Universidade Federal de São Carlos
Centro de Ciências Exatas e de tecnologia
Departamento de Estatística

Rating de jogadores: Sistema de classificação e escoragem para criação de indicadores de rendimento dos jogadores do campeonato brasileiro

Michelangelo Redondo dos Anjos

Trabalho de Conclusão de Curso

Universidade Federal de São Carlos
Centro de Ciências Exatas e de Tecnologia
Departamento de Estatística

Rating de jogadores: Sistema de classificação e escoragem para criação de indicadores de rendimento dos jogadores do campeonato brasileiro

Michelangelo Redondo dos Anjos
Orientador: Márcio Luis Lanfredi Viola

Trabalho de Conclusão de Curso a ser apresentado como parte dos requisitos para obtenção do título de Bacharel em Estatística.

São Carlos
7 de Dezembro de 2018

Michelangelo Redondo dos Anjos

Rating de jogadores: Sistema de classificação e escoragem para criação de indicadores de rendimento dos jogadores do campeonato brasileiro

Trabalho de Conclusão de Curso a ser apresentado como parte dos requisitos para obtenção do título de Bacharel em Estatística.

São Carlos, 7 de Dezembro de 2018

Banca Examinadora

- Márcio Luis Lanfredi Viola (Orientador)
- Daiane Aparecida Zuanetti
- José Carlos Fogo

Dedicatória

"Dedico esse trabalho primeiramente à Deus que em nenhum momento me deixou desistir. Ao meus pais, que estiveram ao meu lado mesmo nos momentos mais difíceis. A minha namorada Crislaine que sempre me deu força e motivação. Ao professor Márcio Luis Lanfredi Viola que me deu a oportunidade de chegar até aqui. E, por fim, aos meus amigos, que foram minha base, estrutura e teto durante esse longo caminho."

Agradecimentos

Agradeço a Deus por me dar força, saúde e inspiração para alcançar meus objetivos.

Agradeço especialmente aos meus amigos da Diretoria 012, Lucas Diniz, Arthur Rossi, Danilo Goto, Rafael Catoia, Guilherme Perim, Lucas Foresti que foram a principal motivação e inspiração para concluir o curso. Por todos os momentos que passamos juntos, pelas dificuldades, pelas conquistas, pelas alegrias, tristezas, brigas, pelas histórias, viagens, festas, estudos, noites viradas, jogos, por duvidarem que chegaríamos onde chegamos, por não acreditarem em nós. Hoje eu tenho certeza que tenho excelentes amigos, profissionais e irmãos que levarei para vida toda. Vocês foram a minha base, minha estrutura e meu teto durante esses longos 7 anos. República Janelas, Santo grau e Tunelas.

Agradeço ao meu professor, Márcio Luis Lanfredi Viola, que me deu a oportunidade de seguir em frente e de continuar no curso nos momentos mais difíceis, dando chances e acreditando em nosso potencial. Por mais professores e pessoas como você dentro das Universidades, que não se importam apenas com notas, mas sim dando valor ao lado humano, esforço e dedicação dos alunos.

Agradeço aos meus pais, pela educação e oportunidade de estudo que me deram. Quando eu não tinha mais forças, estiveram ao meu lado, apoiando minhas decisões e o que fosse melhor para mim, o que me deu um suporte e segurança de saber que poderia sempre contar com eles.

Agradeço a minha namorada Crislaine Spinazola por sempre estar ao meu lado, me incentivando, me trazendo alegria, me aguentando e apoiando em muitos momentos de muito estresse, falta de tempo, falta de atenção, mas sempre compreendendo as dificuldades e me dando suporte para alcançar meus objetivos, comemorando em momentos de alegria e apoiando em momentos de dificuldade.

Agradeço a todos os amigos do melhor grupo do *WhatsApp*, onde estamos sempre nos ajudando, desestressando, rindo, nos apoiando, comemorando as conquistas e sempre mantendo a nossa essência, ajudando uns aos outros a crescer como pessoas e profissionais.

Ao meu primo Fernando e minha vó Helena (in memorian) que não estão mais entre nós, mas continuam sendo minha força e inspiração.

Resumo

A análise de componentes principais, mais conhecida por ACP ou PCA em inglês, é um método estatístico cujo objetivo é a análise multivariada dos dados de forma conjunta, visando a redução de sua dimensionalidade sem perda, ou perda mínima de informação. É uma maneira de identificar a relação entre as características extraídas dos dados. Neste trabalho, a técnica foi explorada com o intuito de agrupar informações sobre os jogadores do campeonato brasileiro de 2017, criando um *rating* com base nos componentes que resuma e pondere as características de cada posição tornando possível o ranqueamento dos mesmos. Por fim, esse ranqueamento foi comparado com a seleção eleita por jornalistas e profissionais do futebol, na qual através dos dados, conseguimos dar um suporte para essas escolhas.

Palavras-chave: *futebol, rating, campeonato brasileiro, componentes principais, jogadores, redução de dimensão.*

Sumário

1 Introdução **1**

2 Objetivo **3**

3 Metodologia **5**

3.1 Análise de Componentes Principais 5

3.1.1 Quantidade de componentes na análise 8

3.1.2 Escolha da matriz de covariância ou de correlação 9

3.2 WebScraping 10

3.2.1 Vantagens e desvantagens 13

4 Resultados **15**

4.1 Tratamento dos dados 15

4.2 Análise Exploratória 20

4.2.1 Centroavante 20

4.2.2 Ponta 22

4.2.3 Meia ofensivo 23

4.2.4 Segundo volante 24

4.2.5 Primeiro volante 26

4.2.6 Lateral 27

4.2.7 Zagueiro 28

4.2.8 Goleiro 29

4.3 *Rating* dos jogadores 30

4.3.1 Centroavante 31

4.3.2 Ponta 33

4.3.3 Meia ofensivo 35

4.3.4 Segundo volante 38

4.3.5 Primeiro volante 40
4.3.6 Lateral 42
4.3.7 Zagueiro 45
4.3.8 Goleiro 47
4.4 Comparação dos resultados 50

5 Considerações finais **55**

Apêndice **57**

A Códigos *WebScraping* **57**

Referências Bibliográficas **59**

Lista de Tabelas

4.1 Variáveis utilizadas para cada posição 17
4.2 Variáveis utilizadas para cada posição 18
4.3 Métricas criadas para análise 19
4.4 Análise descritiva das variáveis de centroavante 21
4.5 Análise descritiva das variáveis de pontas 22
4.6 Análise descritiva das variáveis de meia ofensivo 23
4.7 Análise descritiva das variáveis de segundo volante 25
4.8 Análise descritiva das variáveis de primeiro volante 26
4.9 Análise descritiva das variáveis de laterais 27
4.10 Análise descritiva das variáveis de zagueiros 29
4.11 Análise descritiva das variáveis de goleiros 30
4.12 Autovalores e % de variância explicada: centroavante 31
4.13 Autovetores e pesos dos componentes: centroavante 32
4.14 *Rating* e informações dos centroavantes 33
4.15 Autovalores e % de variância explicada: ponta 33
4.16 Autovetores e pesos dos componentes: ponta 34
4.17 *Rating* e informações dos pontas 35
4.18 Autovalores e % de variância explicada: meia ofensivo 36
4.19 Autovetores e pesos dos componentes: meia ofensivo 37
4.20 *Rating* e informações dos meias ofensivos 37
4.21 Autovalores e % de variância explicada: segundo volante 38
4.22 Autovetores e pesos dos componentes: segundo volante 39
4.23 *Rating* e informações dos segundos volantes 40
4.24 Autovalores e % de variância explicada: primeiro volante 40
4.25 Autovetores e pesos dos componentes: primeiro volante 41

4.26 *Rating* e informações dos primeiros volantes 42
4.27 Autovalores e % de variância explicada: lateral 43
4.28 Autovetores e pesos dos componentes: lateral 44
4.29 *Rating* e informações dos laterais 45
4.30 Autovalores e % de variância explicada: zagueiro 45
4.31 Autovetores e pesos dos componentes: zagueiro 46
4.32 *Rating* e informações dos zagueiros 47
4.33 Autovalores e % de variância explicada: goleiro 48
4.34 Autovetores e pesos dos componentes: goleiro 49
4.35 *Rating* e informações dos goleiros 49
4.36 Comparação dos segundos volantes 51
4.37 Comparação dos primeiros volantes 51
4.38 Comparação dos laterais 52
4.39 Comparação dos zagueiros 52

Lista de Figuras

3.1 Estruta dos dados na página do G1 11
3.2 Ferramenta de desenvolvimento dos navegadores 12
3.3 Estrutura JSON 12

4.1 Correlação entre as variáveis de centroavante. 21
4.2 Correlação entre as variáveis de pontas. 22
4.3 Correlação entre as variáveis de meia ofensivo. 24
4.4 Correlação entre as variáveis de segundo volante. 25
4.5 Correlação entre as variáveis de primeiro volante. 26
4.6 Correlação entre as variáveis de laterais. 28
4.7 Correlação entre as variáveis de zagueiro. 29
4.8 Correlação entre as variáveis de goleiro. 30
4.9 Gráfico de cotovelo para os componente principais: centroavante 31
4.10 Biplot dos centroavantes com relação aos dois componentes analisados . . . 32
4.11 Gráfico de cotovelo para os componente principais: ponta 34
4.12 Biplot dos pontas com relação aos dois componentes analisados 35
4.13 Gráfico de cotovelo para os componente principais: meia ofensivo 36
4.14 Biplot dos meias ofensivos com relação aos dos dois componentes analisados 37
4.15 Gráfico de cotovelo para os componente principais: segundo volante 38
4.16 Biplot dos segundos volantes com relação aos dois componentes analisados 39
4.17 Gráfico de cotovelo para os componente principais: primeiro volante 41
4.18 Biplot dos primeiros volantes com relação aos dois componentes analisados 42
4.19 Gráfico de cotovelo para os componente principais: lateral 43
4.20 Biplot dos laterais com relação aos dois componentes analisados 44
4.21 Gráfico de cotovelo para os componente principais: zagueiro 46
4.22 Biplot dos zagueiros com relação aos dois componentes analisados 47

4.23 Gráfico de cotovelo para os componente principais: goleiro 48
4.24 Biplot dos goleiros com relação aos dois componentes analisados 49
4.25 Seleção sugerida 4-3-3. 53

Capítulo 1

Introdução

A Estatística é uma ciência multidisciplinar que visa o estudo da variabilidade e incerteza com o objetivo de auxiliar em tomadas de decisões por meio de metodologias que permitam obter conclusões, transformando dados em informações. Assim como em todas as áreas de conhecimento, a variabilidade e incerteza também estão presentes no esporte [19].

Com o foco esportivo, a Estatística vem sendo usada em diversas modalidades como o beisebol e o futebol americano, principalmente nos Estados Unidos e em países europeus. Um caso de sucesso nessa área é o do dirigente Billy Beane, que conseguiu levar o *Oakland Athletics*, um dos times de menor orçamento da liga profissional de beisebol, às fases finais do campeonato utilizando estatísticas de jogadores visando encontrar atletas baratos e fundamentais para o time.

Para ter sucesso em seus ideais, Beane se baseou em análises feitas por Bill James, formado em Literatura Inglesa e Economia que, em 1977, publicou um livro contendo análises estatísticas de jogadores, chamado de *The Bill James baseball abstract*. Além das análises, o livro introduziu as estatísticas e fórmulas para medir o desempenho dos jogadores de beisebol dentro de campo. Em 2012 inclusive, o caso veio a se tornar um filme chamado *Moneyball* – O Homem que mudou o jogo [5].

Nos dias atuais a estatística vem buscando se adentrar no futebol a procura de tendências que possam fazer a diferença dentro de campo e, dessa forma, vem cada vez mais ganhando seu espaço atraindo empresas, clubes e universidades.

Os clubes da Europa usam, com muito mais frequência e ferramentas, a Estatística no futebol. No Brasil temos empresas como a *Footstats* que já possui software para análise de desempenho e realiza consultorias para diversos times do Brasil, como por exemplo o

Red Bull Brasil.

O caso de sucesso mais recente do uso da estatística no futebol se deu através da empresa equatoriana *Kin Analytics* [1], que antes de se adentrar no meio futebolístico, trabalhava no ramo de serviços bancários como concessão de crédito, análise de risco e análise de fraude. Ao perceberem que os algoritmos desenvolvidos também poderiam ser aplicados em esportes, aproveitaram as oportunidades para expandir seus produtos, inclusive, foram responsáveis pela montagem do elenco do Grêmio, campeão da Copa do Brasil em 2016 e da Libertadores em 2017. No começo do ano de 2018 a *Kin Analytics* foi contratada pela equipe do Palmeiras [23] que foi campeão do campeonato brasileiro de 2018.

Ainda em casos de sucesso no futebol brasileiro temos o de Alex Bourgeois, um matemático com carreira no mercado financeiro que decidiu trabalhar com o futebol criando um algoritmo que conseguiu prever para os anos de 2012, 2013 e 2014 os quatro últimos e os quatro primeiros times do campeonato brasileiro, não necessariamente em ordem correta. Além disso, Bourgeois diz ser possível prever quantos pontos na tabela cada jogador acrescentaria a determinado time ao final da temporada [13].

Uma metodologia que vem ganhando espaço nesse mercado são os *ratings*, também rotulados como *marks*, índices ou classificações dependendo do local. Os *ratings* já vêm sendo implantados a algum tempo no futebol virtual, como nas franquias Fifa e *Pro Evolution Soccer*. Esses indicadores de desempenho precisos são baseados nas estatísticas dos jogadores. Nos jogos virtuais, ao final de cada partida temos essas notas atribuídas a cada jogador a fim de apresentar ao público quais foram os que mais se destacaram.

Hoje em dia temos alguns *ratings* muito respeitados no mundo do futebol que além de atribuir uma pontuação aos jogadores ao final das partidas também oferece uma pontuação ao final da temporada e identifica quais jogadores mais se destacaram durante todo ano por posição. No Brasil temos referências em relação a esses índices como o índice *Footstats* e de maneira mundial temos o *rating WhoScored*.

Ao final de cada temporada do campeonato brasileiro, temos uma seleção eleita por um colégio eleitoral formado por jogadores, técnicos, jornalistas e ex-jogadores que realizam a votação dos atletas que atuaram nos clubes do país em seu respectivo ano [6].

Capítulo 2

Objetivo

O trabalho tem como objetivo identificar as variáveis que estão mais correlacionadas com cada posição em campo existente, tais como, atacantes, meias, volantes, zagueiros, laterais e goleiros a fim criar um sistema de ranqueamento dos jogadores do campeonato brasileiro baseado em suas diversas variáveis coletadas durante uma temporada, atribuindo uma nota variando de 0 a 10 e gerando um *rating* para cada um. Através de um ranking será compararado os melhores jogadores de cada posição indicados pelo *rating* com a seleção do campeonato brasileiro eleita por jornalistas e profissionais do futebol, com o intuito de auxiliar de maneira analítica a escolha dessa seleção.

