar variáveis relacionadas à caracterísicas de atletas profissionais de esqui. Fizeram parte dessa pesquisa somente esquiadores noruegueses, sendo alguns deles os melhores do mundo. O estudo tinha como objetivo utilizar diversas covariávies desses atletas, a fim de identificar quais eram as mais importantes, as que melhor explicassem o sucesso dos esquiadores. Segundo o autor, foi possível encontrar 5 movimentos corporais que influenciam em aproximadamente 96% da variação presente na postura dos atletas, a qual é fator determinantes para o sucesso no esporte.

Outro esporte onde as análises multivariadas são muito utilizadas é o basquete, no artigo (Hoffman e Joseph, 2017) o autor utiliza um banco de dados com 12 variáveis na análisa dos 87 times da NBA, com o objeitvo de identificar quais clubes

irão se classificar para os *playoffs*, o autor utilizada técnicas de componentes principais, procurando diminuir a diminesionalidade da base de dados. Utilizou-se 5 componentes, os quais explicam aproximadamente 70% da variação do banco.

3.1 Análise Componentes Principais

Principal Component Analysis (PCA) é um procedimento matemático que utiliza uma transformação ortogonal que produz uma representação em baixa dimensão de um conjunto de dados correlacionados linearmente, essa nova representação é formada por um conjunto de componentes não correlacionados (James et al., 2013).

Os componentes principais Z_i, são combinações lineares padronizadas de um conjunto de recursos $X_1, X_2, ..., X_p$:

$$\begin{aligned} Z_1 &= \phi_{11}X_1 + \phi_{21}X_2 + \ldots + \phi_{1p}X_p \\ Z_2 &= \phi_{21}X_1 + \phi_{22}X_2 + \ldots + \phi_{2p}X_p \\ &\vdots \\ Z_p &= \phi_{p1}X_1 + \phi_{p2}X_2 + \ldots + \phi_{pp}X_p. \end{aligned}$$

Padronizando queremos dizer que $\sum_{j=1}^{p} \phi_{1j}^2 = 1$, os elementos ϕ referem-se como as cargas são distribuídas entre as covariáveis X_i. O primeiro componente principal Z_1, representa a maior quantidade de variabilidade presente nos dados. Uma das principais vantagens da técnica de análise de componentes principais possibilita é reduzir a dimensionalidade de tais conjuntos de dados, perdendo a interpretabilidade, mas ao mesmo tempo minimizando o número de covariáveis do modelo, concentrando a variabilidade contida nos dados em um número $k < p$ de componentes. Esse novo conjunto de covariáveis são não correlacionadas, representam diferentes proporções da variância presente nos dados (**?**).

3.2 Análise de *Cluster*

A análise de *cluster* tem como objetivo buscar uma partição dos dados em grupos distintos, espera-se que as observações dentro de cada grupo sejam bastante semelhantes entre si e elementos em grupos diferentes sejam heterogêneos em relacão a essas mesmas características (James et al., 2013). A medida de similariedade é utilizada para avaliar o quanto duas observações são distantes, o quanto são diferentes dado alguma métrica. Citaremos algumas:

3.2.1 Distância Euclidiana

É a medida de distância mais comum, a distância euclidiana pode ser explicada como o comprimento de um segmento conectando dois pontos. A fórmula é bastante direta, pois a distância é calculada a partir das coordenadas cartesianas dos pontos, usando o teorema de Pitágoras. Uma das desvantagens da distância euclidiana é que conforme a dimensionalidade dos dados aumenta, menos a métrica consegue captar a dissimilaridades entre os dados. Embora muitas outras medidas tenham sido desenvolvidas para compensar as desvantagens da distância euclidiana, ela ainda

é uma das medidas de distância mais usadas, alguns dos motivos são: facilidade matemática, seu uso é muito intuitivo e mostra ótimos resultados em muitos casos (Grootendorst, Maarten, 2021).

Dado pontos bidimensionais (x_1,y_1) e (x_2,y_2) a distância é calculada como:

$$d_e(x_1, x_2) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \tag{3.1}$$

3.2.2 Distância generalizada (ponderada)

A Distância generalizada (ponderada) entre duas observações x_1 e y_1

$$d_g(x_1, y_1) = \sqrt{(x_1 - y_1)' A (x_1 - y_1)} \tag{3.2}$$

onde A é matriz de ponderações, positiva definida.

- Quando $A = I$, tem-se a distância euclidiana.
- Quando $A = S^{-1}$, tem-se a distância de Mahalanobis.
- Quando $A = diag(1/p)$, tem-se a distância euclidiana média.

3.2.3 Distância Minkowsky

A Distância generalizada (ponderada) entre duas observações x_l e x_k

$$d_m(x_1, y_1) = \sqrt{\sum_{i=1}^{p} (w_1 |x_{1l} - y_{i1}|^{\lambda})^{1/\lambda}} \tag{3.3}$$

onde $w_i's$ são os pesos de ponderação para as variáveis

- Quando $\lambda = 2$ e $w_i = 1$ para todo $i = 1, 2, .., p$, tem-se a distância Euclidiana.

3.2.4 Similaridade por cosseno

A Similaridade por cosseno é uma medida entre dois vetores num espaço vetorial que avalia o valor do cosseno do ângulo compreendido entre eles. Esta função trigonométrica proporciona um valor igual a 1 se o ângulo compreendido é zero, isto é se ambos vetores apontam a um mesmo lugar

$$cos(\Theta) = \frac{x.y}{||x||.||y||} = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2} \sqrt{\sum_{i=1}^{n} y_i^2}} \tag{3.4}$$

3.2.5 *Clusters* Hierárquicos

A análise de *cluster* hierárquica produz um conjunto exclusivo de categorias ou grupos, fazendo com que em cada etapa, começando com a matriz de correlação, todos os *clusters* e variáveis não agrupadas são testados em todos os pares possíveis, e aquele par que produzir a intercorrelação média mais alta dentro do agrupamento de ensaio é escolhido como o novo agrupamento Jarman (2020). Em contraste com outros tipos de análise de agrupamento em que um único conjunto de *clusters* mutuamente exclusivos e exaustivos é formado, esta técnica prossegue sequencialmente,

de *clusters* mais estreitos e menos inclusivos, para clusters maiores mais inclusivos e continua até que todas as variáveis sejam agrupadas em um único grupo (Jarman, 2020). Os métodos hierárquicos são técnicas onde os dados são particionados sucessivamente, produzindo uma representação hierárquica dos agrupamentos, esses métodos requerem uma matriz contendo as métricas de distância entre os agrupamentos em cada estágio do algoritmo.(Jarman, 2020)

3.2.6 *Clusters* Não Hierárquicos

Diferentemente dos procedimentos de agrupamento hierárquico, os métodos de agrupamento não hierárquico precisam que o usuário especifique com antecedência o número de *clusters*. O algoritmo de agrupamento não hierárquico mais utilizado é o *k -Means* (Giordani et al., 2020).

Método K-Means

Com o método *k-means* procura-se encontrar *clusters* onde a variação *intra-cluster* seja a menor possível.

$$WCV(C_k) = (\sum_{k=1}^{k} \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2), \tag{3.5}$$

onde $|C_k|$ denota o número de observações no k-ésimo cluster e p é o número de variáveis.

Esta variação é uma medida $WCV(C_k)$ do montante pelo qual as observações dentro de um *cluster* diferem uma da outra. Normalmente para calcular essa medida, usamos a distância Euclidiana.

O artigo (Gómez et al., 2019) utiliza análises de *cluster* para separar os jogadores em suas diferentes funções e posteriormente comparar se há diferença de *performace* entre jogadores com contratos recém renovados e jogadores com contratos perto do vencimento. O autor utiliza *k-means* para encontrar os *clusters* mais homogêneos, para calcular a semelhança entre os grupos foi escolhida a medida de distância log-verossimilhança. O estudo possibilitou um melhor entendimento dos efeitos da assinatura de um novo contrato por temporadas consecutivas. Em contraste com a percepção comum entre os fãs de esportes de que os jogadores se tornam "preguiçosos"e se esforçam menos, após assinarem um contrato longo ou próximo do vencimento do contrato atual, os resultados do estudo não fornecem evidências para confirmar essa hipótese (Gómez et al., 2019).

Método Silhueta

Dentre as abordagens existentes para auxiliar na decisão do número de grupos que melhor separa o banco de dados, será utilizado o método Silhueta, o método também permite avaliar os particionamentos encontrados, além de visualizar graficamente os agrupamentos.

A silhueta é um gráfico onde para cada objeto (indivíduo) do cluster, representado por i, calcula-se o valor $s(i)$, composta pelos valores $a(i)$, que representa a dissimilaridade média do objeto i em relação a todos os objetos do mesmo grupo C, e $b(i)$ é a dissimilaridade média entre o objeto i em relação a todos os objetos do grupo de vizinhos mais próximos à ele.

$$s(i) = \frac{b(i) - a(i))}{max(a(i), b(i))} \tag{3.6}$$

O valor de $s(i)$ varia no intervalo entre -1 e 1, sendo adimensional. Quando $s(i) \approx 1$, significa que o objeto i foi bem classificado no grupo C, então:

$$a(i) < b(i)$$

Se o valor de $s(i) \approx$ -1 , significa que o objeto foi mal classificado, então:

$$a(i) > b(i).$$

4 Banco de Dados

Para este estudo, foram observado 720 partidas de futebol, 650 jogadores diferentes, as ligas selecionados foram o campeonato Espanhol e Francês, os quais representam 2 das 5 principais ligas da europa e com os maiores valores de mercado do mundo. O valor de uma liga de futebol é calculado pela soma do valor de mercado dos clubes que nela pertencem, segundo o site (Vyshakh K, 2021) as ligas de maior valor no futebol são:

1. Premier League (Inglaterra) - €8.9 Bilhões
2. Serie A (Italia) - €5.1 Bilhões
3. La Liga (Espanha) - €5 Bilhões
4. Bundesliga (Alemanha) - €4.3 Bilhões
5. Ligue 1 (France) - €3.5 Bilhões

O autor do artigo (Yi et al., 2019) aborda quais os fatores diferenciam uma liga de futebol da outra, cita que a La Liga espanhola, Premier League inglesa, Serie A italiana, Bundesliga alemã e Ligue 1 francesa são as cinco ligas de futebol profissional mais bem classificadas na União das Associações Europeias de Futebol (UEFA) e foram reconhecidas como as ligas de futebol de maior sucesso do mundo (Lago-Peñas et al., 2016).

4.1 *Web Scrapping*

Para extração da base de dados desse estudo, foi utilizado a técnica de *webscraping*, atráves do software R, os pacotes utilizados foram o RSelenium e XML. *Web scrapping* é uma técnica para extrair dados do *World Wide Web (WWW)* e salvá-los em um arquivo sistema ou banco de dados, para posterior análise. Devido à uma enorme quantidade de dados heterogêneos constantemente gerada no WWW, o método de web scraping é amplamente conhecido como uma técnica eficiente e poderosa para coletar o chamado *big data* (Bar-Ilan, 2001).

O processo de extração de dados da Internet pode ser dividido em duas etapas sequenciais, coletar recursos da web e em seguida, extrair as informações desejadas dos dados coletados. *Web scraping* pode ser usado para uma ampla variedade de cenários, como monitoramento da mudança de preços, coleta de avaliações de produtos, coleta de listagens de imóveis. Por exemplo, em uma microescala, o preço de

uma ação pode ser regularmente reduzido a fim de visualizar a mudança de preço ao longo do tempo (Case et al., 2005).

Outra grande utilidade do *web scraping* são nas notícias de mídia social, pois podem ser coletados para investigar opiniões públicas e identificar líderes de opinião (Liu e Zhao, 2017).

Segundo o autor (Zhao, 2017), embora a técnica de *web scraping* seja poderosa na coleta de grandes conjuntos de dados, dependendo do tipo de informação extraída da web, pode-se levantar questões legais relacionadas à direitos autorais.

4.2 A liga Fracesa

Fundada em 1932, a Ligue 1 (anteriormente Nacional e Divisão 1) é a principal divisão do futebol francês, com 20 equipes, totalizando um campeonato de 38 jogos, em que cada uma das equipes jogam entre si, duas vezes, uma como mandante e outra como visitante. A Ligue 1 tem promoção e rebaixamento vinculados à Ligue 2 francesa, a segunda divisão. A cada temporada três times são rebaixados da Ligue 1 e três times são promovidos da Ligue 2. Setenta e três clubes diferentes competiram desde a sua criação, com o Saint-Etienne conquistando o maior número de títulos, com o total de 10 títulos. O maior artilheiro de todos os tempos da competição é Delio Onnis (299 gols) (Sapp et al., 2018).

4.3 A liga Espanhola

Fundada em 1929, a La Liga ou Primera Division da Espanha é a principal divisão do futebol espanhol, com 20 equipes. A temporada vai de agosto a maio, e as equipes jogam entre si em casa e fora para cumprir um total de 38 jogos. La Liga tem promoção e rebaixamento vinculados à Segunda Divisão Espanhola. Três times são rebaixados da La Liga e três times são promovidos do Campeonato, a cada temporada. Sessenta e dois clubes diferentes competiram desde a sua criação, com o Real Madrid conquistando o maior número de títulos. O maior artilheiro de todos os tempos da competição é Lionel Messi (Sapp et al., 2018).

4.4 Variáveis

Para este estudo foram selecionadas 380 partidas do Campeonato Frânces da temporada de 2018 e 380 partidas do campeonato Espanhol de 2018, os dados foram retirados do site *WhoScored*. Devido a grande quantidade de partidas e de jogadores, o banco de dados foi obtido através de um algoritmo, utilizando a metodologia de *web scrapping*. Em cada uma das bases constam informações indivuais dos jogadores de cada equipe que participaram dessa partida, esse número pode variar de 11 até 18 jogadores, pois o técnico pode ter optado por não fazer nenhuma substituição ou ter feito o número máximo de substituições permitido.

Inicialmente os atletas foram divididos em três grupos principais: Atacantes, Meio-Campos e Jogadores Defensivos. O goleiro foi retirado da análise, por possuir métricas de avaliação muito opostas aos demais. O banco de dados é composto por 19 variáveis, indicadas a seguir:

- *Toques* - Número de toques na bola.
- LB Certos - *Line Break* são as jogadas que o passe do atleta "quebra"a linha defensiva do time adversário, gerando uma possível chance de atacar.
- *Intercept* - Número de Interceptações
- *Tack* - Número de Combates/"carrinho"
- Cabeceio - Número de cabeceios executados.
- *AereoDuelo* - Número de duelos aéreos vencidos.
- Bloqueios - Número de bloqueios executados.
- Chances - Número de Chances Criadas.
- CruzCertos - Número de Cruzamentos Certos.
- Cruzamentos - Número de Cruzamentos.
- PassesChaves - Número de passes decisivos.
- *Winx1* - Número de Duelos 1 vs 1 vencidos.
- FaltaSofrida - Número de faltas sofridas.
- Dribles - Número de dribles.
- *Thb* Certos - Número de Roubos de Bola seguidos de posse.
- *Thb* - Número de Roubos de Bola
- Disputas - Número de Disputas
- Chutes - Número de chutes ao gol.
- Assistências - Número de assistências para gol.
- ChutesLongos - Número de chutes executados de longe.
- Faltas - Número de faltas cometidas.
- *Team* - Equipe do atleta.
- *Minutes* - Minutos jogados.

Foram retirados da base todos atletas que jogaram menos de 500 minutos na temporada, também é importante ressaltar que as variáveis foram relacionadas com os minutos jogados por cada atleta, a fim de realizarmos comparações mais coerentes. As variáveis foram padronizadas utilizando o quartil e a variância, para posteriormente utilizarmos técnicas de componentes principais.

5 Resultados

Inicialmente foi realizada uma análise descritiva, a fim de entender algumas características do banco de dados. Foram selecionadas 3 variáveis de origem ofensiva (Chutes, Passes Chaves e Dribles), 2 variáveis defensivas (Interceptação e Roubo de Bola) e 1 varíavel geral (Passes), podemos perceber na Figura 5.1 que nas duas ligas, as características por posições são parecidas. A mediana da variável chutes, para os jogadores ofensivos, possui valores muito próximos. Entre as diferenças detaca-se a variável dribles, dentre os jogadores do meio-campo, a mediana da variável é superior para atletas do campeonato espanhol, além disso, os jogadores com maior destaque para essa variável, representados pelos *outliers*, são mais frequentes, sugerindo que os jogadores dessa liga sejam mais habilidosos, pois tendem a executar mais dribles. Na variável Roubos de Bola, notamos que os atletas do campeonato espanhol também são mais efetivos que os do campeonato frânces.

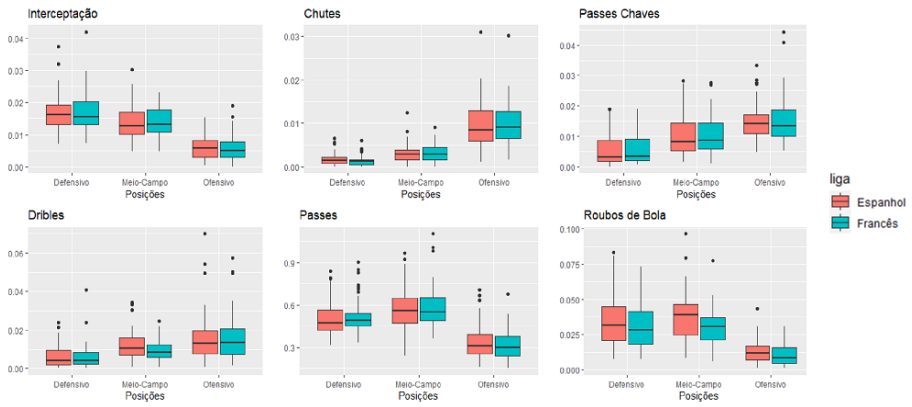

Figura 5.1: Box Plot de algumas variáveis, dado a posição do jogador e a liga a qual ele pertence.

5.1 Análise Jogadores Ofensivos

5.1.1 Componentes Principais

Observando a Figura 5.2 percebemos que existem variáveis com alta correlação, os dados apresentam multicolinearidade. Uma das técnicas para lidar com esse desafio é o PCA, onde é calculada a decomposição espectral da matriz de correlação dos dados observados, essa transformação gera um novo conjunto, composto por

combinações lineares das variávies, os componentes principais. Tais componentes, são ortogonais entre si, ou seja, linearmente independentes. Outro benefício do PCA, é a possibilidade de redução de dimensionalidade, pois é possível representar a variabilidade contida nos dados, atráves de um número de componentes inferior ao número de variáveis inicialmente observada. Podemos perceber na Tabela 5.1, que no componente 8 temos aproximadamente 85% de toda variância dos dados. Por isso, as variáveis das análises a seguir, serão referentes aos 8 primeiros componentes principais. Aaaaa

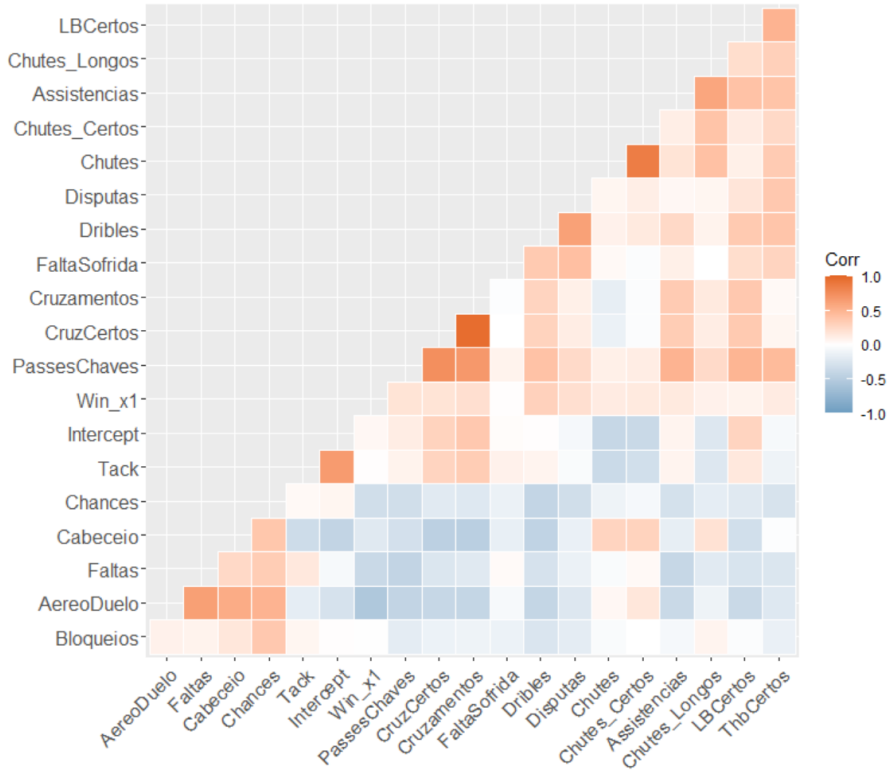

Figura 5.2: Correlação das variáveis entre os jogadores ofensivos.

Podemos avaliar a carga de cada componente, a fim de identificar, quais as principais características dos jogadores predominam dentro dos componentes. Na Figura 5.3 podemos destacar que no primeiro componente as varíaveis de criação de jogadas, como assistências, passes chaves são consideradas muito importantes. Já no cluster dois temos uma correlação muito forte entre o componente e as variávies de finalização, como chutes e cabeceio.

Tabela 5.1: Importância Pca.

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
Proporção da Variância	0,291	0,163	0,114	0,074	0,068	0,057	0,048	0,039
Proporção Acumulada	0,291	0,454	0,568	0,642	0,711	0,768	0,816	0,855

5.1.2 K-means

Após a criação dos componentes principais, buscou-se identificar, por meio do algoritmo *k-means*, diferentes grupos de jogadores ofensivos. Para auxiliar na decisão

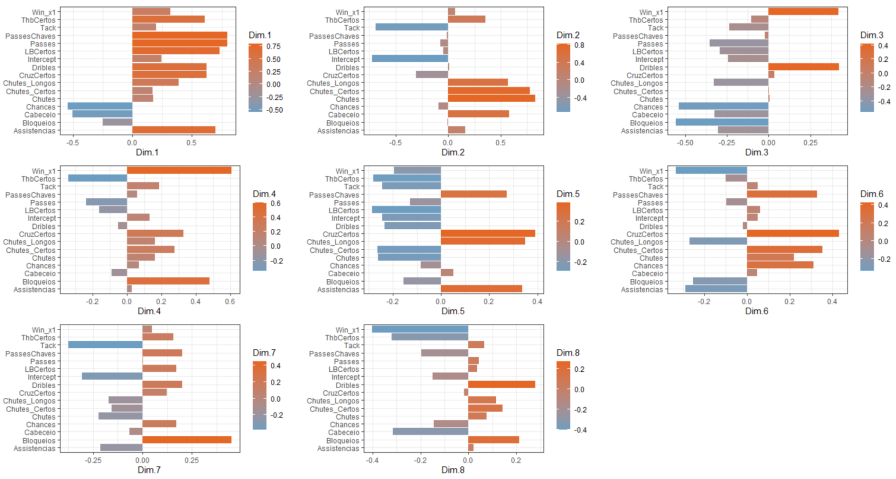

Figura 5.3: Importância das variáveis dentro dos componentes principais para os jogadores ofensivos.

da quantidade de *clusters*, o método de *silhouette*, aponta um número ótimo de grupos igual a 2 (Figura 5.4).

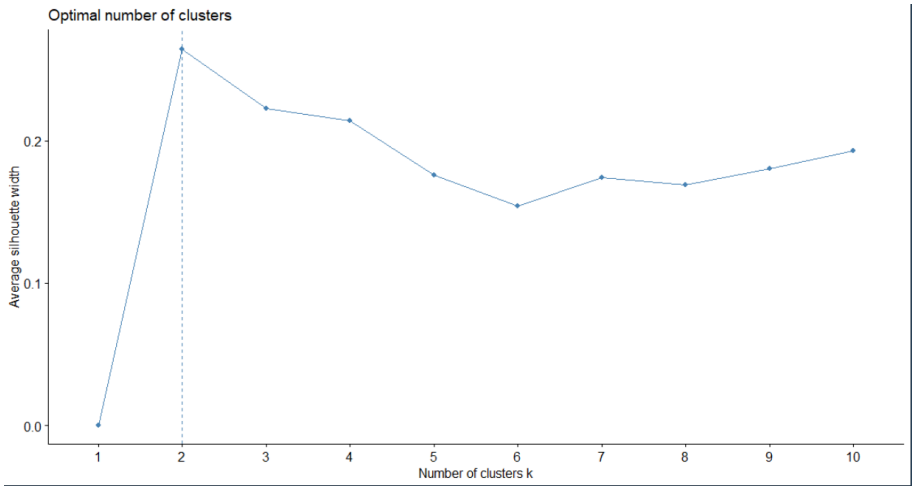

Figura 5.4: Gráfico de silhouette para identificação de número ótimo de *clusters* para os jogadores ofensivos.

Na Figura 5.5, é possível analisar os *clusters* dos jogadores ofensivos com um grande destaque para os Jogadores Messi e Neymar, os dois estão muito distantes do centroides do seu cluster.

Notamos que no *cluster* 2 estão os jogadores mais criativos, atacantes com maior capacidade de criar jogadas, o que fica evidente quando alinhamos a importância de cada componente (Figura 5.6). O *cluster* 2 é impactado positivamente pelo PC1, componente representado principalmente por variáveis de criação de jogadas como: Passes Chaves e Asssitência.