Capítulo 3

Metodologia

Neste capítulo será apresentada a técnica estatística utilizada para a análise do conjunto de dados do campeonato brasileiro relativo à temporada 2017.

3.1 Análise de Componentes Principais

A técnica utilizada nesse estudo para análise dos dados é chamada de análise de componentes principais (ACP), uma metodologia de análise multivariada que foi idealizada por Karl Pearson [30] e está documentada por Hotelling [27].

O objetivo principal da técnica é explicar a estrutura de variância e covariância de um vetor aleatório de p-variáveis quantitativas através da construção de combinações lineares das variáveis originais. Considere o vetor aleatório $\mathbf{X} = (X_1, ..., X_p)^t$ e o vetor de constantes $\mathbf{a_i} = (a_{i1}, ..., a_{ip})^t$. Temos então combinações lineares da forma:

$$Y_i = \mathbf{a_i^t X} = a_{i1}X_1 + a_{i2}X_2 + ... + a_{ip}X_p \quad para \quad i = 1, 2, ..., p.$$

Essas combinações lineares são chamadas de componentes principais, não são correlacionadas entre si e explicam a multicolinearidade das variáveis originais. Os componentes são obtidos de forma que os primeiros descrevem o máximo de informação possível através da variabilidade [28].

Através da ACP conseguimos reduzir a dimensão do conjunto de dados facilitando a análise e interpretação, ou seja, resumir a informação contida nas p variáveis originais através de k, $k < p$, componentes que contém a maior parte da variabilidade do conjunto original. Dessa maneira, se há p variáveis é possível obter p componentes principais.

Por outro lado, a qualidade da análise depende do número de componentes escolhidas. Uma técnica para a escolha desse número utiliza a proporção de variância total explicada pelos componentes de maneira que os escolhidos expliquem o comportamento das variáveis e os que fiquem de fora sejam apenas ruído.

Os componentes principais dependem somente da estrutura da matriz de covariâncias $\mathbf{\Sigma}$ ou da matriz de correlações $\mathbf{R}$ das variáveis originais. Dessa forma, não é necessária nenhuma suposição inicial de normalidade, apenas que as variáveis originais tenham certa correlação significativa entre si [26].

Algumas situações nas quais a ACP perde poder são:

- O conjunto de dados possui mais variáveis do que unidades amostrais;
- As variáveis originais são pouco correlacionadas ou;
- A matriz de correlação $\mathbf{R}$ é igual a matriz identidade $\mathbf{I}$. Nesse caso, os componentes principais são as próprias variáveis originais [32, 26]. Nesse caso não se aplica a ACP.

Para melhor entendimento da metodologia vamos considerar o vetor de variáveis $(X_1, X_2, ..., X_p)^t$ observado em N unidades experimentais e essas variáveis não são independentes. Dessa forma, temos uma matriz N x p de dados, que denotaremos por $\mathbf{X}$ e definida como

$$\mathbf{X} = \begin{bmatrix} X_{11} & \dots & X_{1p} \\ \vdots & \ddots & \vdots \\ X_{n1} & \dots & X_{np} \end{bmatrix}.$$

De forma geral, considere o conjunto $\mathbf{X}$, cujo vetor de médias é $\mu = (\mu_1, \mu_2, ..., \mu_p)^t$ e o vetor de variâncias é $(\sigma_{11}, \sigma_{22}, ..., \sigma_{pp})^t$, respectivamente. Como esse conjunto de variáveis não é independende, possuímos covariâncias entre a i-ésima e k-ésima variável, denotadas por σ_{ik}, para $i \neq k = 1, 2, ..., p$.

Temos então uma estrutura de variâncias e covariâncias expressa pela matriz simétrica:

$$\mathbf{\Sigma} = \begin{bmatrix} \sigma_{11} & \dots & \sigma_{p1} \\ \vdots & \ddots & \vdots \\ \sigma_{p1} & \dots & \sigma_{pp} \end{bmatrix}.$$

Podemos obter a matriz de correlação através da matriz de covariância, sabendo que:

$$\rho_{ik} = \frac{\sigma_{ik}}{\sqrt{\sigma_{ii}\sigma_{kk}}} \quad i,k = 1,2,...,p \; com \; i \neq k.$$

Dessa maneira, teremos a matriz de correlação:

$$\mathbf{R} = \begin{bmatrix} 1 & \dots & \rho_{p1} \\ \vdots & \ddots & \vdots \\ \rho_{p1} & \dots & 1 \end{bmatrix}.$$

Devido a diferença de escalas entre as variáveis, muitas vezes encontradas na prática, é comum padronizá-las com o objetivo de eliminar as diferenças entre elas. A padronização mais comum é chamada de z-escore, onde subtraímos cada valor da variável pela sua média e dividimos pelo seu desvio padrão como:

$$Z_{ij} = \frac{X_{ij} - \mu_i}{\sqrt{\sigma_{ii}}}, \quad i = 1,2,...,p \quad e \quad j = 1,2,...,N. \tag{3.1}$$

Em notação matricial, temos:

$$\mathbf{Z} = \mathbf{V}^{-\frac{1}{2}}(\mathbf{X} - \mu),$$

em que $\mathbf{V}^{\frac{1}{2}}$ é a matriz diagonal de desvios padrões. Dessa forma:

$$Cov(\mathbf{Z}) = \mathbf{V}^{-\frac{1}{2}} \, \mathbf{\Sigma} \, \mathbf{V}^{-\frac{1}{2}} = \mathbf{R}.$$

Mostra-se, assim, que a matriz de correlação $\mathbf{R}$ é igual a matriz de covariância $\mathbf{\Sigma}$ das variáveis padronizadas utilizando a padronização da Equação (3.1). Da mesma forma que obtemos a matriz $\mathbf{R}$ através de $\mathbf{\Sigma}$, podemos também encontrar a matriz $\mathbf{\Sigma}$ através de $\mathbf{R}$:

$$\mathbf{\Sigma} = \mathbf{V}^{\frac{1}{2}}\mathbf{R}\mathbf{V}^{\frac{1}{2}}.$$

Com isso os componentes principais podem ser obtidos tanto da matriz de correlação $\mathbf{R}$ quanto da matriz de covariâncias $\mathbf{\Sigma}$, sendo que todos os resultados vistos posteriormente se aplicam a matriz de correlação.

Os componentes principais são encontrados a partir do teorema da decomposição espectral que nos mostra que a melhor representação de $\mathbf{R}$ é dada pela direção dos autove-

tores associados a cada um dos autovalores da matriz $\mathbf{R}$ cujo comprimento é dado pelo respectivo autovalor.

Seja $\mathbf{\Sigma}$ uma matriz simétrica positiva definida de ordem p. A mesma pode ser reescrita a partir dos seus autovalores e autovetores da seguinte forma:

$$\mathbf{\Sigma} = \mathbf{P}\Delta\mathbf{P}^t,$$

em que $\mathbf{P}$ é a matriz composta pelos autovetores normalizados, $\mathbf{e}_i^t\mathbf{e}_i = 1$, $\mathbf{e}_i^t\mathbf{e}_j = 0$, de $\mathbf{\Sigma}$ em suas colunas, Δ é a matriz diagonal de autovalores de $\mathbf{\Sigma}$. A partir disso, temos que:

$$tr(\mathbf{\Sigma}) = tr(\mathbf{P}\Delta\mathbf{P}^t) = tr(\Delta\mathbf{P}^t\mathbf{P}) = tr(\Delta I) = tr(\Delta) = \sum_{i=1}^{p} \lambda_i.$$

Portanto, a variabilidade total contida nas variáveis originais é igual a variabilidade total contida nos componentes principais [28], ou seja,

$$tr(\mathbf{\Sigma}) = \sum_{i=1}^{p} \sigma_{ii} = \sum_{i=1}^{p} Var(X_i) = \sum_{i=1}^{p} \lambda_i = \sum_{i=1}^{p} Var(Y_i).$$

Com base na matriz de covariância obtemos então os pares de autovalores e autovetores $(\lambda_1, \mathbf{e}_1), (\lambda_2, \mathbf{e}_2), ..., (\lambda_p, \mathbf{e}_P)$, em que $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p$ onde o i-ésimo componente principal da matriz $\mathbf{\Sigma}$ é definido por [28]:

$$\begin{cases} Y_i = \mathbf{e_1^t}\mathbf{X} = e_{11}X_1 + e_{12}X_2 + ... + e_{1p}Xp \\ \vdots \\ Y_p = \mathbf{e_p^t}\mathbf{X} = e_{p1}X_1 + e_{p2}X_2 + ... + e_{pp}Xp. \end{cases}$$

Essas novas variaveis Y_i são consideradas variáveis latentes, ou seja, não são mensuradas a partir do experimento ou levantamento amostral.

O motivo pelo qual a variância da primeira componente é sempre maior que a da segunda e assim por diante se dá pelo lema da maximização que está detalhado no livro *Applied multivariate statistical analysis* [28].

3.1.1 Quantidade de componentes na análise

A quantidade de componentes a ser escolhidos durante a análise é uma questão sempre em discussão. Porém, não existe uma resposta correta apenas alguns aspectos que devem

ser considerados:

- Um componente associado a um autovalor próximo de zero será pouco importante;
- A magnitude dos autovalores que indica a variância dos componentes principais.

Com isso, buscamos manter apenas os componentes que expliquem a maior parte da variação no conjunto de dados. A contribuição do i-ésimo componente principal Y_i é dada por:

$$\frac{\lambda_i}{\lambda_1 + \lambda_2 + ... + \lambda_p} \quad , i = 1, 2, ..., p.$$

Através dessa proporção temos o quanto cada componente explica da variabilidade total e dessa forma conseguimos determinar a quantidade de componentes a ser considerada durante a análise.

Outro método muito utilizado pelos pesquisadores é o critério de Kaiser. Ele seleciona apenas os componentes cujos autovalores sejam maiores que 1 ($\lambda_i > 1$), o que significa uma contribuição maior que a esperada de cada componente.

Temos ainda o critério adotado por Jolliffe que descarta os componentes cuja variância seja inferior a 0.7 ($\lambda_i < 0.7$).

Por fim, um dos métodos mais difundidos é o método do "cotovelo" que consiste da análise de um gráfico (Scree Plot) construído com os valores de $\lambda_i \, vs \, i$. Nesse caso, o número de componentes é escolhido a partir do momento em que os autovalores restantes são relativamente pequenos e tem tamanhos aproximadamente iguais [28].

Os métodos apresentados para a escolha dos componentes não são excludentes, ou seja, podemos utilizar todos eles para definir a quantidade de componentes que serão utilizados na análise. Para ilustrar esse tipo de gráfico, veja, por exemplo, a Figura 4.9.

3.1.2 Escolha da matriz de covariância ou de correlação

A matriz de covariâncias reflete a variabilidade dos dados e a variabilidade de uma variável pode ser muito superior as demais, consequência da sua unidade de medida. Essa presença de variáveis com diferentes escalas tem impacto significativo no cálculo da variância total e consequentemente no dos componentes principais, tendo em vista que as variáveis com maior variabilidade tendem a dominar os primeiros componentes.

Na matriz de correlação seus valores independem da unidade de medida e os valores fora da diagonal reflete a correlação entre as variáveis.

Na literatura a recomendação para construir os componentes é que a matriz de covariância seja utilizada apenas quando as variáveis possuem a mesma escala de medida e que as variâncias entre elas estejam muito próximas. Em geral, isso é muito difícil de ocorrer nos levando na maioria das vezes a utilizar a matriz de correlação.

3.2 WebScraping

Para a obtenção dos dados foi utilizada a técnica denominada *Web Scraping* que é utilizada para extração de dados disponíveis em *websites*. Através dela é possível coletar grande quantidade de dados de maneira automática, implementando *boots* que são robôs programados para fazerem a captura de informações de forma direta usando protocolos HTTP ou por meio do navegador.

Atualmente, muitas empresas têm utilizado essa ferramenta devido ao fato de que cada vez é maior o número de *sites* e instituições que disponibilizam dados relevantes em suas páginas. Essa técnica necessita de alguns conhecimentos a priori de linguagens de programação na web como HTML, PHP, Javascript, entre outras, pois é necessário inspecionar como o site foi desenvolvido e a maneira com que os dados são alimentados nele, o que influencia em quais procedimentos executar, tais como:

- Extração direta por tabelas em formato HTML;
- Extração por meio de *requests* por PHP;
- Extração através de API's alimentadas via Javascript.

As APIs são um tipo de "ponte" que conectam aplicações podendo ser utilizadas para os mais variados tipos de negócio. Contextualizando com a aplicação desse estudo, as API's conectam o banco de dados a sites, aplicativos, ou qualquer meio que forneça um suporte para a visualização dos mesmos. Em geral, temos uma sequência de passos que podemos seguir para essa coleta.

O primeiro passo é identificar o tipo de informação que iremos coletar (tabelas, textos, entre outros) e entender qual a estrutura das páginas onde elas estão armazenadas para depois traçar uma estratégia de extração.

Para exemplificar o uso da técnica, vamos considerar o *site* do G1 [25] com informações sobre a denúncia contra o atual presidente Michel Temer. A Figura 3.1 mostra como os dados estão disponibilizados no site e concluímos que estão em um formato de tabela.

Figura 3.1: Estruta dos dados na página do G1

O segundo passo é descobrir de onde vêm os dados que queremos extrair. Para isso, podemos utilizar ferramentas de desenvolvedor do nosso navegador para encontrar essa fonte de dados. As três principais e mais utilizadas são: HTML, PHP e JSON. Para essa identificação podemos analisar a aba de *networking* do navegador e entender as chamadas HTTP que estão sendo carregadas no site, ou em alguns casos, conseguimos essa informação simplesmente analisando o código fonte da página.

No caso da página do G1 ilustrado na Figura 3.1 não encontramos a tabela no seu código fonte o que nos levar a nos aprofundar um pouco mais na estrutura do site. Abrindo a ferramenta de desenvolvimento dos navegadores (no Google Chrome [9] o atalho é dado pelas teclas CTRL+SHIFT+I) na aba *network* verificamos os dados trafegados pela rede ao recarregar a página utilizando a tecla F5 e encontramos o link do formato JSON que basicamente é um formato de troca de informações/dados entre sistemas que alimenta os dados na página como mostra a Figura 3.3.