Em contrapartida no cluster 1 concentram-se jogadores de maior poder de finalização como: Luis Suárez, Karim Benzema e Mbappé, esse grupo de atletas é impactado positivamente pelo componente PC2. As principais varíáveis desse componente são relativas à finalização, como chutes e cabeceio.

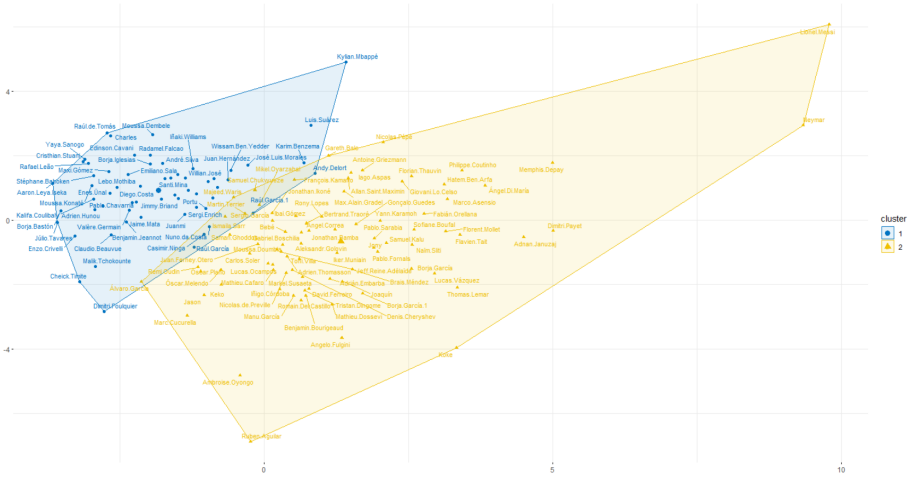

Figura 5.5: *Cluster* entre os jogadores ofensivos.

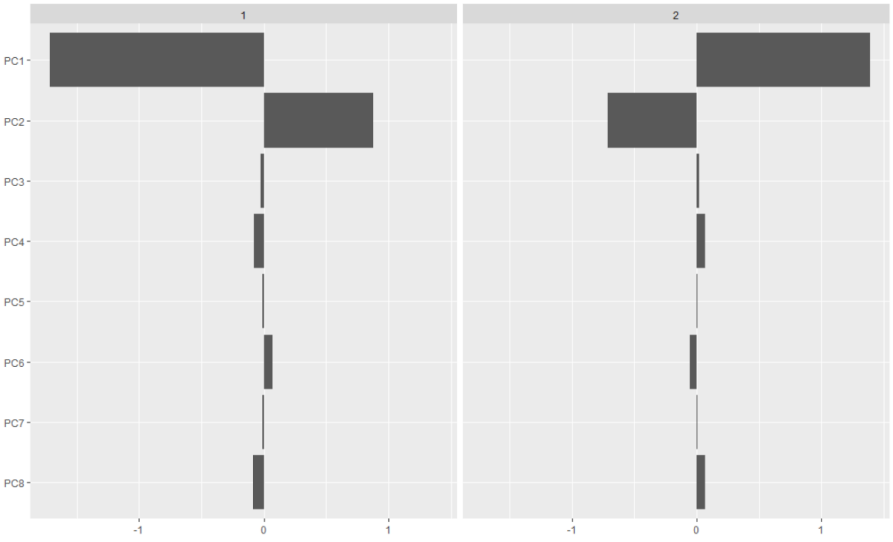

Figura 5.6: Importância de cada componente dentro dos *clusters* de jogadores ofensivos.

5.1.3 Análise Cluster 1

No *cluster* 1 dos jogadores ofensivos destacam-se aqueles com bom poder de finalização, geralmente os artilheiros dos times, os centroavantes. O jogador destaque foi o frânces Mbappé, do Paris Saint Germain, duas das suas principais características são o Chute e os Dribles.

Na Figura 5.7 estão selecionados os jogadores que compõem o grupo 1, as coordenadas x e y indicam o primeiro e segundo componentes, respectivamente. Nesse gráfico também estão destacados os dois jogadores mais próximos do jogador de referência, Mbappé, utilizando como métrica a distância euclidiana.

Na Figura 5.8 foi utilizado o gráfico de radar a fim de comparar os 3 jogadores em destaque. Percebe-se que mesmo considerandos os 3 melhores jogadores dentro do cluster, o Frânces Mbappé, é superior no quesito chutes e chutes certos. Importante ressaltar que as varíaveis com alta correlação, as quais visualizamos na Figura 5.2, não foram representadas no gráfico de radar.

Atráves da distância euclidiana encontramos jogadores com qualidades seme-

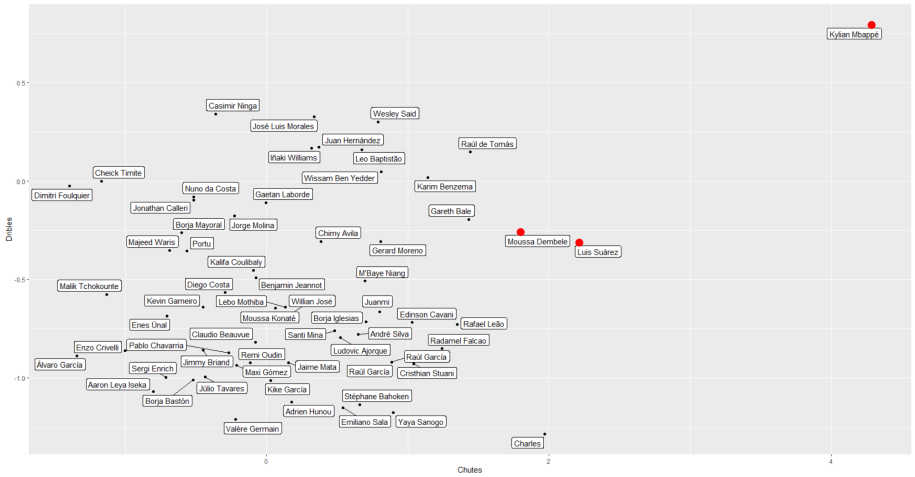

Figura 5.7: Gráfico dispersão entre atletas ofensivos do cluster 1. Com destaque para os jogador mais semelhantes atráves da distância euclidiana.

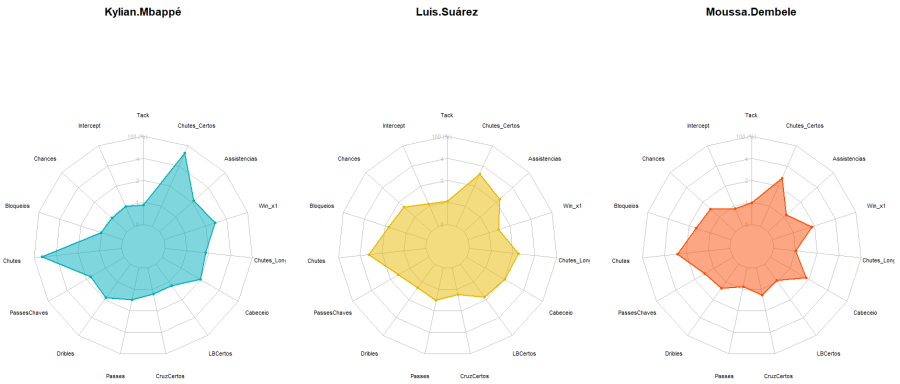

Figura 5.8: Gráfico de radar entre atletas ofensivos do cluster 1. Com destaque para os jogador mais semelhantes atráves da distância euclidiana.

lhantes, por essa métrica, se um time visa contratar um centroavante, as melhores opções seriam Mbappé e Suárez, atletas com alto desempenho e alto valor de mercado. Entretanto, em casos que o time deseja fazer alguma contratação com um menor investimento, deve-se procurar alternativas parecidas com os atletas em questão, porém com um menor custo.

Uma nova abordagem, é calcular a similaridade entre os componentes dos difenrentes jogadores, por meio da distância cosseno, onde procura-se atletas com o valor de similaridade mais próximo de 1 com o jogador de referência, isto é, se ambos vetores apontam a um mesmo lugar. A similaridade por cosseno, proposta nesse estudo pode ser útil na busca de jogadores, com mesmas características que o jogador de referência do cluster, mas com menor valor de mercado. Visualizando a Figura 5.9 além do jogador frânces Mbappé, destacam-se jogadores um pouco inferiores, com características técnicas parecidas, comparação que pode ser visualizada na Figura 5.10.

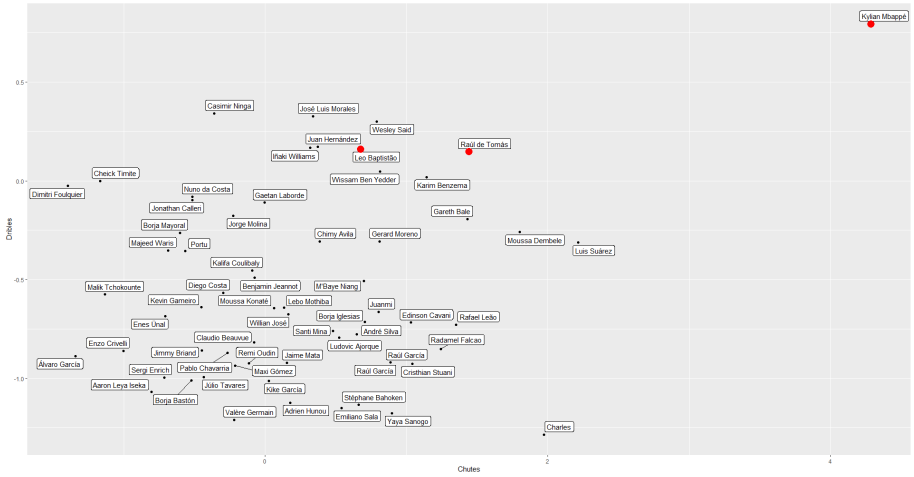

Figura 5.9: Gráfico dispersão entre atletas ofensivos do cluster 1. Com destaque para os jogador mais semelhantes atráves da similaridade cosseno.

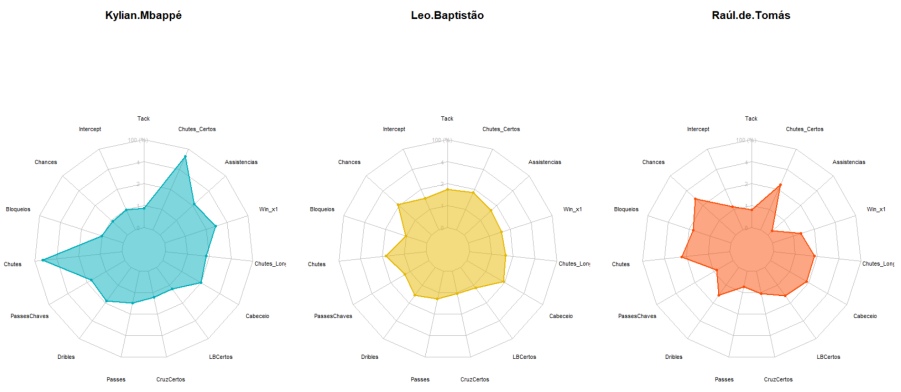

Figura 5.10: Gráfico de radar entre atletas ofensivos do cluster 1. Com destaque para os jogador mais semelhantes atráves da similaridade do cosseno.

Abaixo a Tabela 5.2 compara as similaridades dos dois jogadores com menores distâncias euclidianas e dos dois jogadores com menores distâncias cosseno entre os atletas desse *cluster*.

Tabela 5.2: Jogadores com as 2 menores distâncias para o jogador de referência do *cluster* K.Mbappé.

	Distâncias do jogador de referência K.Mbappé	
Jogador	Euclidiana	Cosseno
Luis Suárez	**2,3456**	0,9481
Dembele	2,6903	0,9457
Léo Baptistão	3,6641	**0,9998**
Raul de Tomás	3,8053	0,9997

5.1.4 Análise Cluster 2

No cluster 2 dos jogadores ofensivos destacam-se aqueles com maior poder de criação de jogadas, geralmente os atacantes que jogam um pouco atrás dos centro-avantes, atletas com uma maior movimentação durante as partidas. Neste cluster o atleta com maior performance é o argentino Lionel Messi, atacante do Barcelona. Duas das suas principais características são as assistências e Lb Certos, na Figura 5.11 fica evidente que Messi possui valores mais altos em comparação com os outros atletas, tanto para o componente 1 (criação de jogadas) quanto para o componente 2 (finalizações).

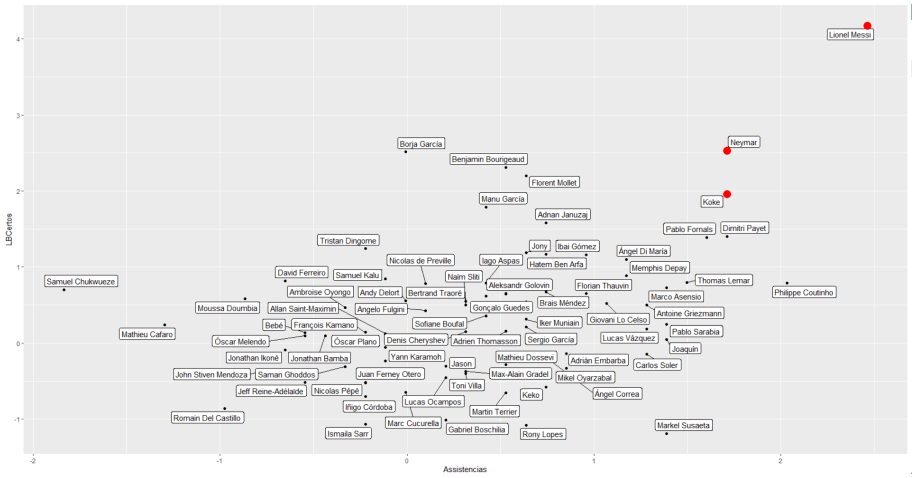

Figura 5.11: Gráfico dispersão entre atletas ofensivos do cluster 2. Com destaque para os jogador mais semelhantes atráves da distância euclidiana.

Na Figura 5.11 foram sinalizados os jogadores mais semelhantes atráves da distância euclidiana e na Figura 5.12 está o gráfico de radar dos 3 jogadores de maior significância. Percebemos que o jogador brasileiro Neymar é o mais próximo de Lionel Messi, apesar de uma vantagem do argentino, ambos possuem excelente desempenho no quesito dribles e Passes Chaves. O brasileiro fica bem atrás quando olhamos a variável LB Certos e Chutes Certos.

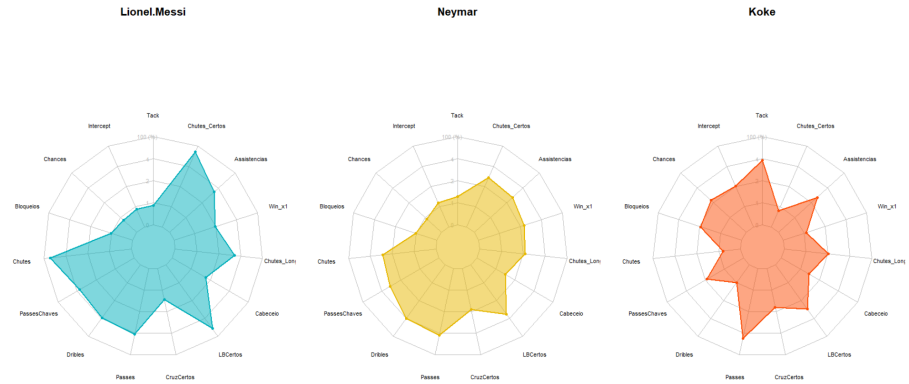

Figura 5.12: Gráfico de radar entre jogadores mais semelhantes utilizando a distância euclidiana.

A Figura 5.13 apresenta os jogadores mais similares ao jogador de referência dado a distância cosseno, em vermelho estão atletas menos conhecidos, como Naïm Sliti, que exibe maior similaridade com L.Messi. Investigando a performance do atleta percebemos que ele teve um grande destaque na temporada, comparando com outros atletas da mesma equipe, a média de minutos necessários para efetuar 1 passe chave é cerca de 135 minutos. Essa média, entre os atletas ofensivos da liga espanhola ou francesa é de 70 minutos, enquanto o atleta Naïm Sliti precisou, em média, de apenas 35 minutos para efetuar um passe chave, um desempenho muito bom.

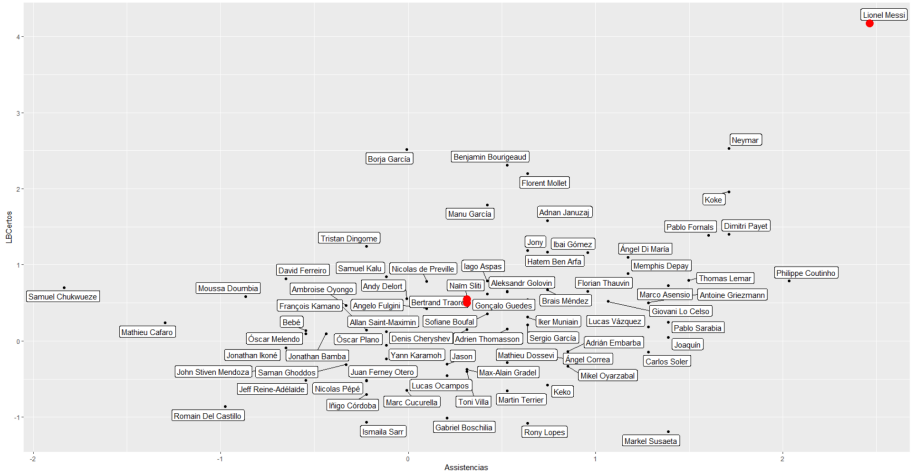

Figura 5.13: Gráfico dispersão entre atletas ofensivos do cluster 2. Com destaque para os jogador mais semelhantes atráves da similaridade cosseno.

No gráfico de radar 5.14 Naïm Sliti teve uma performance próxima de L.Messi para passes chaves, mas essa paridade não se confirma em outras medidas, apesar de Sliti possuir características técnicas semelhantes, a qualidade é muito inferior à Messi. Se analisarmos o valor de mercado de cada um, segundo o site Transfermarket.com, o jogador argentino tem valor estimado em 80 milhões de euros, o jogador tunisiano Naïm Sliti tem valor de mercado de 6,5 milhões de euros.

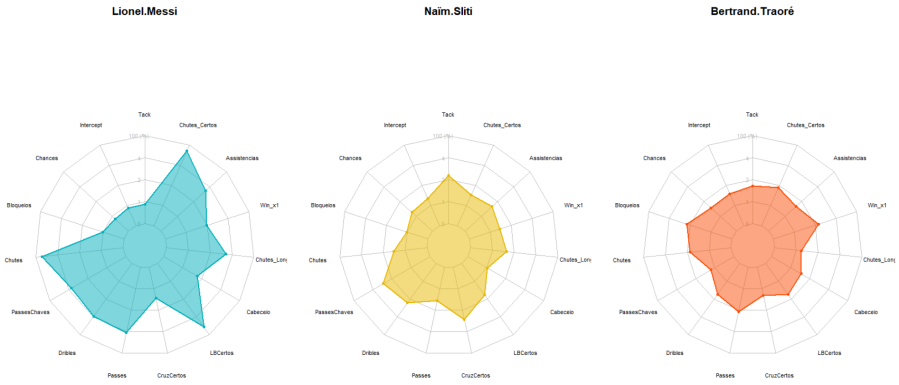

Figura 5.14: Gráfico de radar entre jogadores mais semelhantes utilizando a similaridade cosseno.

Abaixo a Tabela 5.3 compara as similaridades dos dois jogadores com menores distâncias euclidianas e dos dois jogadores com menores distâncias cosseno entre os atletas desse *cluster*.

Tabela 5.3: Jogadores com as 2 menores distâncias para o jogador de referência do *cluster* Lionel Messi.

	Distâncias do jogador de referência Lionel Messi	
Jogador	Euclidiana	Cosseno
Neymar	**1,8062**	0,9981
Koke	2,3381	0,9832
Naim Sliti	4,2115	**0.9998**
Bertrand T.	4,2534	09994

5.2 Análise Jogadores do Defensivos

5.2.1 Componentes Principais

Agora iremos analisar os jogadores defensivos, nessa região do campo estarão os zagueiros do time e os laterais, onde a principal função é desarmar e interceptar jogadas. Além dessas funções defensivas, os laterais têm grandes responsabilidades nas criações de jogadas, principalmente na métrica de cruzamentos. Observando a Figura 5.15 percebemos que existem variáveis com alta correlação, assim como foi discutido no cluster dos jogadores ofensivos, também foram utilizadas técnicas de componentes principais, a fim de diminuir a dimensionalidade e novamente será utilizado 8 componentes principais nas análises.

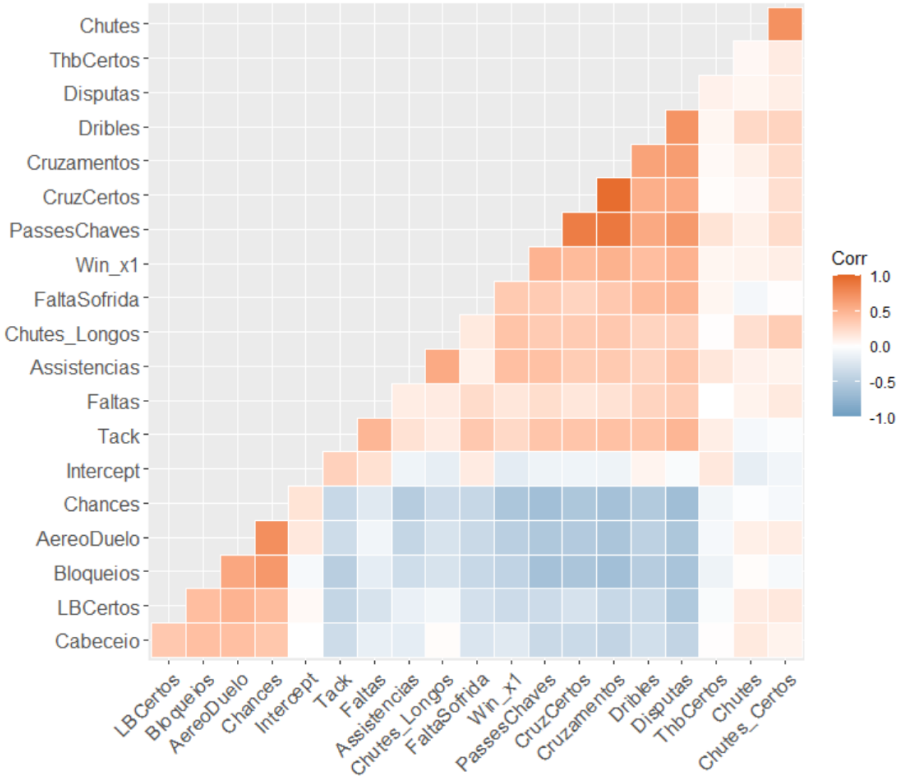

Figura 5.15: Correlação das variáveis entre os jogadores defensivos.

Na Figura 5.16 o primeiro componente contém como varíaveis mais correlacionadas os cruzamentos certos, dribles e passes chaves, características esperadas para os laterais, já o segundo componente tem uma alta correlação negativa com as principais características defensivas, são elas: Interceptação e Combates ('Tack').

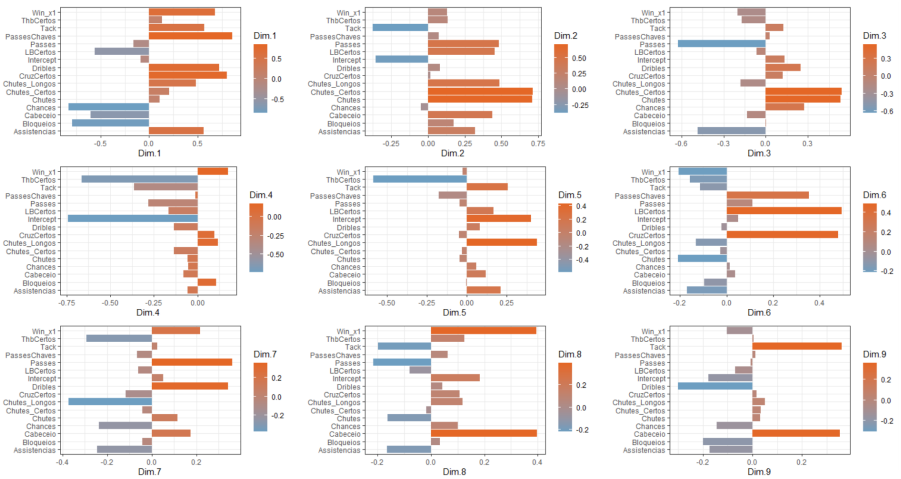

Figura 5.16: Importância das variáveis entre os jogadores defensivos.

5.2.2 K-means

Após a criação dos componentes foi feito um agrupamento para identificar diferentes *clusters* entre os jogadores defensivos, o número ótimo escolhido por meio

do método silhouete foi 2. Na Figura 5.17, é possível identificar claramente que o cluster 1 tem como grande destaque os laterais, sendo o brasileiro Marcelo, o jogador com valores que indicam melhor desempenho para o PC1. No Cluster 2 há uma forte concetração de zagueiros, com destaque para o espanhol Sérgio Ramos.