Ao acessar o link encontrado na aba *network* conseguimos analisar a estrutura desse armazenamento exbibida na Figura 3.3

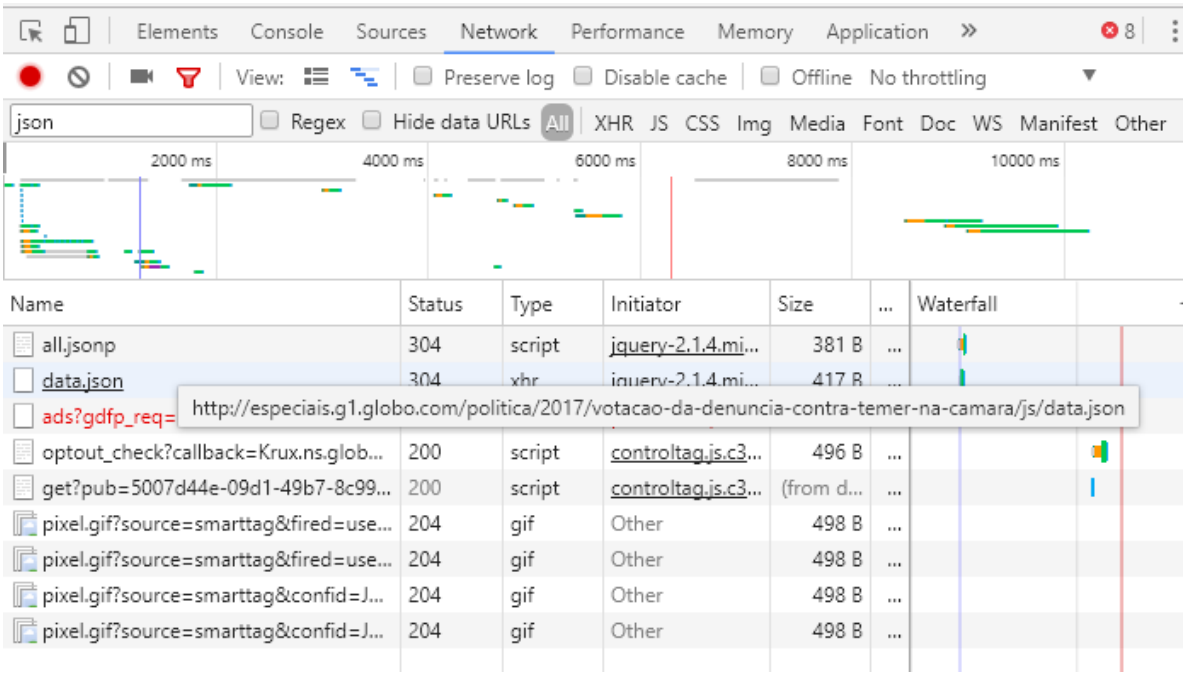


Figura 3.2: Ferramenta de desenvolvimento dos navegadores

Figura 3.3: Estrutura JSON

O terceiro passo, que podemos ou não executar, são iterações no site caso os dados estejam dispersos em diversas páginas. Buscamos encontrar um padrão entre elas e começar a iteração através de um *loop* programado até que todas tenham obtido algum retorno. Caso as páginas não possuam um padrão, podemos ir mais a fundo e fazer o acesso através do navegador, fazendo assim com que o computador acesse as páginas automaticamente e, em seguida, realize as extrações.

O *R* possui alguns pacotes que nos ajudam nessas tarefas, são eles: *rvest* [17], *httr* [24], *xml2* [18], *jsonlite* [22], *rjson* [10] e *RSelenium* [20]. Além disso, diversos cursos em plataformas que abordam esse tema e alguns livros sobre o assunto estão surgindo. No Apêndice A apresentamos o código usado para a obtenção dos dados do *site* do G1.

3.2.1 Vantagens e desvantagens

Apesar das facilidades e benefícios que essas ferramentas nos trazem, principalmente nos dias atuais em que os dados estão sendo considerados como o "novo petróleo"[3], essas técnicas vêm sendo utilizadas também para práticas maliciosas, utilizadas para roubar conteúdos protegidos, cometer fraudes ou saturar servidores tornando-os indisponíveis para os clientes.

Algumas soluções estão sendo tomadas por parte da Tecnologia da Informação (TI) a fim de evitar essa captura de informações de forma a construir sites com capacidade de detectar e desconsiderar os *boots* nas páginas. Portanto, muitas vezes a técnica de *scraper* acaba sendo confundida como atividades *hackers* o que é um assunto de muita discussão por, hoje em dia, não ter leis muito específicas relacionadas a isso.

Vale então a questão de ética e respeito aos termos e condições de uso impostos pelos sites e, principalmente, respeitar a privacidade no sentido de que se um usuário comum não tenha acesso as informações, *boots* também não deveriam obtê-las. [14, 16].

Temos alguns casos a favor e contra a prática de *WebScraping* como do engenheiro de software Pete Warden que construiu um rastreador da Web para coletar dados do Facebook. Ele coletou dados de aproximadamente 200 milhões de usuários tais como nomes, informações de localização, amigos e interesses, prática essa que infringe os termos de uso do site e que trouxe notificações e avisos por meio da equipe do Facebook para que os ataques parassem [29].

Um contraexemplo ao de Pete Warden é o caso do Linkedln que teve problema por impedir o scraping em dados que são públicos na plataforma. A justiça considerou errado

a impedição e ordenou que o *scraping* fosse liberado pelo Linkedln em até 48 horas [11].

Existem três critérios que precisam ser atendidos, ou pelo menos alguns deles, para que um *scraping* seja considerado uma invasão. São eles:

- Falta de consentimento: A maioria dos servidores da web estão abertos a todos, sendo assim, dão consentimentos ao *WebScraping*. Porém, conforme já citado anteriormente, os termos de serviços de muitos websites proíbem especificamente o uso de scrapers, ou, em alguns casos, notificações e avisos também revogam esse consentimento;

- Dano: Servidores possuem custos, sendo assim, caso os sites sejam derrubados por scrapers e sua capacidade de atender os usuários seja limitada, o proprietário pode ser prejudicado;

- Intencionalidade: A partir do momento que você escreve um código você é responsável por ele e sabe o que está fazendo, sendo assim, todo ato passa a ser intencional [29].

Os termos de uso do *WhoScored*, site no qual foi realizado a captação dos dados, estão disponibilizados no próximo parágrafo. Durante esse estudo, tentamos estabelecer contato com o site para pedir permissão de acesso, porém não tivemos resposta. Mas devido seus termos de uso e o objetivo do estudo que não tem fins lucrativos ou de reprodução, a coleta dos dados não infringe os mesmos.

Termos de uso do *WhoScored*

"O WhoScored.com tem como objetivo fornecer conteúdo o mais preciso e rápido possível, no entanto, não somos responsáveis por conteúdo impreciso e atrasado. A cópia, download, reprodução, republicação, enquadramento, transmissão de conteúdo WhoScored.com, incluindo, mas não se limitando a todas as estatísticas, dados, produtos, tabelas, gráficos e outras informações é proibida sem uma licença oficial. Se você deseja obter permissão para usar o conteúdo do WhoScored.com, entre em contato conosco aqui." Data sources - Opta Sports, eNetPulse Getty Images. Copyright © 2018 WhoScored.com.

Capítulo 4

Resultados

Neste capítulo, o conjunto de dados e o tratamento realizado a ele são detalhados. Em seguida, algumas análises descritivas são feitas afim de se conhecer melhor as variáveis em estudo. Por fim, a escolha e construção dos indicadores utilizados na análise e criação dos ratings dos atletas são apresentadas.

4.1 Tratamento dos dados

Os dados foram obtidos através da técnica de *Webscrapping* realizada no site *Whoscored* [6] para a obtenção de algumas váriaveis explicativas acerca dos jogadores de interesse. Ao todo 40 variáveis foram retiradas dessa fonte de dados com informações sobre o desempenho dos jogadores ao longo da temporada, sumarizadas por rodada, por exemplo, porcentagem de passes certos por jogo e dribles certos por jogo. Algumas com os totais também foram capturadas como o número de gols na temporada, total de assistências, entre outras.

Para agregar mais conteúdo as análises, também foram obtidos dados da API do *fantasy* CartolaFC, principalmente pelo fato de termos poucas informações sobre goleiros. Nessa fonte de dados 18 varíaveis foram coletadas, todas com valores totais, como por exemplo, número de defesas difícieis, número de gols sofridos, entre outras.

Dentre as 58 variáveis coletadas foram utilizadas apenas as definidas como relevantes, de forma subjetiva e conhecimento a priori do meio futebolístico, para o nosso objetivo. Para cada posição as variáveis podem ou não ser as mesmas, dependendo das suas características.

A população de estudo continha incialmente 730 jogadores inscritos no campeonato

brasileiro de 2017, porém, devido ao fato de que os times utilizam apenas 11 jogadores titulares, podem fazer apenas 3 substituições por jogo e que 7 jogadores ficam no banco de reservas, quase metade dos inscritos por time não chegam nem a ser relacionados para partida. Outra questão é que o número de inscritos no campeonato brasileiro não é limitado, ocasionando times com muitos jogadores com pouca participação nos jogos.

Com o intuito de evitar que esses jogadores com pouca informação participassem das análises, o que acabaria atrapalhando devido a discrepância e falta de dados, um filtro foi realizado selecionando apenas aqueles com um total de jogos maior que o número médio de partidas dos jogadores no campeonato (14 partidas). Dessa forma a população final de estudo foi composta por 326 jogadores, dentre eles 76 atacantes, 24 goleiros, 56 laterais, 118 meias e 52 zagueiros.

Os dados foram coletados através do site *WhoScored* e da API do *fantasy* Cartola FC. Devido à divergência entre as fontes de informações, números diferentes nos indicadores como o total de assistências, a média entre as duas variáveis foram calculadas. Após a coleta, os dados foram processados e análisados através da linguagem de programação R [31] e com o auxílio da interface de desevolvimento RStudio [15]. Para a criação do *rating* geral por posição foi utilizada a análise de componentes principais e quando necessária mais de um componente para explicar a maior parte da variabilidade dos dados, a média ponderada entre elas foi calculada.

De maneira a deixar esse *rating* mais interpretável, após a obtenção do escore para cada jogador, o mesmo foi padronizado pela função acumulada da distribuição uniforme, que aplicado ao nosso objetivo é calculado da seguinte forma:

$$R_i = \frac{J_i - Min(J)}{Max(J) - Min(J)}, \quad i = 1, 2, ..., N \tag{4.1}$$

em que:

- R_i : Valor do *rating* padronizado do jogador i;
- J_i : Valor do jogador gerado pelos componentes (escore do jogador i);
- $Max(J)$: Máximo valor dos jogadores gerado pelos componentes (escore máximo);
- $Min(J)$: Mínimo valor dos jogadores gerado pelos componentes (escore mínimo).

Essa padronização força os valores a ficarem entre 0 e 1 e depois os mesmos são multiplicado por 10, sendo 0 a menor pontuação possível e 10 a máxima.

Tinhamos inicialmente 5 posições (atacante, meio campo, laterais, zagueiros, goleiros), porém, entre elas os estilos de jogo divergiam. Dessa forma, foi realizada uma maior segmentação afim de otimizar nossas análises de maneira que os grupos finais se dividiram em: Atacantes (centroavante e pontas), meio campo (meia ofensivo, primeiro volante e segundo volante), laterais, zagueiros e goleiros.

As variáveis utilizadas em cada posição são exibidas na Tabela 4.1 e 4.2.

Tabela 4.1: Variáveis utilizadas para cada posição

Posição	Variável/Descrição
Centroavante	DAV = duelos aéreos vencidos
	LET = letalidade
	GPM = gols por minutos jogados
	FG = finalizações no gol
Pontas	TASS = total de assistências
	POG = passe para oportunidade de gol
	TXD = taxa de dribles
	LET = letalidade
	GPM = gols por minutos jogados
	ACC = acurácia no chute
Meia ofensivo	TASS = total de assistências
	POG = passe para oportunidade de gol
	DRC = dribles certos
	CRP = cruzamentos precisos
	GPJ = gols por jogo
Segundo volante	ETRB = efetividade nas roubadas de bola
	POG = passe para oportunidade de gol
	EDAV = efetividade nos duelos aéreos vencidos
	TXD = taxa de dribles
	EINT = efetividade nas interceptações
	EDES = efetividade nos desarmes
	PPC = porcentagem de passes certos

Tabela 4.2: Variáveis utilizadas para cada posição

Posição	Variável/Descrição
Primeiro volante	ETRB = efetividade nas roubadas de bola
	EDAV = efetividade nos duelos aéreos vencidos
	EINT = efetividade nas interceptações
	EDES = efetividade nos desarmes
	EREB = efetividade nas rebatidas
	EBLO = efetividade nos bloqueios
Laterais	ETRB = efetividade nas roubadas de bola
	TSG = jogos sem sofrer gols
	TASS = total de assistências
	POG = passe para oportunidade de gol
	EINT = efetividade na interceptação
	EDES = efetividade nos desarmes
	CRP = cruzamentos precisos
Zagueiros	ETRB = efetividade nas roubadas de bola
	TSG = jogos sem sofrer gol
	EINT = efetividade nas interceptações
	EDES = efetividade nos desarmes
	EREB = efetividade nas rebatidas
	EDAV = efetividade nos duelos aéreos vencidos
	EBLO = efetividade no bloqueio
	PPC = porcentagem de passes certos
Goleiros	TDP = totais de defesas de pênaltis
	TSG = jogos sem sofrer gols
	PDEF = poder defensivo

Algumas variáveis foram utilizadas de maneira natural e outras foram criadas através de combinações buscando uma melhor interpretação e ponderamento. Dentre essas, temos métricas bastante utilizadas no meio futebolístico tais como a média de gols por jogo e média de gols por minuto (normalmente chamada de minutos para marcar). Além destas, outras também foram utilizadas sendo julgadas relevantes para o objetivo do estudo. A fórmula como elas foram cálculadas é exibida na tabela 4.3

Tabela 4.3: Métricas criadas para análise

Variáveis	Fórmula
Letalidade	Gols/(Finalizações no gol + Finalizações para fora)
Gols por minutos jogados	Gols/Minutos jogados
Gols por jogo	Gols/Jogos
Taxa de dribles	Dribles certos/(Dribles errados + Dribles certos)
Finalizações no gol	Finalizações na trave + Finalizações defendidas
Acurácia no chute	Finalizações no gol / Finalizações
Poder defensivo	Defesas difíceis / Gols sofridos
Efetividade nas interceptações	Interceptações / Gols sofridos
Efetividade nos desarmes	Desarmes / Gols sofridos
Efetividade nos duelos aéreos vencidos	Duelos aéreos vencidos / Gols sofridos
Efetividade nas rebatidas	Rebatidas / Gols sofridos
Efetividade no bloqueio do chute	Bloqueios de chute / Gols sofridos
Efetividade na roubada de bola	Roubadas de bola / Gols sofridos

A seguir será apresentado a descrição das variáveis de modo a explicar o funcionamento de cada uma no meio futebolístico e na aplicação nesse estudo. Dessa forma, é considerado:

- Assistência: Passe que gera o gol de um companheiro. Foi utilizado o valor total em todo campeonato;
- Bloqueio de chutes: Prevenir que um oponente dê um chute ao gol;
- Rebatidas: Em lances em que a bola é lançada na área quando o zagueiro afasta a mesma temporariamente;
- Cruzamentos precisos: Lançamentos na área na qual resultou em toque de um companheiro de time;
- Dribles: Certos quando é executado sem perda de bola. Errado quando o adversário evita a passagem;
- Interceptação: Quando um jogador intercepta o passe do adversário para um companheiro evitando que o passe seja executado recuperando assim a posse de bola;

- Passe para oportunidade de gol: O passe final que leva a um chute no gol de um jogador da equipe;
- Desarme: Retirada da bola do oponente fazendo o desarme saindo ou não com a posse de bola.