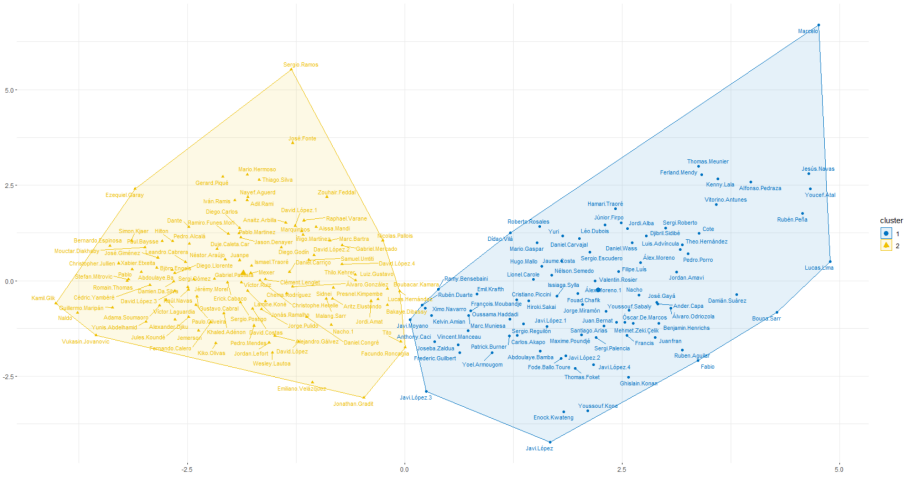

Figura 5.17: *Cluster* entre os jogadores defensivos.

5.2.3 Análise Cluster 1

Notamos que no cluster 1 estão os laterais, onde se espera jogadores com poder de criação de jogadas, bons cruzamentos e com bons fundamentos defensivos, observando 5.18, podemos notar que o cluster 1 é altamente correlacionado com o PC1, componente com grande correlação com as varíaveis de criação, por exemplo: Cruzamentos Certos, Passes Certos. Já no cluster 2, formado em grande parte por zagueiros, esperamos atletas com grande poder de desarme de jogadas. Na figura 5.18, notamos que o cluster 2 tem uma grande correlação negativa com o primeiro componente, o que confirma que nesse *cluster* temos jogadores com bons índices de bloqueios de jogadas e cabeceio, outra característica muito importante para um zagueiro.

No cluster 1 dos jogadores defensivos, concentram-se laterais. Neste cluster o jogador destaque foi o lateral brasileiro Marcelo, do Real Madrid. Duas das suas principais características são o Chute e os Dribles, sendo assim podemos observar na Figura 5.19 que comparando duas das principais características para os laterais, ter um bom índice de passes certos e ser bom driblador, para facilitar a criação de jogadas. Não há jogadores com grande destaque. Sendo assim, destacamos 3 jogadores com características muito semelhantes, utilizando como métrica para identificar semelhança a distância euclidiana 3.1

Na Figura 5.7 foram selecionados os jogadores mais semelhantes, e posteriormente, na figura 5.20 foi utilizado o gráfico de radar a fim de comparar os 3 jogadores em destaque. Percebemos que o atleta Marcelo possui altos valores nas variáveis referentes a chutes, porém nas outras características os outros jogadores são muito próximos, na variável de cruzamentos certos, o jogador Marcelo perde para os demais, onde há predomínio do lateral Cote, com o desempenho em cruzamentos superior aos demais.

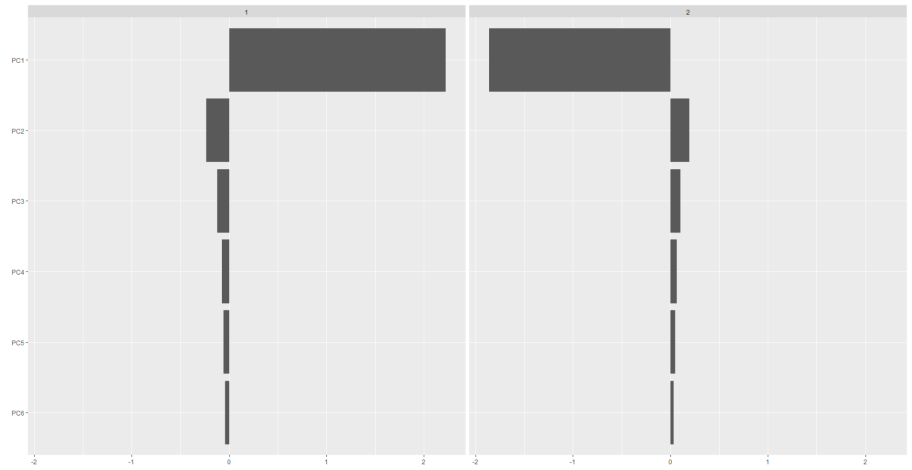

Figura 5.18: Relação entre os componentes e clusters dos jogadores defensivos.Importância de cada componente dentro dos *clusters*.

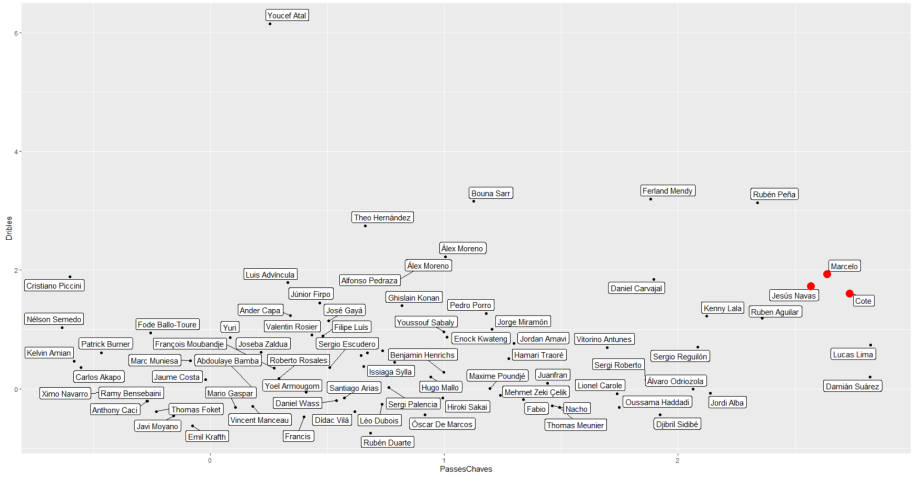

Figura 5.19: Gráfico de dispersão entre jogadores defensivos mais semelhantes utilizando a distância euclidiana.

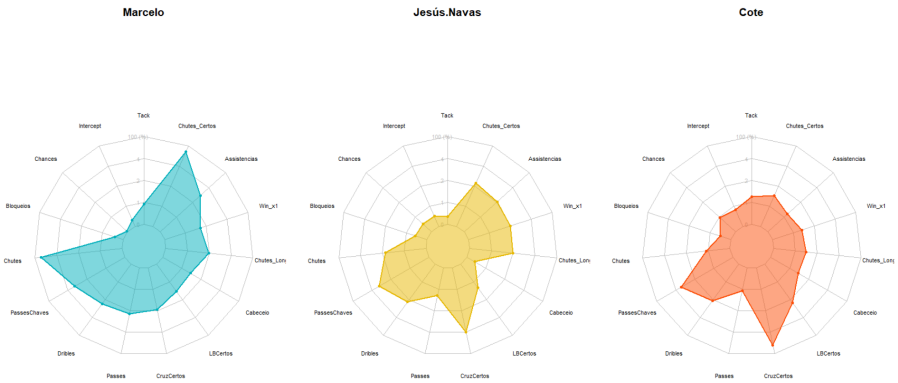

Figura 5.20: Gráfico de radar entre jogadores defensivos mais semelhantes utilizando a distância euclidiana.

Visualizando a Figura 5.21 percebemos que o lateral Jesus Navas também também é muito similar ao atleta de referência Marcelo. Observando o gráfico 5.21

notamos que o atleta Sergio Escudero é consideravelmente inferior aos demais, porém esse atleta tem uma tendência a ser um bom finalizador, característica muito marcante no atleta Marcelo.

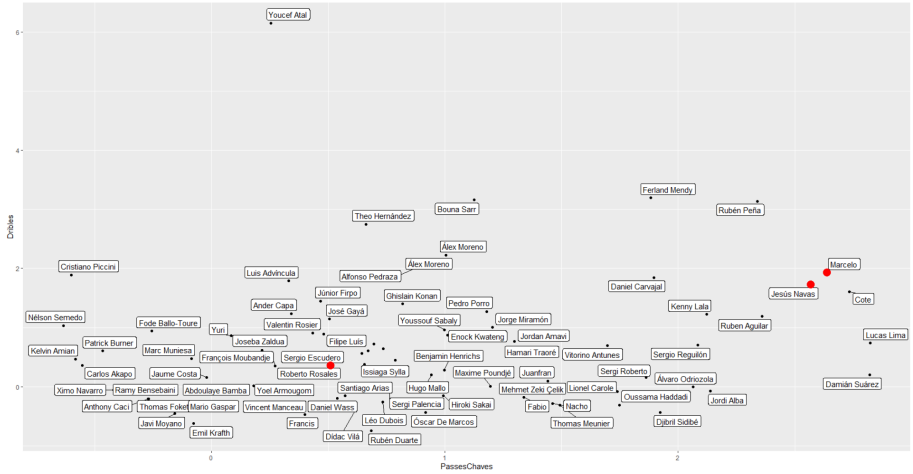

Figura 5.21: Gráfico de dispersão entre jogadores defensivos mais semelhantes utilizando a similaridade cosseno.

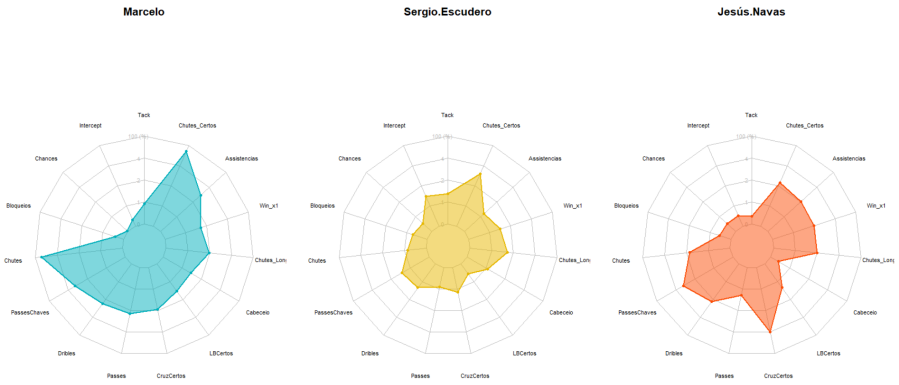

Figura 5.22: Gráfico de radar entre jogadores defensivos mais semelhantes utilizando a similaridade cosseno.

Abaixo a Tabela 5.4 compara as similaridades dos dois jogadores com menores distâncias euclidianas e dos dois jogadores com menores distâncias cosseno entre os atletas desse *cluster*.

Tabela 5.4: Jogadores com as 2 menores distâncias para o jogador de referência do *cluster* Marcelo.

	Distâncias do jogador de referência Marcelo	
Jogador	Euclidiana	Cosseno
Jesus Navas	**0,2134**	0,9992
Cote	0,3325	0,9952
Sergio Escudero	2,6338	**0.9997**

5.2.4 Análise Cluster 2

No cluster 2 dos jogadores defensivos destacam-se jogadores com maior poder de desarmes de jogadas, principalmente formado por zagueiros. Neste cluster o jogador destaque foi o espanhol Sergio Ramos, zagueiro do Real Madrid. Duas das suas principais características são as roubos de bola(LBCertos) e Cabeceio, podemos observar na Figura 5.23 o quão destacado ele é perante os demais jogadores desse cluster.

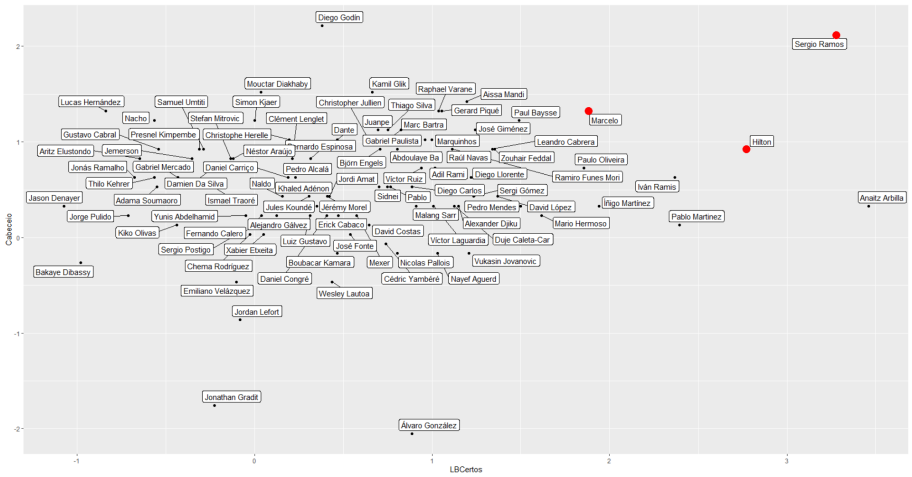

Figura 5.23: Gráfico de dispersão entre jogadores defensivos do cluster 2 mais semelhantes utilizando a distância euclidiana.

Na Figura 5.23 foram destacados os jogadores mais semelhantes, utilizando a a distância Euclidiana. E posteriormente, na figura 5.24 foi utilizado o gráfico de radar a fim de comparar os 3 jogadores em destaque. Percebemos que o jogador espanhol Sergio Ramos possui maiores valores para a variável roubos de bola, outro ponto que chama atenção é a sua grande capacidade nos passes e nos chutes. Os outros 2 atletas também tem um bom desempenho nos roubos de bola.

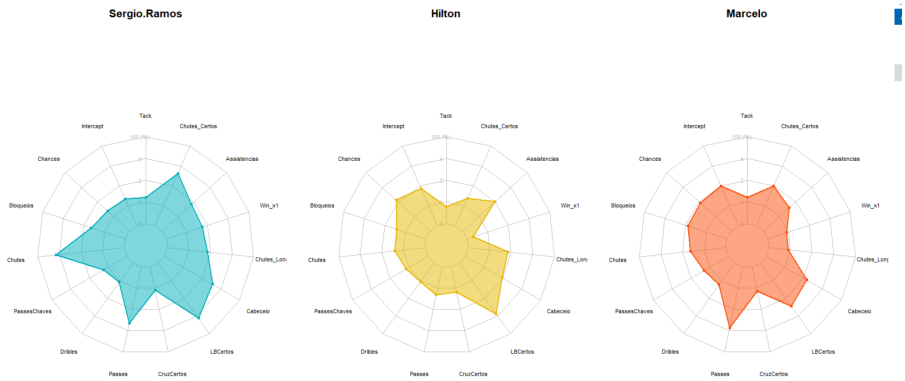

Figura 5.24: Gráfico de radar entre jogadores defensivos do cluster 2 mais semelhantes utilizando a distância euclidiana.

Visualizando na Figura 5.25, podemos observar que além do Sergio Ramos, há 2 jogadores menos conhecidos que possuem características similares, sendo que a *performance* do atleta Raúl Navas, possui a menor distância cosseno de Sergio Ramos.

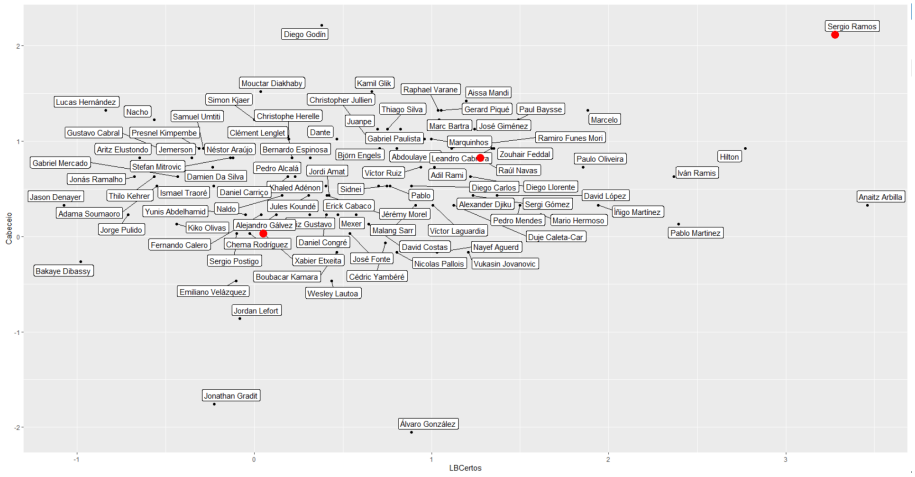

Figura 5.25: Gráfico de dispersão entre jogadores defensivos do *cluster* 2 mais semelhantes utilizando a similaridade cosseno.

No gráfico de radar 5.26 percebemos que a característica principal do atleta Raúl Navas são os passes e os roubos de bola. Em comparação com o jogador de referência, Sergio Ramos, Raul Navas possui quantidades muito menores para esses variávies, mas é importante destacar que segundo o site Transfermarkt (2021) o atleta Sergio Ramos é em torno de 28x mais caro que o atleta Raúl Navas, com isso percebemos que a similaridade do cosseno apresentou uma solução, procurando jogadores menos conhecidos.

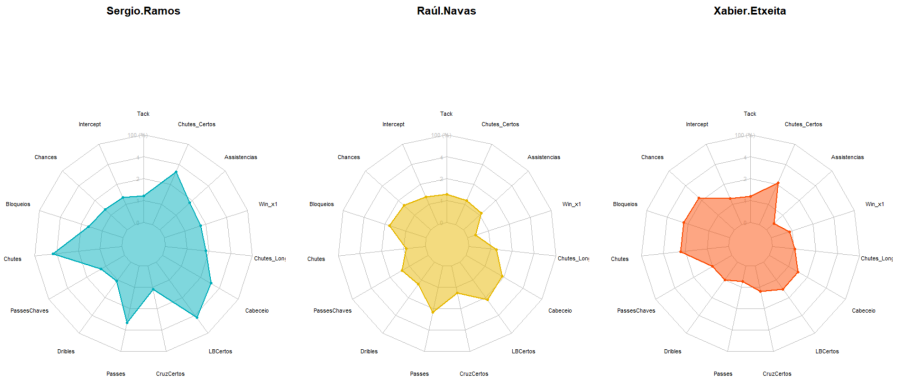

Figura 5.26: Gráfico de radar entre jogadores defensivos do *cluster* 2 mais semelhantes utilizando a similaridade cosseno.

Abaixo a Tabela 5.5 compara as similaridades dos dois jogadores com menores distâncias euclidianas e dos dois jogadores com menores distâncias cosseno entre os atletas desse *cluster*.

Tabela 5.5: Jogadores com as 2 menores distâncias para o jogador de referência do *cluster* Sergio Ramos.

	Distâncias do jogador de referência Sergio Ramos	
Jogador	Euclidiana	Cosseno
Hilton	**1,8062**	0,9687
Marcelo	2,3381	0,9992
Raul Navas	2,3847	**0.9999**
Xabier E.	3,8423	0.9994

5.3 Análise Jogadores do Meio-Campo

5.3.1 Componentes Principais

Após a análise dos jogadores de ataque dos jogadores de defesa, iremos analisar os jogadores do meio-campo. Região do campo composta por jogadores com objetivos defensivos e ofensivos. Dentre os jogadores com o características defensivas, como os volantes, onde a principal função é desarmar jogadas e facilitar a saida de bola para os jogadores do meio-campo mais ofensivos, onde esses tem como objetivo criar jogadas, procurar passes chaves e fazer gols. Observando a Figura 5.27 percebemos que existem variáveis com alta correlação, sendo assim iremos utilizar a técnica de componentes principais. Com o objetivo de analisarmos o quão correlacionada são as variáveis com aquele componente, a fim de identificar, no caso desse estudo, quais as principais características do jogadores predominam dentro dos componentes.

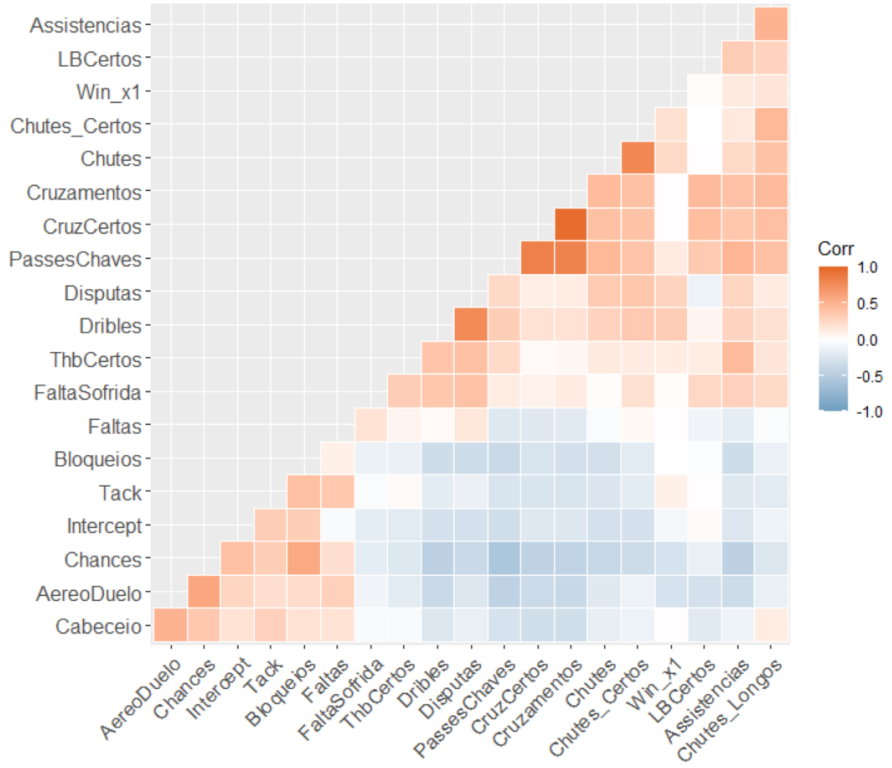

Figura 5.27: Correlação das variáveis entre os jogadores do meio-campo.

Na Figura 5.28 percebemos muito claro a diferença entre os dois primeiros componentes. No primeiro as varíaveis mais correlacionadas são interceptação, bloqueio e em contra-partida as variáveis de criação de jogadas como passes chaves, são negativamente correlacionadas com o componentes 1. Já no componente 2, é o oposto, as variávies de criação, passes, LB Certos são altamente correlacionadas com o componente. Porém há um ponto bem interessante, as variáveis de chutes ficaram com correlação negativa nos dois primeiros cluster. Essas variáveis de finalizações ficaram correlacionadas com o terceiro componente.

Após a criação dos componentes foi feito um agrupamento para identificar diferentes *clusters* dentro dos jogadores do meio-campo, foi utlizado o método *k-means* e como instrumento para decisão da quantidade de *cluster* foi utilizado o método

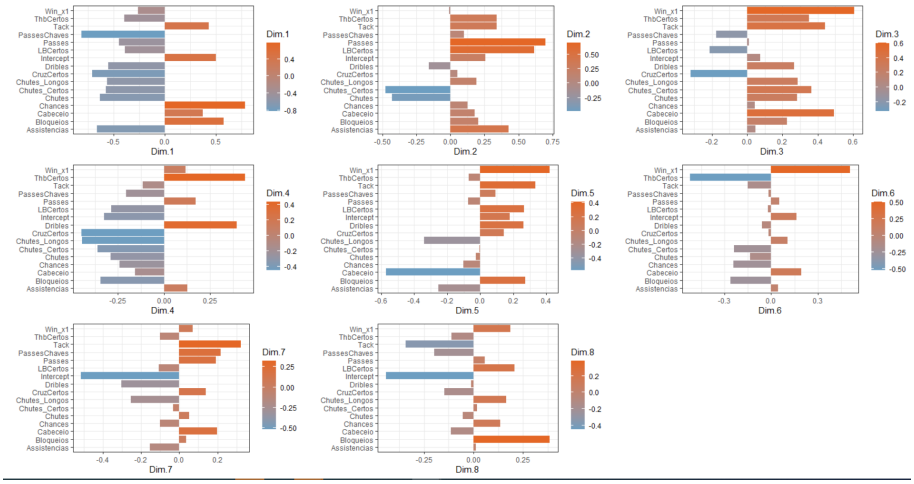

Figura 5.28: Importância das variáveis dentro dos componentes principais.

silhouette, 3.6 o número ótimo de *clusters* escolhido, foi 2.

Na Figura 5.29, é possível analisar os *clusters* dos jogadores do meio-campo podemos perceber que o atleta Toni Kroos, uma das referências do time do Real Madri, é um dos destaques.

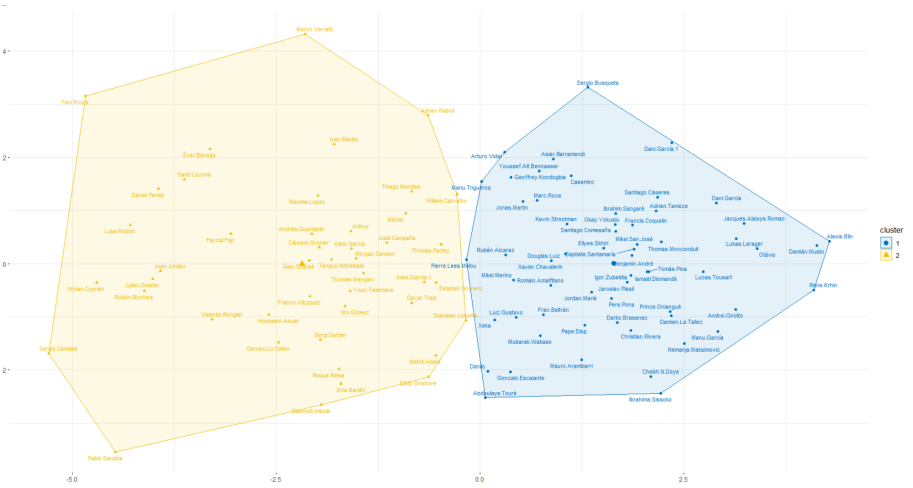

Figura 5.29: *Cluster* jogadores do meio campo.