As variáveis criadas como efetividade descritas na Tabela 4.3 se deve ao fato de que muitas vezes o jogador é muito acionado e não temos a quantidade de bolas que chegaram até ele, apenas as que ele efetuou a ação. Por exemplo no desarme, não temos quantos desarmes ele deixou de efetuar, apenas quanto foi executado. Muitas vezes o aumento desse número não necessariamente significa ser melhor.

Dessa forma, essas variáveis foram divididas pelo número total de gols sofridos na temporada a fim de tentar explicar o quanto aquele desarme está evitando que ocorra gol. Com a mesma ideia foi criada a varíavel "poder defensivo"para os goleiros, pois ele pode realizar muitas defesas difíceis, porém pode ser devido ao fato dele estar sendo muito acionado por causa de uma fraca marcação dos zagueiros. Além disso ele pode estar sofrendo muitos gols, ou seja, essas defesas não estão sendo efetivas.

4.2 Análise Exploratória

Será apresentado nessa seção uma análise descritiva dos dados a fim de conhecê-los e explorá-los. Os dados foram sumarizados por posição de jogadores. As medidas utilizadas são de tendência central como a média e medidas de variabilidade (desvio padrão, valor máximo e mínimo). Além disso, análises bivariadas foram realizadas através da correlação de Pearson afim de verificar o grau de correlação linear entre as variáveis. Esse grau de correlação pode ser interpretado da seguinte maneira:

- $\rho = 1$: Perfeita correlação positiva entre as variáveis;
- $\rho = -1$: Perfeita correlação negativa entre as variáveis;
- $\rho = 0$: Variáveis não dependem linearmente uma da outra.

4.2.1 Centroavante

Na Tabela 4.4 podemos varificar uma alta variabilidade na variável passe para oportunidade de gol (POG), cerca de 9.50 e duelos aéreos vencidos (DAV), 34.51. Observamos

também uma média de 15.89 e 40.85, respectivamente.

Com relação as variáveis de índices criadas nesse estudo, temos algumas com valores entre 0 e 1 como a letalidade (LET), onde quanto maior o valor maior o aproveitamento do atleta, ou seja, caso seja 1 todos os chutes dados pelo jogador resultam em gol. Podemos ver que temos um máximo de 0.6 para esse indicador e um mínimo de 0.09 com um desvio de 0.15 em torno da média. O mesmo comportamento se dá para a váriável acurácia no chute (ACC). A varíavel gols por minutos (GPM) possui valores baixos, o que já era esperado e uma variabilidade pequena.

Tabela 4.4: Análise descritiva das variáveis de centroavante

Medidas	POG	DAV	LET	GPM	ACC
Mínimo	1.000	0.000	0.090	0.001	0.294
Média	15.890	40.850	0.279	0.004	0.458
Desvio	9.500	34.510	0.150	0.001	0.090
Máximo	37.000	155.000	0.600	0.007	0.666

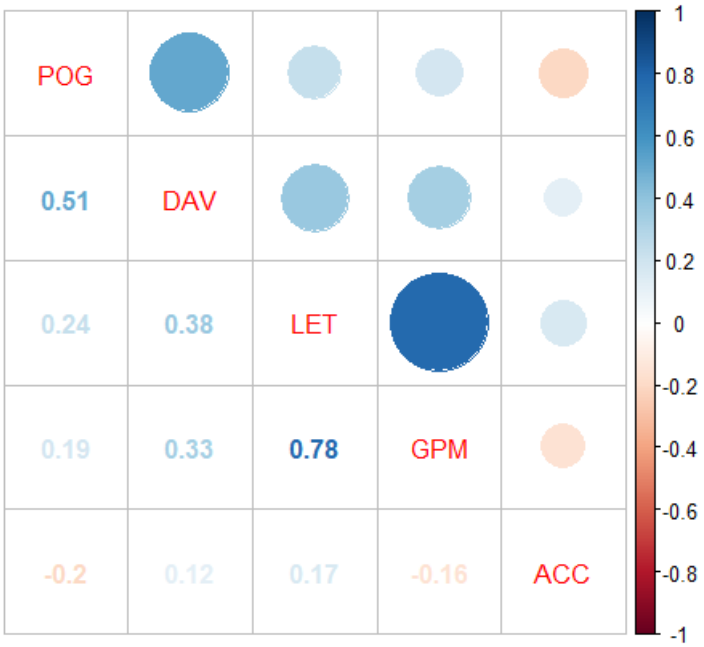


Figura 4.1: Correlação entre as variáveis de centroavante.

Na Figura 4.1 observamos uma forte correlação entre as variáveis gols por minuto (GPM) e letalidade (LET), 0.78, o que indica que um jogador que precisa de menos tempo em campo para marcar, também tem um melhor aproveitamento nas chances de gols. Observamos uma moderada correlação entre DAV e POG de 0.51 e algumas fracas

entre as demais, o que nos da indício da utilização de mais de um componente para o cálculo do *rating*.

4.2.2 Ponta

Verificamos na Tabela 4.5 que os pontas parecem ter uma média próxima a dos centroavantes na acurácia no chute. Eles tem um papel mais de auxiliar no ataque além do de fazer gols e possuem uma média de 2.48 de passes para oportunidade de gol, que é superior a dos centroavantes. Vemos uma certa variabilidade no total de assistências (TASS) com um mínimo de 0 e um máximo de 10. Em questão da letalidade e gols por minutos jogados a média dos centroavantes parecem ser maiores, o que já era de se esperar.

Tabela 4.5: Análise descritiva das variáveis de pontas

Medidas	TASS	POG	TXD	LET	GPM	ACC
Mínimo	0.000	6.000	0.102	0.000	0.000	0.200
Média	2.480	23.610	0.386	0.213	0.002	0.439
Desvio	2.282	13.703	0.125	0.122	0.001	0.123
Máximo	10.000	57.000	0.666	0.545	0.006	0.800

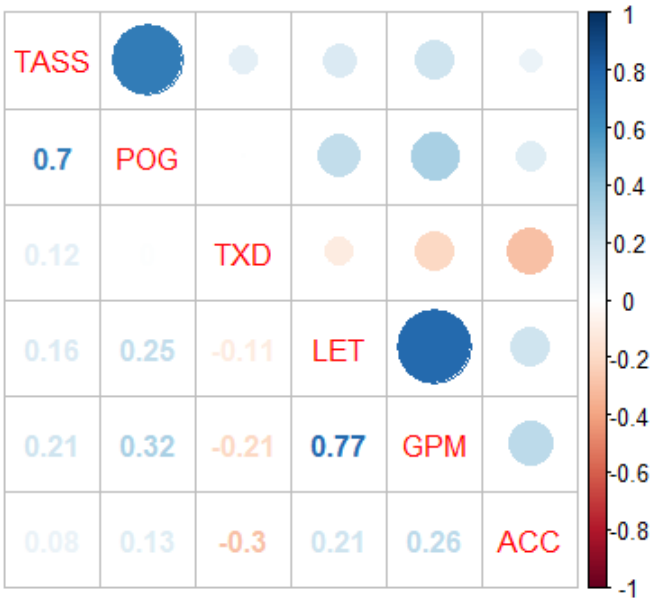


Figura 4.2: Correlação entre as variáveis de pontas.

Na Figura 4.2 observamos uma forte correlação entre letalidade e gols por minutos jogados de 0.77 e também entre total de assistências e passe para oportunidade de gol de

0.7, ou seja, jogadores que dão mais passes para oportunidade de gols também tem um maior número de assistências. Porém, no geral vemos correlações fracas entre as variáveis o que nos da indício do uso de mais de um componente para a criação do *rating*.

4.2.3 Meia ofensivo

Os meias tem como papel principal a criação de jogadas para os atacantes completarem para o gol. Assistências e passes para oportunidade de gol são características comuns entre os meias ofensivos e pontas. Podemos observar na Tabela 4.6 uma média de assistências de 2.88 maior para os meias ofensivos do que os pontas já que essa é sua principal função, assim como para passes para oportunidade de gols com 34.91.

Outra variável que também está ligada a assistências e passes para oportunidade de gols são os cruzamentos precisos (CRP) para os atacantes marcarem com uma média de 19.80 e um desvio de 15.12 em torno da média.

O drible também é uma característica fundamental dos meias ofensivos com o intuito de quebrar a marcação e criar espaço para jogadas, podemos observar os dribles certos (DRC) com uma média de 21.66 e uma grande variabilidade com desvio de 14.16 em torno da média.

A letalidade não é tão importante para um meia ofensivo mas fazer gols é uma característica secundária para o seu perfil. Sendo assim foi decidido analisar gols por jogo (GPJ) e não gols por minutos jogados. Observamos uma média de 0.15 gols por jogo com um máximo de 0.40, ou seja, quase 1 gol marcado a cada 2 jogos e um mínimo de 0.

Tabela 4.6: Análise descritiva das variáveis de meia ofensivo

Medidas	TASS	POG	DRC	CRP	GPJ
Mínimo	0.000	5.000	2.000	3.000	0.000
Média	2.886	34.910	21.660	19.800	0.155
Desvio	2.176	18.479	14.164	15.122	0.101
Máximo	7.500	82.000	64.000	55.000	0.407

Observamos na Figura 4.3 uma forte correlação entre passes para oportunidade de gol e cruzamentos precisos de 0.77, assim como passes para oportunidade de gols e total de assistências, cerca de 0.68. Observamos também uma média correlação entre total de assistências e cruzamentos precisos de 0.52, ou seja, quanto maior as chances de gols

criadas maior o número de assistências realizadas.

Conseguimos identificar uma baixa correlação entre gols por jogo e as outras variáveis já que é uma característica secundária dos meias e nos dá indício de um componente único para explicar essa variável.

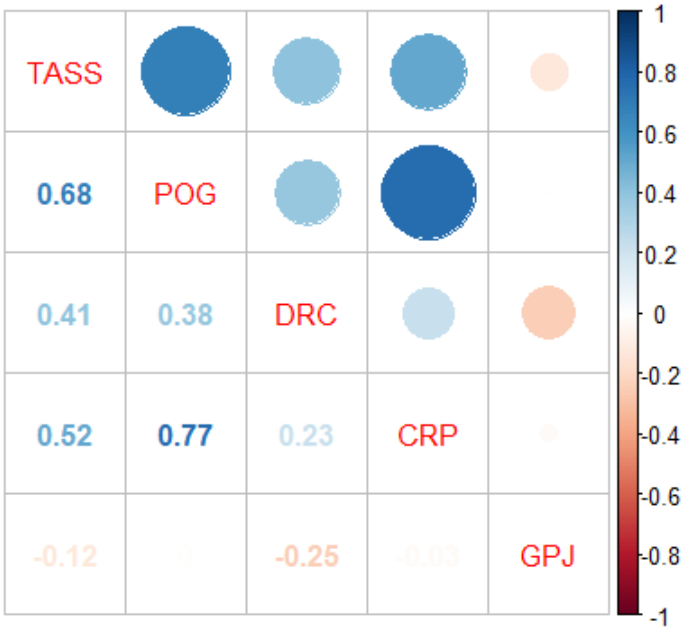


Figura 4.3: Correlação entre as variáveis de meia ofensivo.

4.2.4 Segundo volante

O segundo volante tem características tanto de armação de jogadas como de marcação para auxiliar o primeiro volante. Temos portanto um maior grupo de variáveis para essa posição.

Podemos ver um número de assistência menor do que o dos meias ofensivos na Tabela 4.7 com uma média de 0.6 e um desvio padrão de 0.77 em torno da média. As outras variáveis como POG e DRC também possuem médias menores que dos meias ofensivos, o que pode ser pelo fato deles possuirem praticamente duas funções principais.

No futebol moderno o que se espera de um segundo volante também é a chegada surpresa por trás da "zaga"a fim de surpreender os marcadores e fazer gols. Marcar gols nessa posição é mais raro de acontecer do que para um meia ofensivo e atacante já que ele tem que se preocupar também com a defesa. Dessa forma, o número de gols marcados (GOL) foi levado em consideração com uma média de 0.95, um desvio de 1.24 e um máximo de 5.

Com relação às variáveis defensivas temos o índice de efetividade criado que quanto maior seu valor, mais efetivo está sendo aquela ação para evitar gols já que o número de gols se encontra no denomidador. Temos uma efetividade média de 1.14 nos desarmes (EDES) e 0.59 nos duelos aéreos vencidos, indicando que para essa posição os desarmes estão sendo mais efetivos do que os duelos aéreos vencidos.

Tabela 4.7: Análise descritiva das variáveis de segundo volante

Medidas	TASS	POG	DRC	EINT	EDES	EDAV	ETRB	GOL
Mínimo	0.000	4.000	2.000	0.125	0.339	0.089	0.040	0.000
Média	0.607	12.620	9.714	0.820	1.140	0.593	0.929	0.952
Desvio	0.777	6.821	5.794	0.500	0.628	0.399	0.567	1.248
Máximo	2.500	29.000	26.000	1.966	2.744	1.607	2.333	5.000

Com relação as correlações entre as variáveis, a Figura 4.4 nos mostra uma forte correlação entre as de defesa e uma fraca entre as de ataque o que nos da indício de uso de mais de um componente para criação do *rating*.

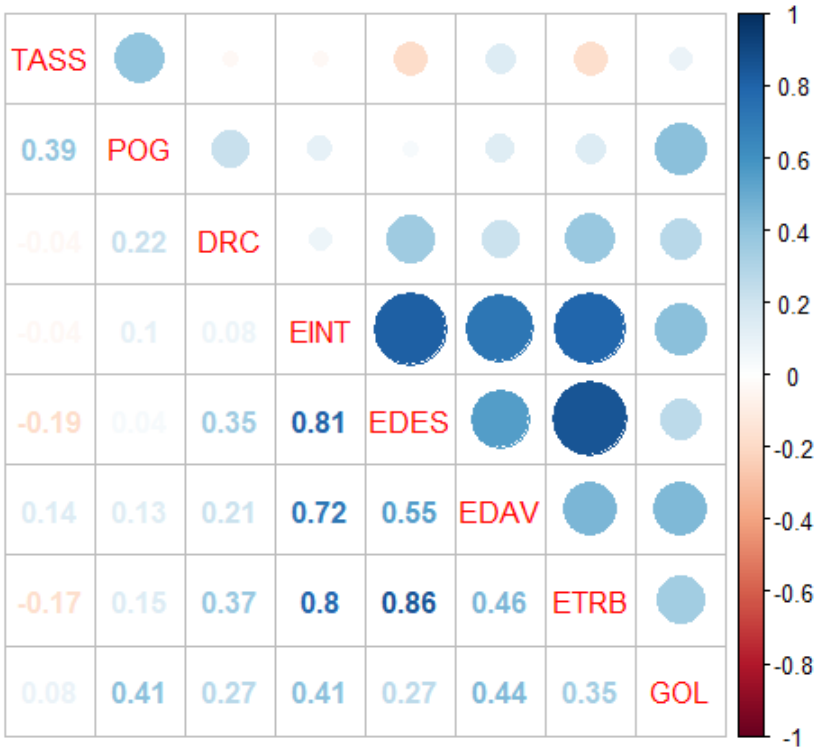


Figura 4.4: Correlação entre as variáveis de segundo volante.

4.2.5 Primeiro volante

O primeiro volante possui como característica principal a marcação com o intuito de tirar a pressão sobre os zagueiros e muitas vezes fazer a cobertura dos laterais quando os mesmos sobem para auxiliar no ataque.

Apenas os índices de efetividade na marcação foram utilizados na análise para esses jogadores, podemos ver na Tabela 4.8 uma efetividade média maior para desarmes de 0.940, interceptações (EINT), 0.702 e roubadas de bola (ETRB), 0.781 do que para bloqueios (EBLO), 0.084, rebatidas (EREB), 0.539 e duelos aéreos vencidos, 0.487 o que já era de se esperar já que as variáveis com menos efetividade afastam o perigo temporariamente.

Tabela 4.8: Análise descritiva das variáveis de primeiro volante

Medidas	EINT	EDES	EDAV	EREB	EBLO	ETRB
Mínimo	0.034	0.208	0.038	0.057	0.000	0.208
Média	0.702	0.940	0.487	0.539	0.084	0.781
Desvio	0.432	0.524	0.307	0.381	0.074	0.421
Máximo	1.868	2.205	1.261	1.562	0.289	1.631

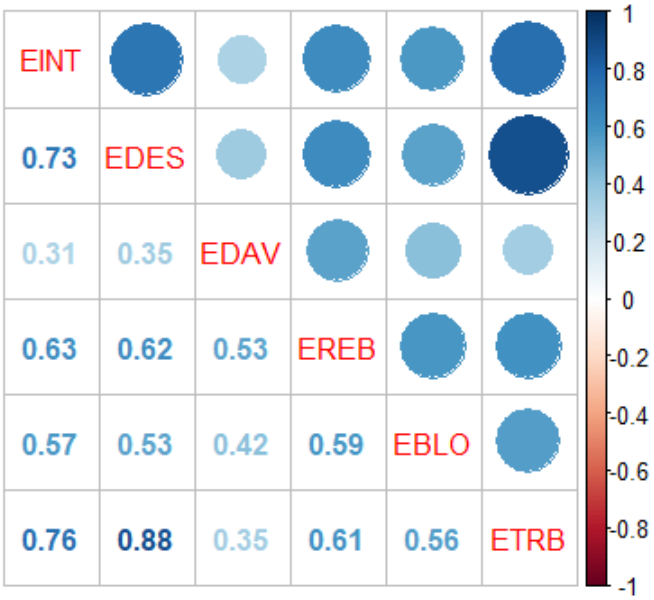


Figura 4.5: Correlação entre as variáveis de primeiro volante.

Na Figura 4.5 conseguimos verificar uma forte correlação entre as variáveis de marcação principalmente entre efetividade nos desarmes e roubadas de bola, 0.88 e efetividade nas

interceptações e roubadas de bola, 0.76 já que essas variáveis estão diretamente relacionadas com o fato de retirar a posse de bola do adversário.

4.2.6 Lateral

Os laterais precisam de um bom preparo físico e velocidade já que tem como funções principais apoiar o ataque e voltar para recompor a defesa. Com relação ao ataque tem como principais características o auxílio no cruzamento, que foi uma variável a ser levada em consideração na análise. Como mostra a Tabela 4.9 a variável cruzamentos precisos tem uma média de 19.2, um desvio de 11.4 em torno da média, um máximo de 51 e mínimo de 0.

Como os cruzamentos estão ligados a assistências e passes para oportunidade de gols, essas também foram varíaveis a ser avaliadas. Temos uma média de 21.96 passes para oportunidade de gols e 1.8 no total de assistências.

Já com relação a defesa, os indicadores de efetividade foram avaliados. Observamos uma média de 0.85 para interceptações e 0.98 para desarmes. A variável roubada de bola (TRB) não foi avaliada como efetividade e sim seu número total devido ao fato de muitas vezes os laterais recuperarem a posse para auxiliar no ataque com cruzamentos e não apenas para evitar o gol, temos uma média de 32.32 roubadas de bola com um desvio de 15.15 em torno da média, um mínimo de 4 e um máximo de 65.

Tabela 4.9: Análise descritiva das variáveis de laterais

Medidas	TRB	TSG	TASS	POG	EINT	EDES	CRP
Mínimo	4.000	2.000	0.000	0.000	0.250	0.244	0.000
Média	32.320	6.625	1.804	21.960	0.853	0.987	19.210
Desvio	15.151	2.844	1.765	11.762	0.403	0.470	11.401
Máximo	65.000	15.000	8.000	67.000	2.187	2.666	51.000

Analisamos na Figura 4.6 uma forte correlação entre o total de assistências e passes para oportunidade de gols, 0.71 e entre passes para oportunidade de gols e cruzamentos precisos, 0.87. Temos correlações fortes também entre as variáveis de defesa como eficiência nos desarmes e interceptações 0.66.

Já as outras variáveis nos mostra uma baixa correlação entre elas o que esperavamos devido ao fato delas medirem diferentes aspectos do jogador. Isso nos dá um indicativo

de que mais de um componente será necessário para a criação do *rating*.

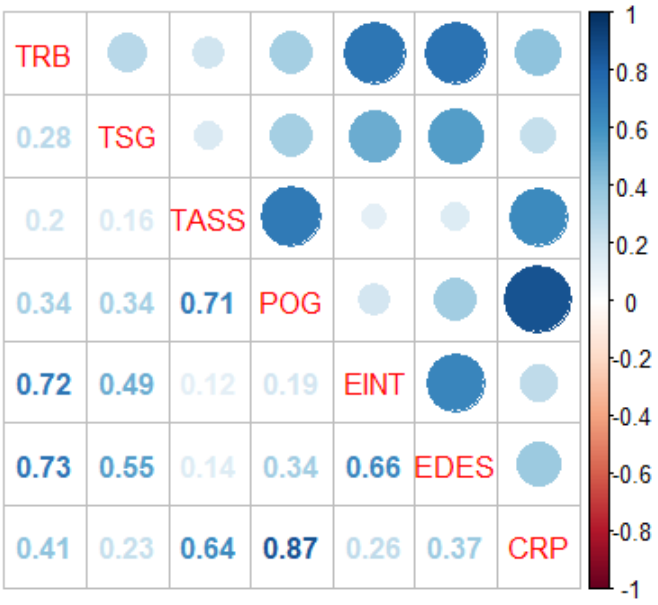


Figura 4.6: Correlação entre as variáveis de laterais.

4.2.7 Zagueiro

Os zagueiros tem como característica principal evitar gols, afastando o perigo de sua área. Dessa forma todas as variáveis de efetividade para evitar gols foram utilizadas. Verificamos na Tabela 4.10 uma média alta de efetividade nas rebatidas, cerca de 3, com um desvio de 1.59 em torno da média.

Os duelos aéros vencidos também são muito importantes para os zagueiros pois demonstra o domínio da sua área evitando que os atacantes ganhem a disputa. Temos uma média de eficiência dessa variável de 1.31. Foi utilizada também o total de jogos sem sofrer gols (TSG) o que mostra a solidez da defesa com uma média de 6.4, um desvio de 3.1 em torno da média, um mínimo de 1 e máximo de 15 jogos sem tomar gols.

Outra característica de zagueiros modernos é saber sair jogando com a bola para começar as jogadas até o ataque e não apenas dar chutões para frente fazendo ligação direta. Com o intuito de análisar essa característica foi utilizada a porcentagem de passes certos, com uma média de 83.6%, um desvio de 4.21%, mínimo de 65.42% e máximo de 90.86%.

Tabela 4.10: Análise descritiva das variáveis de zagueiros

Medidas	TSG	EINT	EDES	EREB	EDAV	EBLO	ETRB	PPC
Mínimo	1.000	0.125	0.107	0.333	0.196	0.000	0.107	65.420
Média	6.423	0.989	0.627	3.006	1.312	0.381	0.570	83.670
Desvio	3.158	0.465	0.308	1.598	0.729	0.275	0.269	4.213
Máximo	15.000	2.300	1.533	9.187	3.333	1.406	1.466	90.860

Na Figura 4.7 observamos uma forte correlação entre praticamente todas as variáveis de efetividade. Vale destacar uma fraca correlação entre a porcentagem de passes certos com as demais o que nos da indício de que será necessário mais de um componente para a criação do *rating*, sendo que um deles será praticamente para explicar essa variável.

Figura 4.7: Correlação entre as variáveis de zagueiro.

4.2.8 Goleiro

A Tabela 4.11 nos da informação sobre as características dos goleiros. Temos uma média de 0.83 defesas de pênaltis com um mínimo de 0 e um máximo de 4 com um desvio de 1.16 em torno da média. Podemos observar também o poder defensivo, em que um

número maior que 1 mostra que temos mais defesas difíceis do que gols tomados e um número menor que 1 o contrário. Para esse indicador temos uma média de 1.34 com um pequeno desvio de 0.43.

Outra variável analisada é o total de jogos sem sofrer gols e que relacionado com o poder defensivo pode vir a explicar a qualidade do goleiro.

Tabela 4.11: Análise descritiva das variáveis de goleiros

Medidas	TDP	TSG	PDEF
Mínimo	0.000	0.000	0.795
Média	0.833	6.000	1.348
Desvio	1.167	3.283	0.434
Máximo	4.000	13.00	2.826

Na Figura 4.8 vemos uma correlação forte entre o número de jogos sem sofrer gols e o poder defensivo. A variável defesa de pênalti tem baixa correlação o que pode indicar o uso de mais de um componente para criação do *rating*.

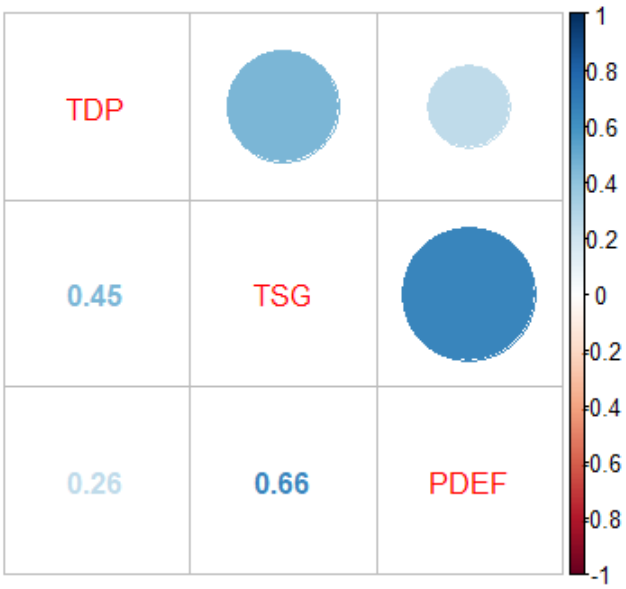


Figura 4.8: Correlação entre as variáveis de goleiro.

4.3 *Rating* dos jogadores

Através da técnica de componentes principais o *rating* foi construído a partir das variáveis descritas anteriormente. Será apresentado a seguir esse escore para cada posição de jogadores e serão listados os 5 melhores por posição a fim de escolher a seleção dos

melhores jogadores do campeonato brasileiro de 2017.

4.3.1 Centroavante

Na Tabela 4.12 encontramos os autovalores e a porcentagem de variância explicada por cada componente referente aos centroavantes, verificamos que dois componentes representam aproximadamente 68.2% da variabilidade total dos dados.

Tabela 4.12: Autovalores e % de variância explicada: centroavante

Componente	Autovalor	% Variância explicada
CP1	2.236	44.722
CP2	1.176	23.525
CP3	1.033	20.667
CP4	0.410	8.207
CP5	0.143	2.877

A Figura 4.9 nos indica o uso de 3 componentes para análise, porém devido ao fato de 2 deles explicarem já 68.2% da variabilidade total dos dados e também estarem representando bem todas as variáveis, foi decidido continuar apenas com 2 componentes para criação do *rating*.

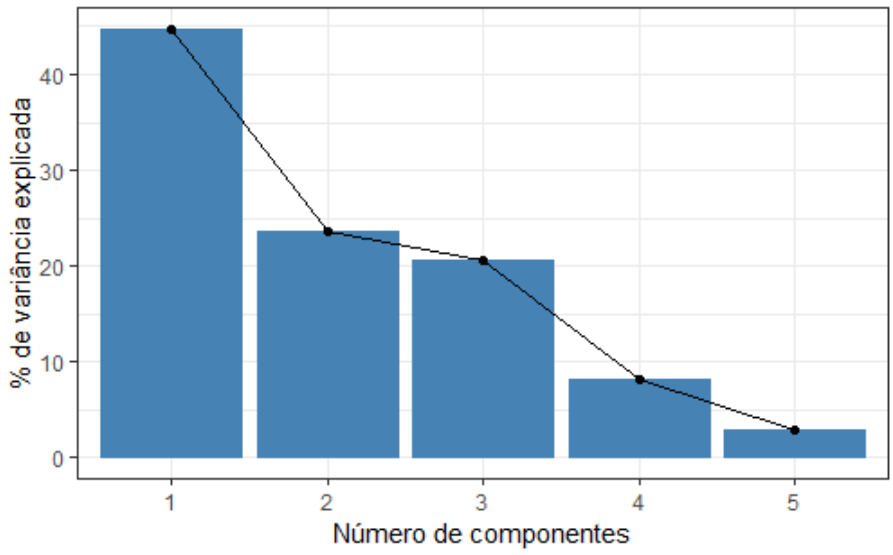


Figura 4.9: Gráfico de cotovelo para os componente principais: centroavante

Na Tabela 4.13 conseguimos verificar o peso de cada variável no componente, observamos que o primeiro componente da um peso maior para as variáveis DAV, LET e GPM e um peso médio para POG que seria uma característica secundária dos centroavantes. Temos que o segundo componente é necessariamente para explicar a variável acurácia no chute.