5.3.2 K-means

Notamos que no cluster 2 estão os jogadores com maior poder de criação, o que fica evidente quando alinhamos a importância de cada componente princial dentro dos *clusters*, na Figura 5.30, observamos que o cluster 2 é impactado negativamente pelo PC1, onde foi possível observar em 5.28, que o componente 1 é correlacionado com variávies mais defensivas, como bloqueio, interceptação, este mesmo componente é negativamente correlacionado com as variáveis de criação de jogadas, como passes chaves, cruzamentos certos. Em contrapartida no cluster 1 concentram-se jogadores com maior poder defensivo, onde os principais atributos devem ser roubos de bola, desarmes, entre outros.

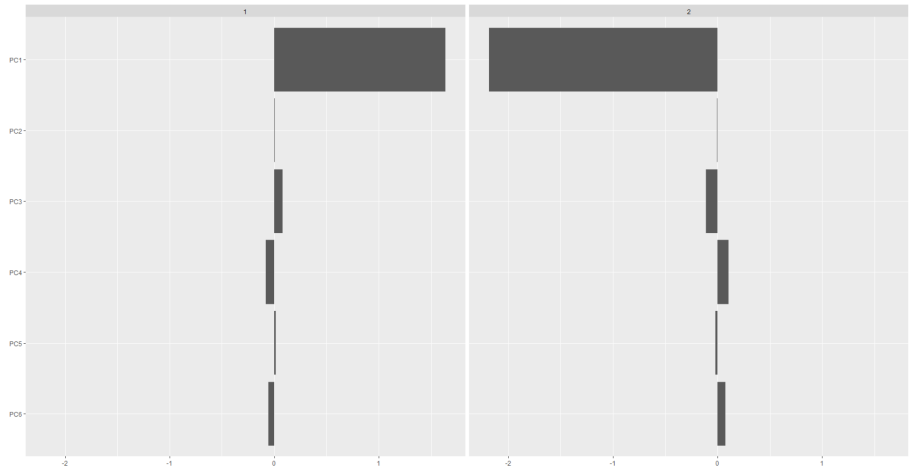

Figura 5.30: Importância de cada componente dentro dos *clusters*.

5.3.3 Análise Cluster 1

No cluster 1 dos jogadores do meio-campo destacam-se jogadores com maior poder de inteceptações, bloqueio de jogadas, jogadores com bom desempenho em características defensivas. Neste cluster o jogador destaque foi meio-campo da seleção Dinamarquesa Lukas Lerager. Duas das suas principais características são as interceptações e bloqueis, sendo assim podemos observar na Figura 5.31 que ele apresenta um certo destaque perante os demais atletas desse cluster. Nesse gráfico também estão destacados os dois jogadores mais próximos do jogador de referência, Lukas Lerager, utilizando como métrica para identificar semelhança a distância euclidiana 3.1

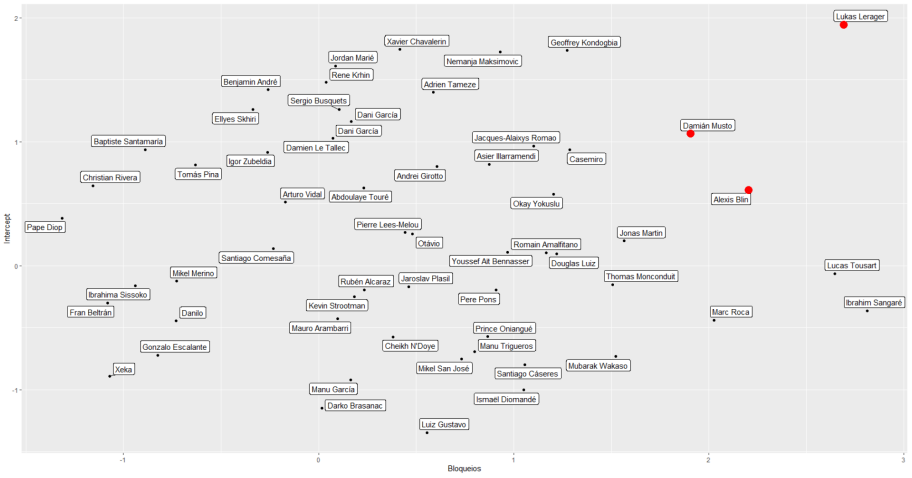

Figura 5.31: Gráfico de dispersão entre jogadores do meio-campo do *cluster* 1 mais semelhantes utilizando a distância euclidiana.

Na Figura 5.31 foram selecionados os jogadores mais semelhantes, e posteriormente, na Figura 5.32 foi utilizado o gráfico de radar a fim de comparar os 3 jogadores em destaque. Percebe-se que os 3 jogadores tem desempenhos semelhantes, com destaque bem definido para as características defensivas.

Posteriormente foi proposto fazer a mesma análise porém utilizando a similaridade por cosseno 3.4. Visualizando na Figura 5.33 percebemos que além do jogador

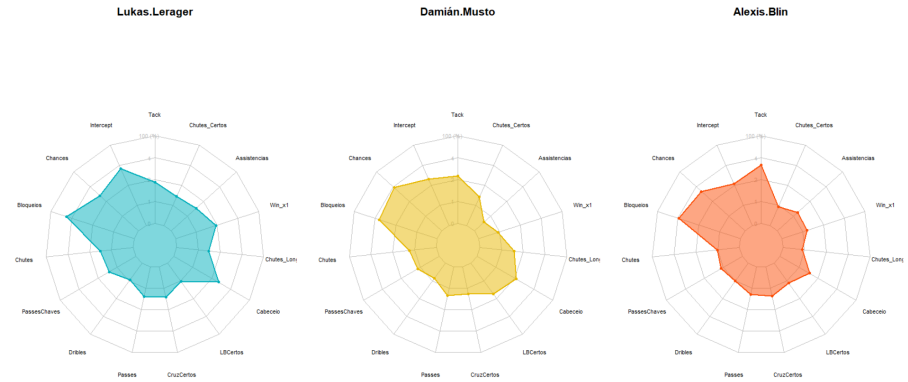

Figura 5.32: Gráfico de radar entre jogadores do meio-campo do *cluster* 1 mais semelhantes utilizando a distância euclidiana.

em destaque desse cluster, o dinamarquês, Lukas Lerager, destaca-se o jogador brasileiro, Casemiro. Visualiazando o gráfico de radar 5.34, percebemos que o atleta Casemiro parece ser um atleta mais completo, porém nas variáveis defensivas o atleta Lukas Lerager tem vantagem,

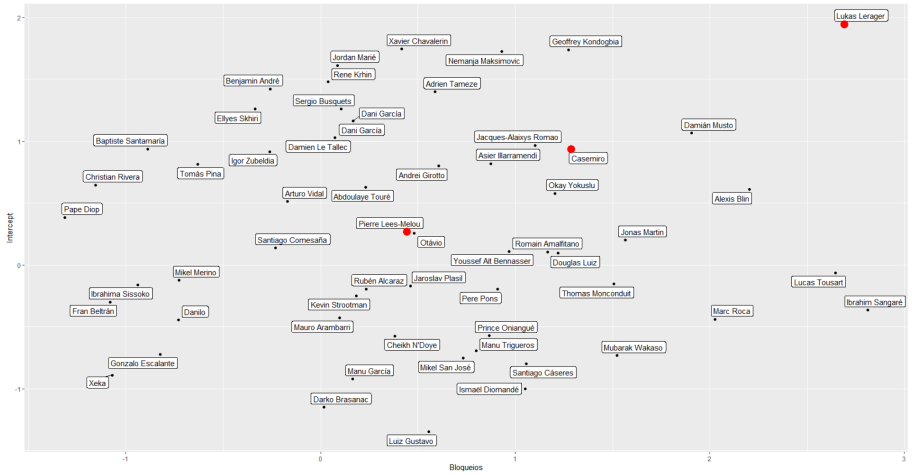

Figura 5.33: Gráfico de dispersão entre jogadores do meio-campo do *cluster* 1 mais semelhantes utilizando a similaridade cosseno

Abaixo a Tabela 5.6 compara as similaridades dos dois jogadores com menores distâncias euclidianas e dos dois jogadores com menores distâncias cosseno entre os atletas desse *cluster*.

Tabela 5.6: Jogadores com as 2 menores distâncias para o jogador de referência do *cluster* Lukas Leraker.

	Distâncias do jogador de referência Lukas Leraker	
Jogador	Euclidiana	Cosseno
D. Musto	**1,1789**	0,9931
Alex Blin	1,4224	0,9371
Casemiro	1,7291	**0.9999**
Pierre L.	2,8401	0.9994

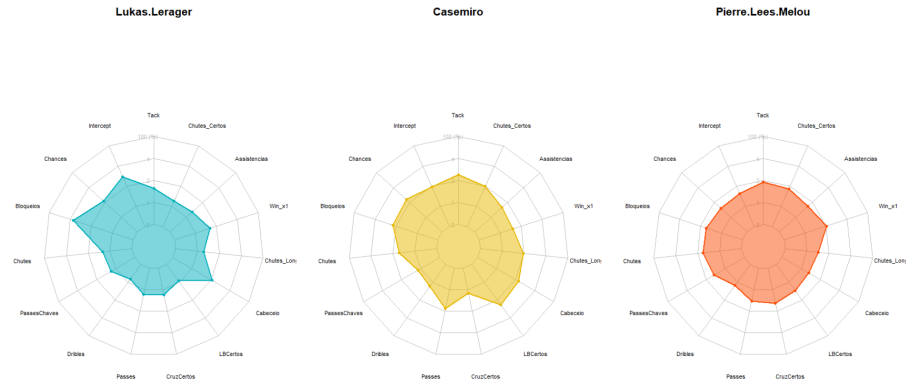

Figura 5.34: Gráfico de radar entre jogadores do meio-campo do *cluster* 1 mais semelhantes utilizando a similaridade cosseno

5.3.4 Análise Cluster 2

No cluster 2 dos jogadores do meio-campo destacam-se jogadores com maior poder de criação de jogadas, geralmente os meio-campos ofensivos. Neste cluster o jogador destaque foi o alemão Toni Kroos, meio-campo do Real Madrid. Duas das suas principais características são as PassesChaves e Chutes Longos, sendo assim podemos observar na Figura 5.35 que o atleta se destaca dentro do cluster.

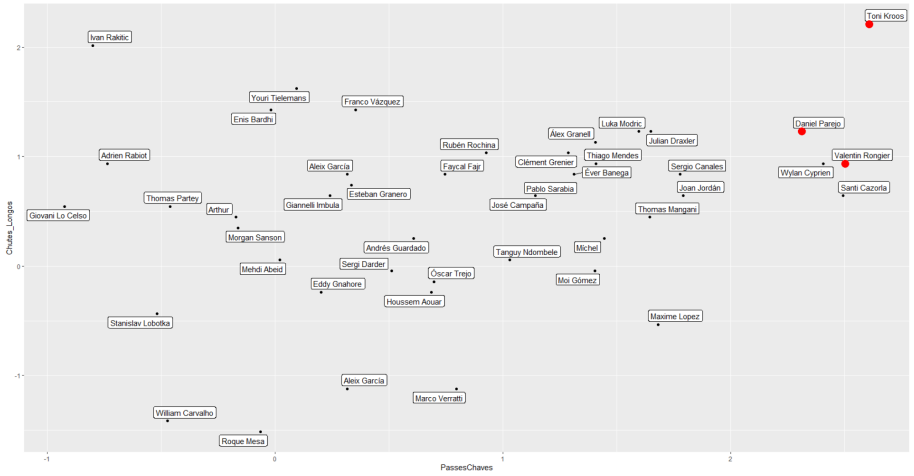

Figura 5.35: Gráfico de dispersão entre jogadores do meio-campo do *cluster* 2 mais semelhantes utilizando a distância euclidiana.

Na Figura 5.35 foram destacados os jogadores mais semelhantes, utilizando a a distância Euclidiana. E posteriormente, na Figura 5.36 foi utilizado o gráfico de radar a fim de comparar os 3 jogadores em destaque. Percebemos os 3 jogadores tem como principal característica os PassesChaves.

Analisando atráves da similaridade cosseno Visualizando na Figura 5.37. Podemos observar que além do Toni Kroos, destacam-se jogadores com desempenho inferior, porém ambos com PassesChaves como sua principal característica

No gráfico de radar 5.38 percebemos que o atleta Álex Granell possui um bom desempenho em PassesChaves e um excelente desempenho em Cruzamentos Certos, desempenho superior ao jogador de referência desse *cluster*.