Podemos verificar como os jogadores se distribuem nos componentes através da Figura 4.10 com destaque pros jogadores Henquique Dourado e Jô no primeiro componente sendo que Henrique Dourado está acima da média com relação a acurácia no chute que é o segundo componente.

Tabela 4.13: Autovetores e pesos dos componentes: centroavante

Variáveis	CP1	CP2
POG	0.588	-0.566
DAV	0.719	-0.079
LET	0.843	0.368
GPM	0.814	0.117
ACC	-0.018	0.837

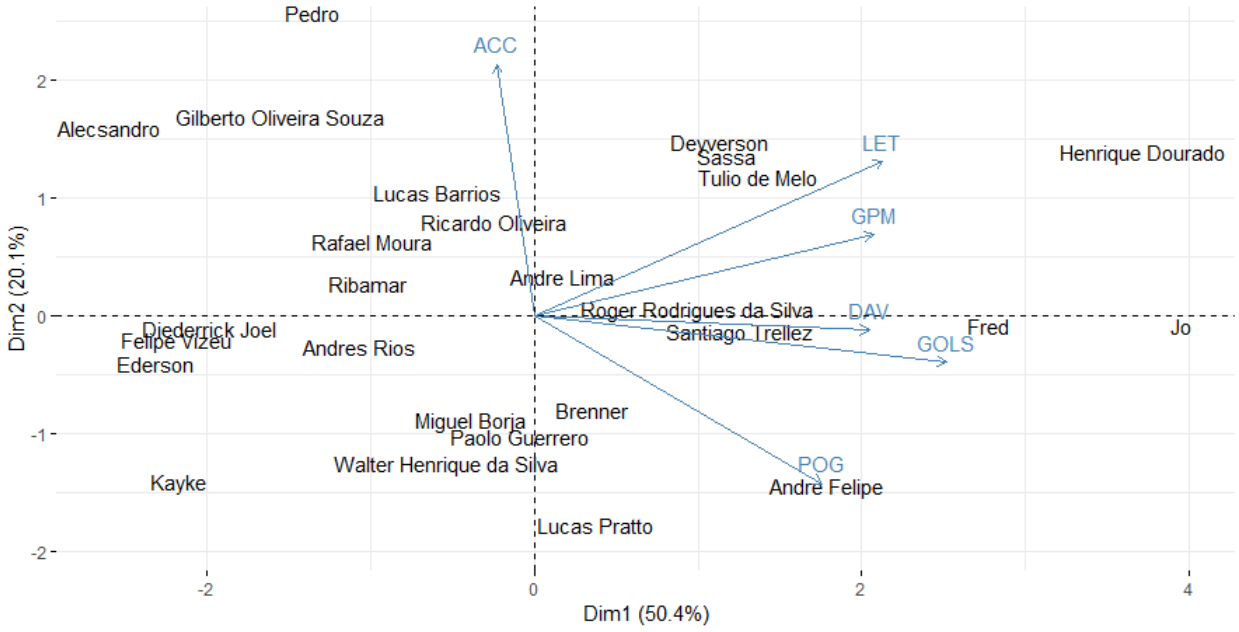


Figura 4.10: Biplot dos centroavantes com relação aos dois componentes analisados

Usando a variância de cada componente como seu respectivo peso, uma média ponde-

rada foi calculada entre os dois componentes para geração do *rating* e exibida na Tabela 4.14 considerando os 5 melhores centroavantes de acordo com nosso indicador.

Para a obtenção desse *rating* temos a seguinte equação:

$$\begin{aligned} Y_i &= 0.44722 * (0.588 * POG_i + 0.719 * DAV_i + 0.843 * LET_i + 0.814 * GPM_i \\ &\quad -0.018 * ACC_i) + 0.23525 * (-0.566 * POG_i - 0.079 * DAV_i + 0.368 * LET_i + \\ &\quad 0.117 * GPM_i + 0.837 * ACC_i)/(0.44722 + 0.23525), \end{aligned} \tag{4.2}$$

em que i representa o centroavante. O valor do *rating* é obtido pela padronização (4.1) do valor resultante de (4.2).

Tabela 4.14: *Rating* e informações dos centroavantes

Jogadores	POG	DAV	LET	GPM	ACC	*Rating*
Henrique Dourado	26	84	0.600	0.006	0.566	10.000
Jô	24	155	0.428	0.005	0.452	9.002
Fred	35	66	0.545	0.005	0.454	8.023
Tulio de Melo	14	80	0.3888	0.006	0.500	7.679
Sassa	12	5	0.600	0.007	0.400	7.620

4.3.2 Ponta

Na Tabela 4.15 encontramos os autovalores e a porcentagem de variância explicada por cada componente referente aos pontas. Verificamos que dois componentes representam aproximadamente 63.5% da variabilidade total dos dados.

Tabela 4.15: Autovalores e % de variância explicada: ponta

Componente	Autovalor	% Variância explicada
CP1	2.331	38.856
CP2	1.470	24.511
CP3	1.010	16.847
CP4	0.681	11.352
CP5	0.289	4.830
CP6	0.216	3.600

A Figura 4.11 nos indica o uso de 2 componentes para análise, pois como podemos observar a partir de 2, não temos muito ganho significativo na explicação da variabilidade. Dessa forma apenas eles foram utilizados para criação do *rating*.

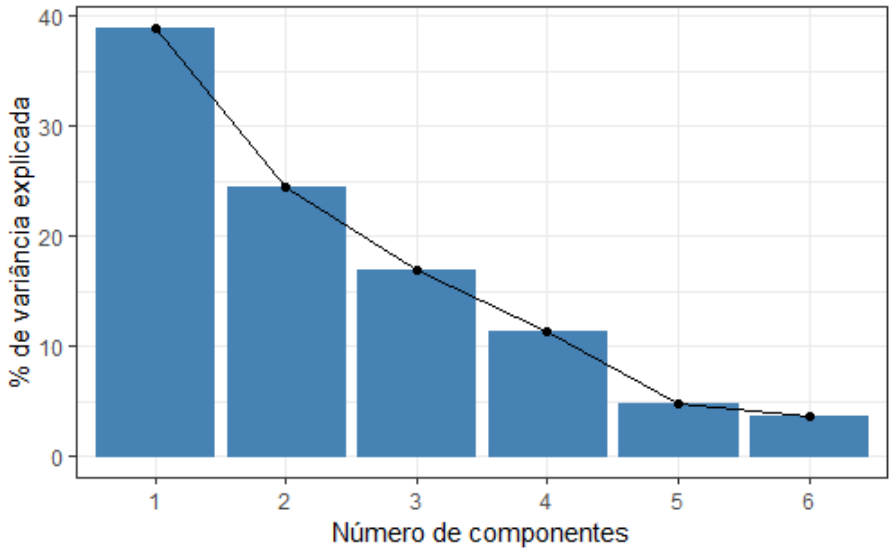


Figura 4.11: Gráfico de cotovelo para os componente principais: ponta

Tabela 4.16: Autovetores e pesos dos componentes: ponta

Variáveis	CP1	CP2
TASS	0.586	0.678
POG	0.691	0.547
TXD	-0.234	0.621
LET	0.764	-0.278
GPM	0.824	-0.290
ACC	0.440	-0.405

Na Tabela 4.16 conseguimos verificar o peso de cada variável no componente. O primeiro componente explica tanto as variáveis de assitência para gols quanto a marcação de gols, sendo a última com maior peso. Os dribles não entraram com muito peso nesse componente sendo explicados pelo segundo que também retira um peso das variáveis referentes a finalização e dá maior peso aos dribles e auxílio aos centroavantes.

Podemos verificar como os jogadores se distribuem nos componentes através da Figura 4.12 com destaque para os jogadores Dudu, Robinho, Bruno Henquique e Edgar Júnior

no primeiro componente e para Rodrigo Pimpão no segundo que representa a taxa de dribles, assistências e passes para oportunidade de gols.

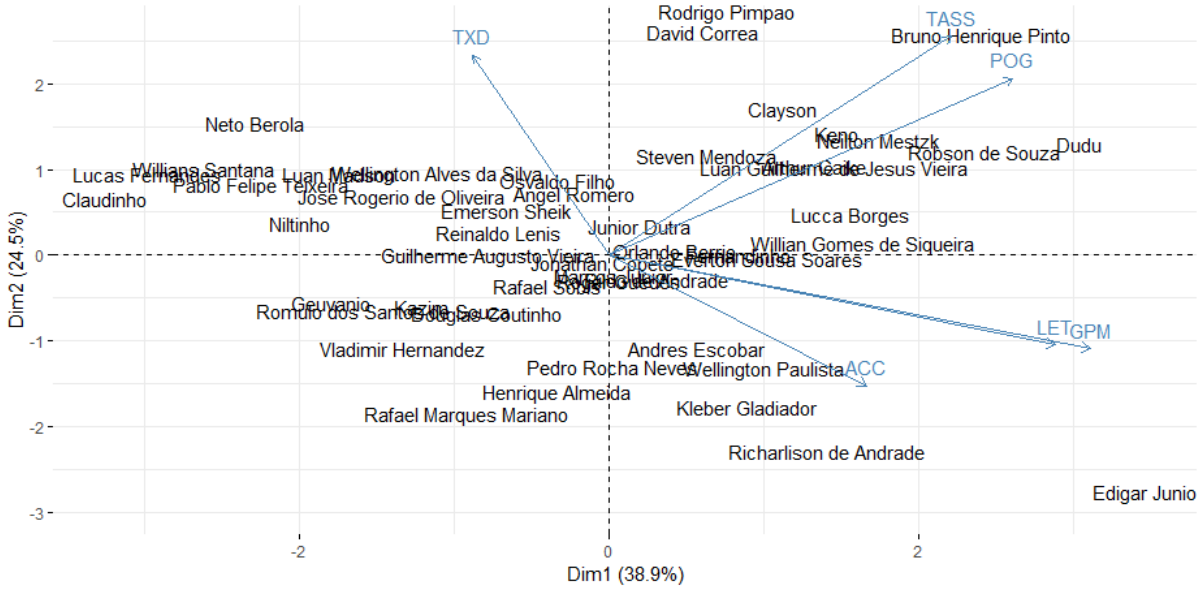


Figura 4.12: Biplot dos pontas com relação aos dois componentes analisados

Usando a variância de cada componente como seu respectivo peso, uma média ponderada foi calculada entre os dois para geração do *rating*. A Tabela 4.17 nos mostra os 5 melhores pontas de acordo com nosso indicador.

Tabela 4.17: *Rating* e informações dos pontas

Jogadores	TASS	POG	TXD	LET	GPM	ACC	*Rating*
Bruno Henrique Pinto	10	42	0.447	0.266	0.003	0.500	10.000
Dudu	5	57	0.473	0.346	0.004	0.615	9.755
Robson de Souza	5.5	44	0.403	0.466	0.003	0.400	8.781
Neilton Mestzk	6	39	0.403	0.269	0.003	0.461	7.902
Rodrigo Pimpão	8.5	50	0.343	0.058	0.001	0.500	7.864

4.3.3 Meia ofensivo

Na Tabela 4.18 encontramos os autovalores e a porcentagem de variância explicada por cada componente referente aos meias ofensivos. Verificamos que dois componentes representam aproximadamente 73.8% da variabilidade total dos dados.

Tabela 4.18: Autovalores e % de variância explicada: meia ofensivo

Componente	Autovalor	% Variância explicada
CP1	2.559	51.199
CP2	1.131	22.636
CP3	0.680	13.615
CP4	0.443	8.877
CP5	0.183	3.670

A Figura 4.13 nos indica o uso de apenas 2 componentes para análise, pois como podemos observar a partir de 2, não temos muito ganho significativo na explicação da variabilidade. Dessa forma a média ponderada pela variância desses componentes foi utilizada para criação do *rating*.

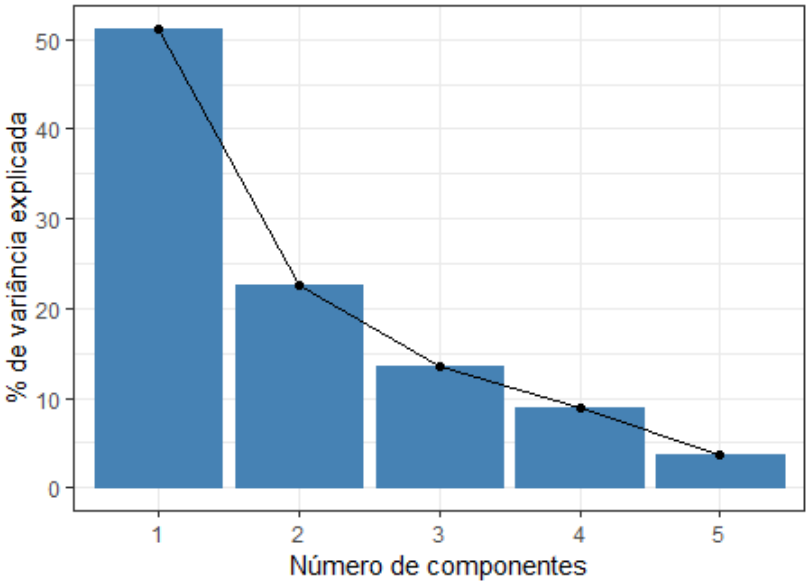


Figura 4.13: Gráfico de cotovelo para os componente principais: meia ofensivo

Na Tabela 4.19 conseguimos verificar o peso de cada variável no componente. O primeiro componente explica em geral as variáveis de assistência para os atacantes e o drible para criação de jogadas e o segundo entra exclusivamente para explicar os gols por jogo.

Podemos verificar como os jogadores se distribuem nos componentes através da Figura 4.14 com destaque para os jogadores Lucas Lima e Cazares no primeiro componente que

estão muito acima da média e para Diego Souza e Thiago Neves que estão acima na média no segundo.

Tabela 4.19: Autovetores e pesos dos componentes: meia ofensivo

Variáveis	CP1	CP2
TASS	0.835	-0.004
POG	0.910	0.227
DRC	0.588	-0.491
CRP	0.813	0.290
GPJ	-0.173	0.869

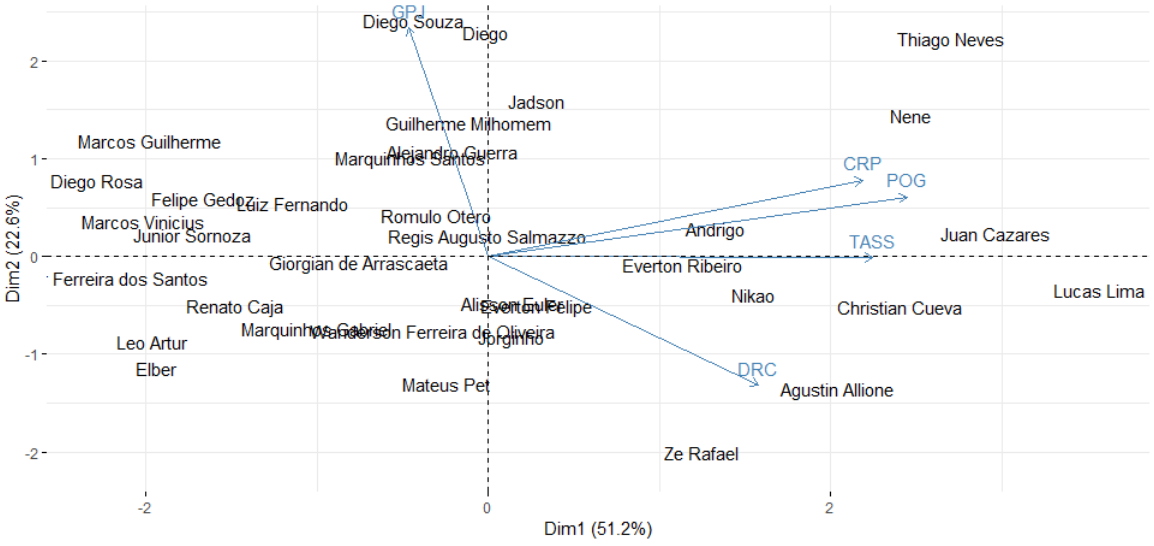


Figura 4.14: Biplot dos meias ofensivos com relação aos dos dois componentes analisados

Tabela 4.20: *Rating* e informações dos meias ofensivos

Jogadores	TASS	POG	DRC	CRP	GPJ	*Rating*
Thiago Neves	6.500	70.000	28.000	42.000	0.333	10.000
Lucas Lima	6.000	82.000	38.000	43.000	0.040	9.561
Nenê	5.000	65.000	15.000	55.000	0.161	9.045
Juan Cazares	7.500	54.000	29.000	49.000	0.100	8.981
Christian Cueva	7.000	54.000	41.000	27.000	0.107	7.528

Usando a variância de cada componente como seu respectivo peso, uma média ponderada foi calculada entre os dois para geração do *rating*. A Tabela 4.20 nos mostra os 5 melhores meias ofensivo de acordo com nosso indicador.

4.3.4 Segundo volante

Na Tabela 4.21 encontramos os autovalores e a porcentagem de variância explicada por cada componente referente ao segundo volante. Verificamos que dois componentes representam aproximadamente 64% da variabilidade total dos dados.

Tabela 4.21: Autovalores e % de variância explicada: segundo volante

Componente	Autovalor	% Variância explicada
CP1	3.529	44.118
CP2	1.599	19.993
CP3	1.0314	12.893
CP4	0.704	8.805
CP5	0.609	7.624
CP6	0.334	4.183
CP7	0.120	1.503
CP8	0.070	0.876

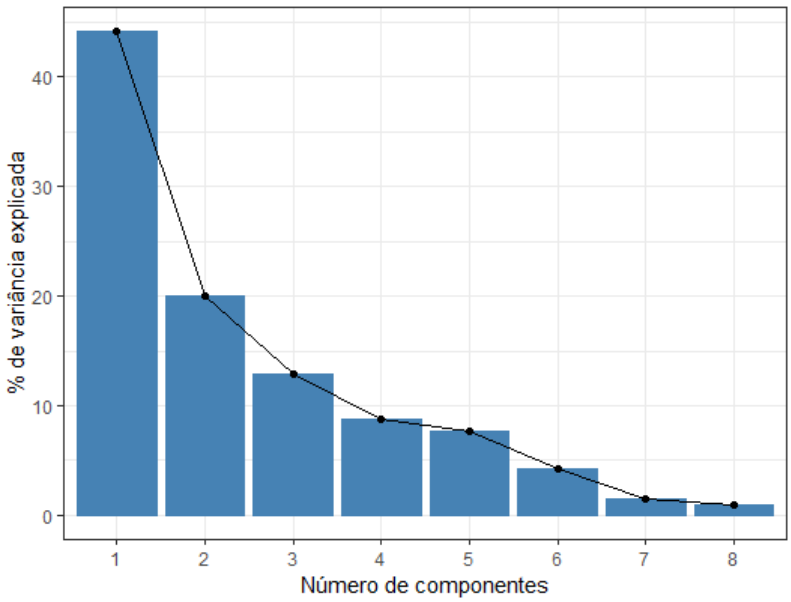


Figura 4.15: Gráfico de cotovelo para os componente principais: segundo volante

A Figura 4.15 nos indica o uso de 2 componentes para análise, pois como podemos observar a partir de 2, não temos muito ganho significativo na explicação da variabilidade. Dessa forma a média ponderada desses componentes foi utilizada para criação do *rating*.

Na Tabela 4.22 conseguimos verificar o peso de cada variável no componente. O primeiro componente explica com maior peso as variáveis de características defensivas e o segundo tira um pouco um peso dessas variáveis e dá valor as variáveis de poder ofensivo.

Podemos verificar como os jogadores se distribuem nos componentes através da Figura 4.16 com destaque para os jogadores Renê Júnior e Michel Ferreira no primeiro componente referente ao poder de marcação e Rodrigo Lindoso e Paulo Roberto no segundo componente com relação ao poder ofensivo.

Tabela 4.22: Autovetores e pesos dos componentes: segundo volante

Variáveis	CP1	CP2
TASS	-0.043	0.763
POG	0.266	0.791
DRC	0.432	0.153
EINT	0.897	-0.147
EDES	0.882	-0.302
EDAV	0.758	0.115
ETRB	0.881	-0.215
GOL	0.582	0.442

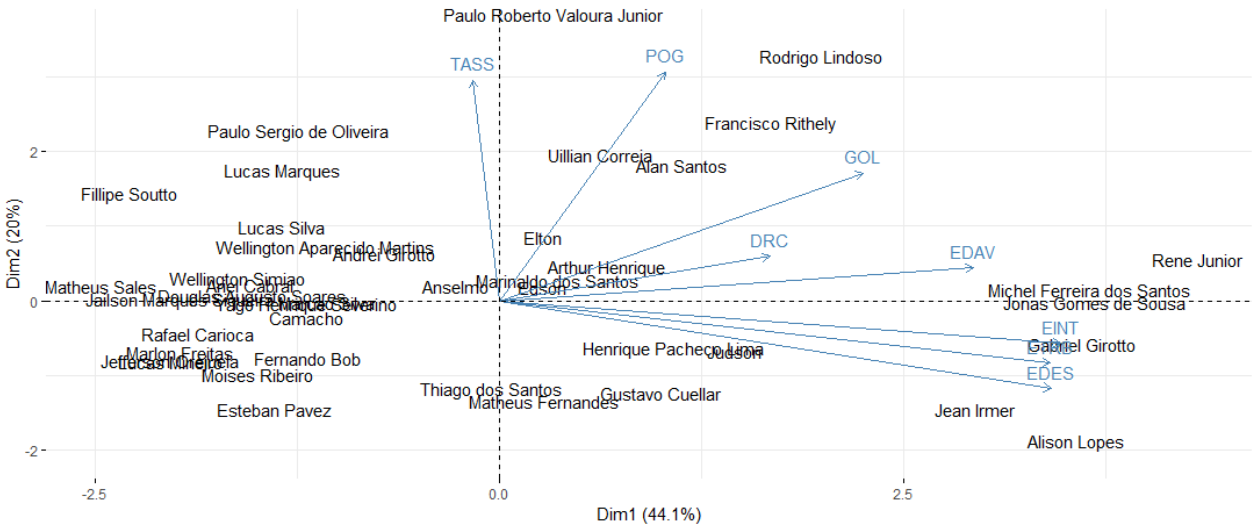


Figura 4.16: Biplot dos segundos volantes com relação aos dois componentes analisados

Usando a variância de cada componente como seu respectivo peso, uma média ponderada foi calculada entre os dois para geração do *rating*. A Tabela 4.23 nos mostra os 5

melhores segundos volantes de acordo com nosso indicador.

Tabela 4.23: *Rating* e informações dos segundos volantes

Jogadores	TASS	POG	DRC	EINT	EDES	EDAV	ETRB	GOL	*Rating*
Renê Junior	0	17	26	1.229	2.479	1.104	1.958	4	10.000
Michel Ferreira	0	8	13	1.805	1.583	1.416	1.472	5	8.782
Jonas	0	16	18	1.274	2.274	1.607	1.725	2	8.716
Rodrigo Lindoso	2	29	5	1.428	1.119	1.214	1.119	3	8.428
Gabriel Girotto	0	21	13	1.966	2.433	0.600	2.333	1	8.236

4.3.5 Primeiro volante

Na Tabela 4.24 encontramos os autovalores e a porcentagem de variância explicada por cada componente referente ao primeiro volante. Verificamos que dois componentes representam aproximadamente 78.9% da variabilidade total dos dados.

Tabela 4.24: Autovalores e % de variância explicada: primeiro volante

Componente	Autovalor	% Variância explicada
CP1	3.870	64.513
CP2	0.867	14.458
CP3	0.505	8.419
CP4	0.357	5.962
CP5	0.278	4.641
CP6	0.120	2.004

A Figura 4.17 nos indica o uso de 2 componentes para análise, pois como podemos observar a partir de 2, não temos muito ganho significativo na explicação da variabilidade. Dessa forma a média ponderada desses componentes foi utilizada para criação do *rating*.

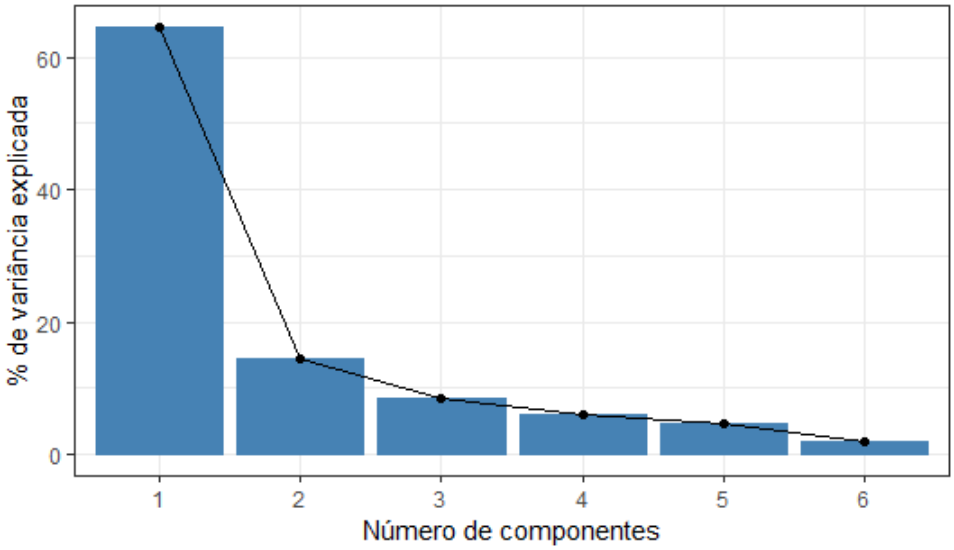


Figura 4.17: Gráfico de cotovelo para os componente principais: primeiro volante

Na Tabela 4.25 conseguimos verificar o peso de cada variável no componente. O primeiro componente explica praticamente todas as variáveis com excessão ao duelo aéreo vencidos que é explicada pelo segundo como um indicador de domínio de meio de campo no jogo aéreo.

Podemos verificar como os jogadores se distribuem nos componentes através da Figura 4.18 com destaque para os jogadores Bruno Silva, William Arão e Lucas Romero no primeiro compoente, sendo que os dois primeiros também tem um valor alto no segundo.

Tabela 4.25: Autovetores e pesos dos componentes: primeiro volante

Variáveis	CP1	CP2
EINT	0.852	-0.258
EDES	0.878	-0.281
EDAV	0.572	0.748
EREB	0.828	0.226
EBLO	0.759	0.151
ETRB	0.886	-0.297

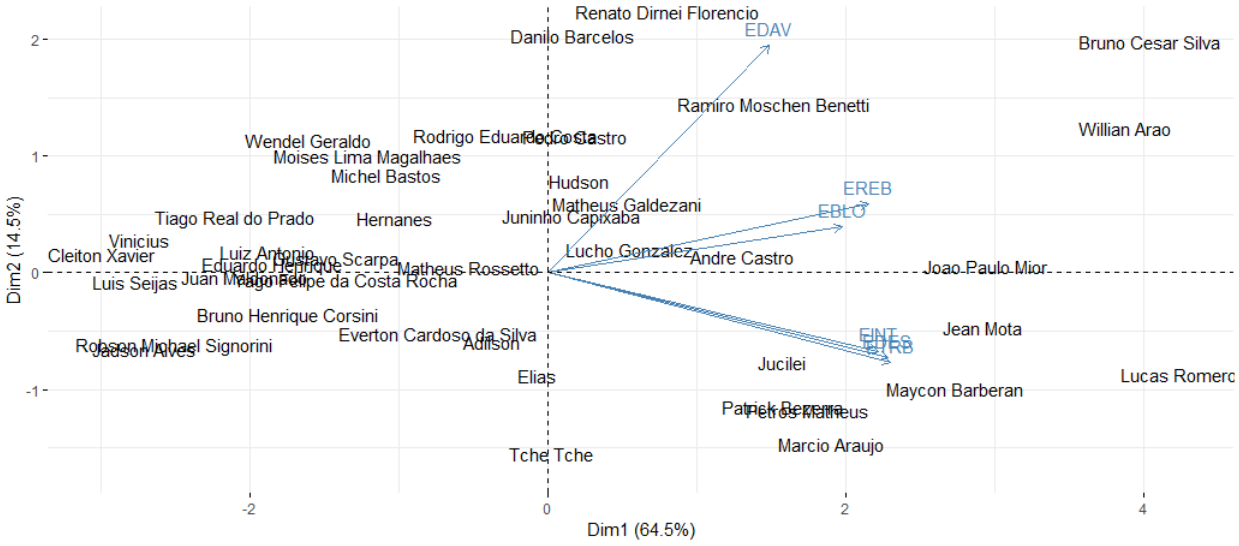


Figura 4.18: Biplot dos primeiros volantes com relação aos dois componentes analisados

Usando a variância de cada componente como seu respectivo peso, uma média ponderada foi calculada entre os dois para geração do *rating*. A Tabela 4.26 nos mostra os 5 melhores primeiros volantes de acordo com nosso indicador.