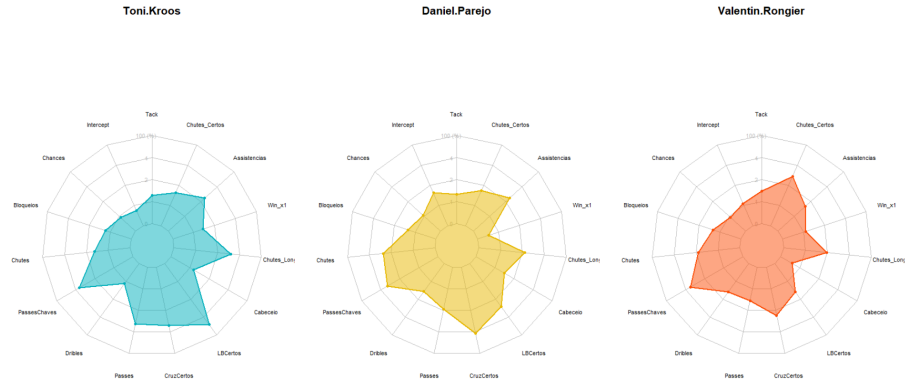

Figura 5.36: Gráfico de radar entre jogadores do meio-campo do *cluster* 2 mais semelhantes utilizando a distância euclidiana

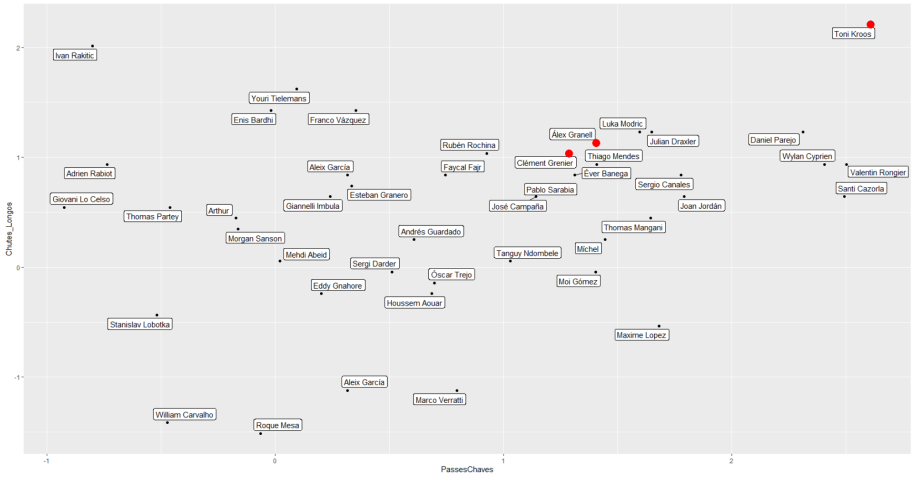

Figura 5.37: Gráfico de dispersão entre jogadores do meio-campo do *cluster* 2 mais semelhantes utilizando a similaridade cosseno

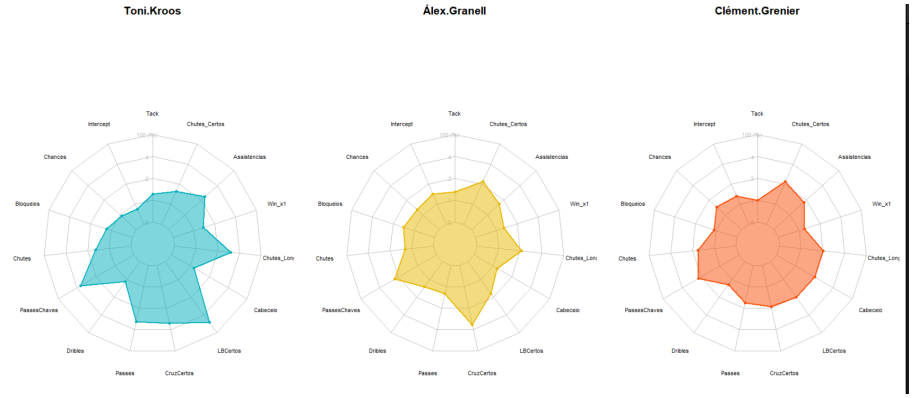

Figura 5.38: Gráfico de radar entre jogadores do meio-campo do *cluster* 2 mais semelhantes utilizando a similaridade cosseno

Abaixo a Tabela 5.7 compara as similaridades dos dois jogadores com menores distâncias euclidianas e dos dois jogadores com menores distâncias cosseno entre os atletas desse *cluster*.

Tabela 5.7: Jogadores com as 2 menores distâncias para o jogador de referência do *cluster* Toni Kroos.

	Distâncias do jogador de referência Toni Kroos	
Jogador	Euclidiana	Cosseno
Daniel P.	**1,0273**	0,9931
Valentin R.	1,2780	0,9371
Alex G.	1,6139	**0.9999**
Ciemente G.	1,7691	0.9964

6 Considerações Finais

No dia-a-dia dos profissionais de análise de desempenho dos clubes de futebol aparecem diversos questionamentos quanto a *perfomance* dos atletas do clube, possíveis jogadores substitutos e soluções externas, atráves de contratações, as quais poderiam agregar pro clube. Sendo assim, esse estudo tem como objetivo propor algumas abordagens as quais a análise de dados pode auxiliar na tomada de decisões em momentos como: Identificar jogadores com características semelhantes, quando houver perda de um atleta, propor uma possível substituição por um atleta do elenco que exerça uma função similar dentro da equipe, entre outras decisões.

Após a divisão entre atletas ofensivos, defensivos e do meio-campo, foi proposto a criação de *clusters* e utilizou-se a distância euclidiana a fim de identificar os jogadores semelhantes.

Percebemos que nosso estudo apresentou bons resultados, onde foi possível destacar os melhores jogadores, porém atráves da distância euclidiana encontramos jogadores com valor de mercado muito semelhantes.E um dos objetivos desse estudo era encontrar jogadores menos conhecidos e com valores de mercado inferior, porém que apresentassem características de jogo parecidas com os atletas de referência, para esse objetivo utilizou-se a similaridade por cosseno, onde foi possível encontrar jogadores com potenciais e com uma relação de custo-benefício mais atrativa.

Percebemos que principalmente no grupo de jogadores ofensivo a estratégia encontrou alternativas interessantes. Como destaque para esse grupo foi encontrado os jogadores mais conhecidos e com alto valor de mercado, como: Messi, Neymar. E através da similaridade cosseno, encontramos o atleta Naim Sliti. O qual apresentou um *performance* muito surpreendente, precisando de apenas 35 minutos para efetuar um passe chave, menos que a metade dos demais atletas do ataque das ligas estudas.

Entretanto, é importante ressaltar que um dos problemas nas análises de futebol, é a disparidade entre a qualidade de dados de jogadores ofensivos em relação aos defensivos, dados defensivos muitas vezes são difíceis de mensurar. Uma das principais características de um jogador defensivo é a marcação aos adversários, porém a marcação não envolve necessariamente apenas roubos de bola, interceptações. Muitas vezes tirar o espaço da linha de passe é uma grande jogada, porém esses detalhes da partida e da movimentação do jogador dado a posição da bola são difíceis de serem mensurados, os quais são disponibilizados apenas por empresas especializadas, onde são vendido aos clubes, não há disponibilidade desses dados públicos.

Há estudos que afirmam que um atleta participa cerca de apenas 5% do tempo do jogo com a bola nos pés o restante do tempo são só movimentações corporais e de posicionamento no campo. Sendo assim, as análises ficam prejudicadas pela menor

quantidade de informações sobre essas funções. No nosso estudo isso ficou evidente, pois entre os jogadores do ataque e jogadores do meio-campo foi possível identificar melhores resultados em comparação com os defensivos.

Outra característica importante que poderia ser considerada em trabalhos futuros é a questão temporal e a interação entre os atletas de uma mesma equipe e do time adversário.

Referências Bibliográficas

Ali, A. (2011). Measuring soccer skill performance: a review. *Scandinavian journal of medicine & science in sports*, 21(2):170–183.

Bangsbo, J. (1994). The physiology of soccer–with special reference to intense intermittent exercise. *Acta Physiologica Scandinavica. Supplementum*, 619:1–155.

Bar-Ilan, J. (2001). Data collection methods on the web for infometric purposes—a review and analysis. *Scientometrics*, 50(1):7–32.

Brooks, J., Kerr, M., e Guttag, J. (2016). Developing a data-driven player ranking in soccer using predictive model weights. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 49–55.

Case, K. E., Quigley, J. M., e Shiller, R. J. (2005). Comparing wealth effects: the stock market versus the housing market. *Advances in macroeconomics*, 5(1).

Dietschy, P. (2013). Making football global? fifa, europe, and the non-european football world, 1912-74. *Journal of Global History*, 8(2):279.

Freitas, L. F. (2017). A história da análise de desempenho no futebol – parte 1. http://futebolanalitico.com.br/analise-estatistica/a-historia-daanalise-de-desempenho-parte-1/. Acesso em: 20 de novembro 2020.

Giordani, P., Ferraro, M. B., e Martella, F. (2020). Introduction to clustering. In *An Introduction to Clustering with R*, pages 3–5. Springer.

Gløersen, Ø., Myklebust, H., Hallén, J., e Federolf, P. (2018). Technique analysis in elite athletes using principal component analysis. *Journal of sports sciences*, 36(2):229–237.

Gómez, M.-Á., Lago, C., Gómez, M.-T., e Furley, P. (2019). Analysis of elite soccer players' performance before and after signing a new contract. *PLoS one*, 14(1):e0211058.

Grootendorst, Maarten (2021). 9 distance measures in data science. https://towardsdatascience.com/9-distance-measures-in-data-science-918109d069fa. Acesso em: 7 de abril 2021.

Guterman, M. (2013). *O futebol explica o Brasil: Uma história da maior expressao popular do país*. Editora Contexto.

Helal, R., Soares, A. J. G., e Lovisolo, H. R. (2001). *A invenção do país do futebol: mídia, raça e idolatria*. Mauad Editora Ltda.

Hoffman, L. e Joseph, M. (2017). A multivariate statistical analysis of the nba. *Advanced Engineering Informatics*, 33:388–396.

James, G., Witten, D., Hastie, T., e Tibshirani, R. (2013). *An introduction to statistical learning*, volume 112. Springer.

Jarman, A. M. (2020). Hierarchical cluster analysis: Comparison of single linkage, complete linkage, average linkage and centroid linkage method.

Lago-Peñas, C., Gómez-Ruano, M., Megías-Navarro, D., e Pollard, R. (2016). Home advantage in football: Examining the effect of scoring first on match outcome in the five major european leagues. *International Journal of Performance Analysis in Sport*, 16(2):411–421.

Lees, A. (2002). Technique analysis in sports: A critical review. *Advanced Engineering Informatics*, 20:813–28.

Liu, G., Luo, Y., Schulte, O., e Kharrat, T. (2020). Deep soccer analytics: learning an action-value function for evaluating soccer players. *Data Mining and Knowledge Discovery*, 34(5):1531–1559.

Liu, J. C.-E. e Zhao, B. (2017). Who speaks for climate change in china? evidence from weibo. *Climatic change*, 140(3-4):413–422.

Mchale, I. G., Scarf, P. A., e Folker, D. E. (2012). On the development of a soccer player performance rating system for the english premier league. *Interfaces*, 42(4):339–351.

Perez, Rafael and Alves, Carlos Eduardo (2020). Santos é o time que mais lucrou com vendas de jogadores da base na década. https://tntsports.com.br/futebolbrasileiro/Santos-e-o-time-que-mais-lucrou-com-vendas-de-jogadores-da-base-na-decada-202002.html. Acesso em: 20 de novembro 2020.

Sally, D. e Anderson, C. (2013). *The Numbers Game: Why Everything You Know About Football is Wrong*. Penguin Books.

Sapp, R. M., Spangenburg, E. E., e Hagberg, J. M. (2018). Trends in aggressive play and refereeing among the top five european soccer leagues. *Journal of sports sciences*, 36(12):1346–1354.

Schultze, S. R. e Wellbrock, C.-M. (2018). A weighted plus/minus metric for individual soccer player performance. *Journal of Sports Analytics*, 4(2):121–131.

Transfermarkt (2021). https://www.transfermarkt.com.br/. Acesso em: 9 de março 2021.

Vyshakh K (2021). Top 10 most valuable football leagues in the world. https://www.sportskeeda.com/football/top-10-most-valuable-football-leagues-in-the-world. Acesso em: 10 de abril 2021.

Yi, Q., Groom, R., Dai, C., Liu, H., e Gómez Ruano, M. Á. (2019). Differences in technical performance of players from 'the big five' european football leagues in the uefa champions league. *Frontiers in psychology*, 10:2738.

Zhao, B. (2017). Web scraping. *Encyclopedia of big data*, pages 1–3.