Tabela 4.26: *Rating* e informações dos primeiros volantes

Jogadores	EINT	EDES	EDAV	EREB	EBLO	ETRB	*Rating*
Bruno Cesar Silva	0.880	1.904	1.261	1.357	0.238	1.309	10.000
Willian Arão	1.157	1.526	1.157	0.789	0.289	1.631	9.563
Lucas Romero	1.743	2.205	0.564	1.179	0.205	1.410	9.418
João Paulo Mior	1.404	1.595	0.595	1.095	0.214	1.023	7.944
Jean Mota	1.187	1.656	0.562	1.562	0.031	1.593	7.762

4.3.6 Lateral

Na Tabela 4.27 encontramos os autovalores e a porcentagem de variância explicada por cada componente referente aos laterais. Verificamos que dois componentes representam aproximadamente 75.1% da variabilidade total dos dados.

Tabela 4.27: Autovalores e % de variância explicada: lateral

Componente	Autovalor	% Variância explicada
CP1	3.533	50.477
CP2	1.731	24.732
CP3	0.760	10.860
CP4	0.413	5.905
CP5	0.296	4.234
CP6	0.165	2.358
CP7	0.100	1.430

A Figura 4.19 nos indica o uso de 2, componentes para análise, pois como podemos observar a partir de 2, não temos muito ganho significativo na explicação da variabilidade. Dessa forma a média ponderada desses componentes foi utilizada para criação do *rating*.

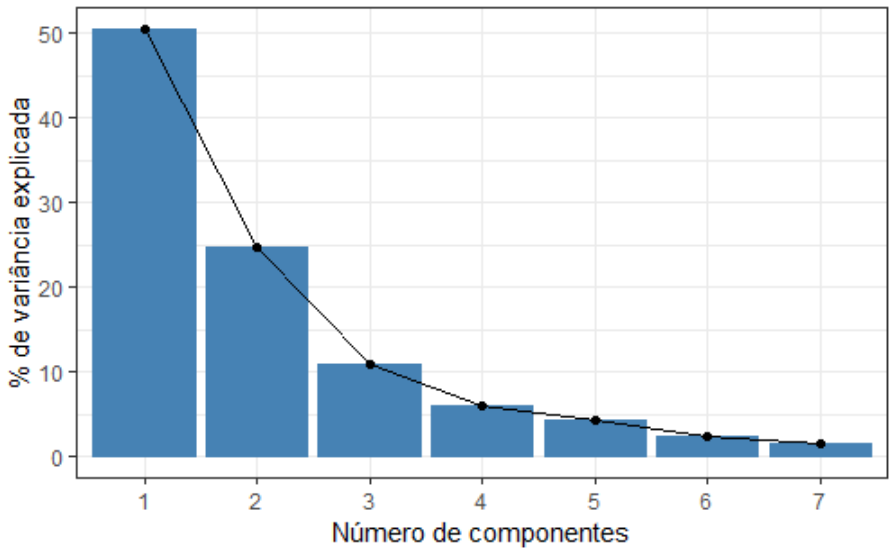


Figura 4.19: Gráfico de cotovelo para os componente principais: lateral

Na Tabela 4.28 conseguimos verificar o peso de cada variável no componente. O primeiro componente explica praticamente todas as variáveis com peso menor para o total de assistência que é explicada pelo segundo.

Podemos verificar como os jogadores se distribuem nos componentes através da Figura 4.20 com destaque para os jogadores Fagner e Reinaldo no primeiro componente estando bem acima da média. Com relação ao segundo, Fagner parece ter um poder maior de

marcação enquanto Reinaldo é um lateral que apoia mais o ataque.

Tabela 4.28: Autovetores e pesos dos componentes: lateral

Variáveis	CP1	CP2
TRB	0.759	-0.381
TSG	0.598	-0.293
TASS	0.579	0.646
POG	0.763	0.565
EINT	0.697	-0.548
EDES	0.779	-0.449
CRP	0.767	0.510

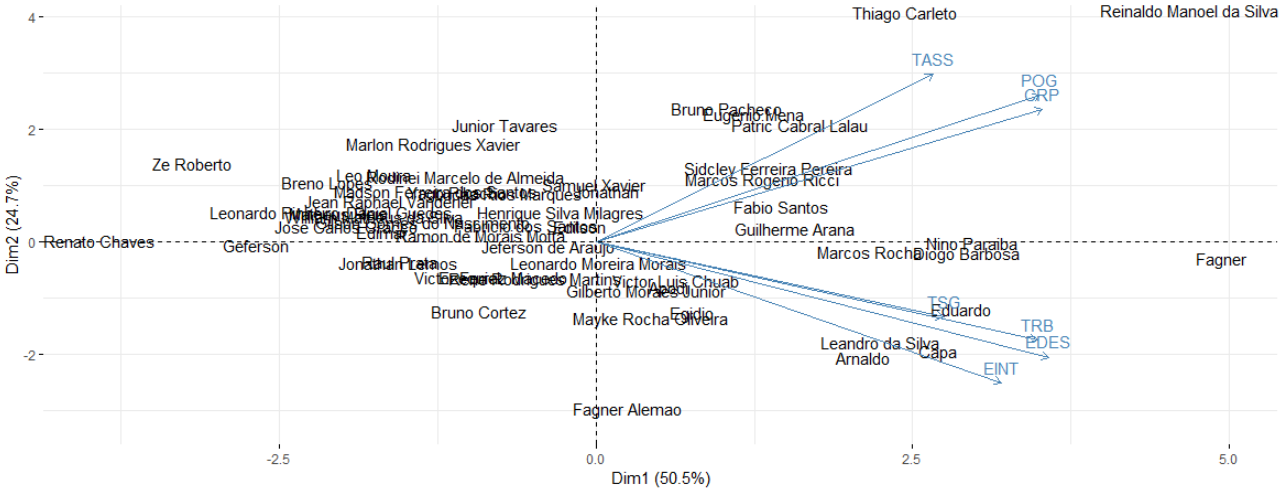


Figura 4.20: Biplot dos laterais com relação aos dois componentes analisados

Usando a variância de cada componente como seu respectivo peso, uma média ponderada foi calculada entre as dois para geração do *rating* na qual a Tabela 4.29 nos mostra os 5 melhores laterais de acordo com nosso indicador.

Tabela 4.29: *Rating* e informações dos laterais

Jogadores	TRB	TSG	TASS	POG	EINT	EDES	CRP	*Rating*
Reinaldo Manoel da Silva	50	8	8	67	0.857	1.265	51	10.000
Fagner	49	13	3	44	1.333	2.666	43	8.197
Thiago Carleto	31	6	8	42	0.784	0.745	48	7.865
Nino Paraíba	53	7	3	26	1.634	1.480	43	6.475
Diogo Barbosa	45	9	4	36	1.461	1.871	22	6.356

4.3.7 Zagueiro

Na Tabela 4.30 encontramos os autovalores e a porcentagem de variância explicada por cada componente referente aos zagueiros, verificamos que dois componentes representam aproximadamente 77.1% da variabilidade total dos dados.

Tabela 4.30: Autovalores e % de variância explicada: zagueiro

Componente	Autovalor	% Variância explicada
CP1	5.264	65.811
CP2	0.905	11.316
CP3	0.739	9.244
CP4	0.394	4.929
CP5	0.329	4.119
CP6	0.185	2.321
CP7	0.114	1.425
CP8	0.066	0.832

A Figura 4.21 nos indica o uso de 2 componentes para análise, pois como podemos observar a partir de 2, não temos muito ganho significativo de variabilidade. Dessa forma a média ponderada desses componentes foi utilizada para criação do *rating*.

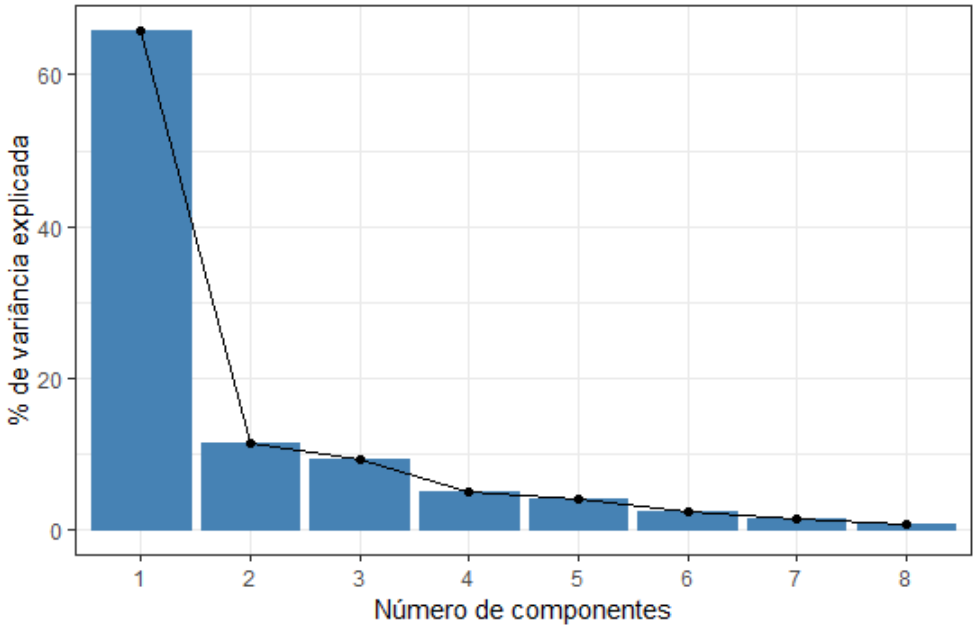


Figura 4.21: Gráfico de cotovelo para os componente principais: zagueiro

Na Tabela 4.31 conseguimos verificar o peso de cada variável no componente. O primeiro componente explica praticamente todas as variáveis com peso menor para a porcentagem de passes certos que é explicada praticamente sozinha no segundo.

Podemos verificar como os jogadores se distribuem nos componentes através da Figura 4.22 com destaque para os jogadores Lucas Veríssimo e Fabian Balbuena no primeiro componente. No segundo componente, Lucas Verissimo parece ter uma melhor qualidade no passe dando destaque também para Pablo e Thiago Heleno.

Tabela 4.31: Autovetores e pesos dos componentes: zagueiro

Variáveis	CP1	CP2
TSG	0.830	0.106
EINT	0.913	-0.207
EDES	0.775	-0.416
EREB	0.927	0.210
EDAV	0.794	0.102
EBLO	0.854	0.146
ETRB	0.835	-0.363
PPC	0.475	0.686

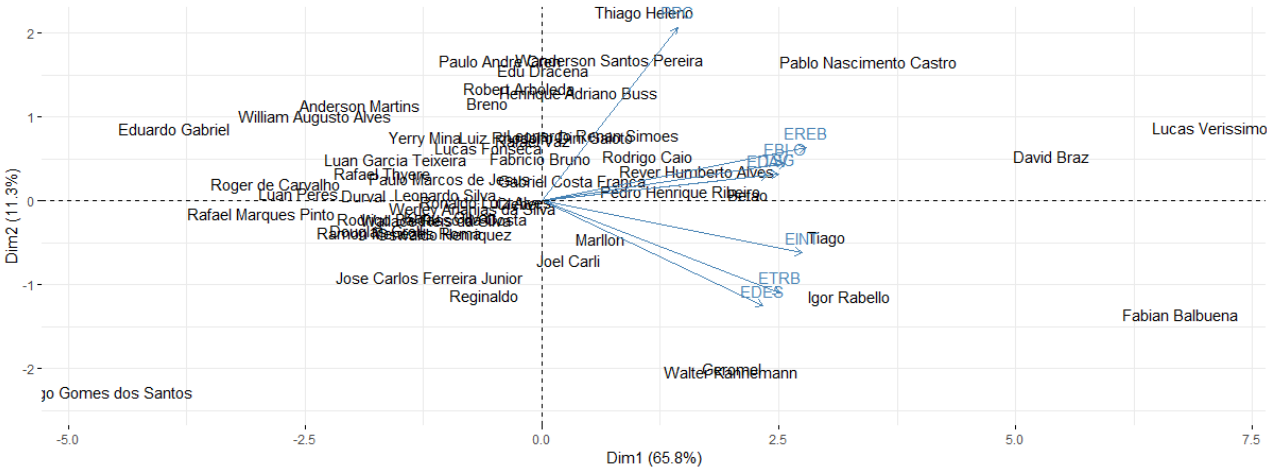


Figura 4.22: Biplot dos zagueiros com relação aos dois componentes analisados

Usando a variância de cada componente como seu respectivo peso, uma média ponderada foi calculada entre os dois para geração do *rating*. A Tabela 4.32 nos mostra os 5 melhores zagueiros de acordo com nosso indicador.

Tabela 4.32: *Rating* e informações dos zagueiros

Jogadores	TSG	EINT	EDES	EREB	EDAV	EBLO	ETRB	PPC	*Rating*
Lucas Verissimo	15.000	2.031	1.156	9.187	3.218	1.187	1.125	88.454	10.000
Fabian Balbuena	14.000	2.300	1.533	7.166	3.333	0.666	1.466	85.679	9.436
David Braz	13.000	1.812	1.125	6.593	2.000	1.406	0.968	87.150	8.584
Pablo Nascimento	12.000	1.366	0.766	6.133	2.200	0.700	0.800	90.860	7.175
Igor Rabello	7.000	1.857	0.952	4.476	2.595	0.928	0.952	80.052	6.621

4.3.8 Goleiro

Na Tabela 4.33 encontramos os autovalores e a porcentagem de variância explicada por cada componente referente aos goleiros. Verificamos que dois componentes representam aproximadamente 89.8% da variabilidade total dos dados.

Tabela 4.33: Autovalores e % de variância explicada: goleiro

Componente	Autovalor	% Variância explicada
CP1	1.930	64.352
CP2	0.763	25.452
CP3	0.305	10.195

A Figura 4.23 nos indica o uso de 2 componentes para análise, pois como podemos observar a partir de 2, não temos muito ganho significativo na explicação da variabilidade. Dessa forma a média ponderada desses componentes foi utilizada para criação do *rating*.

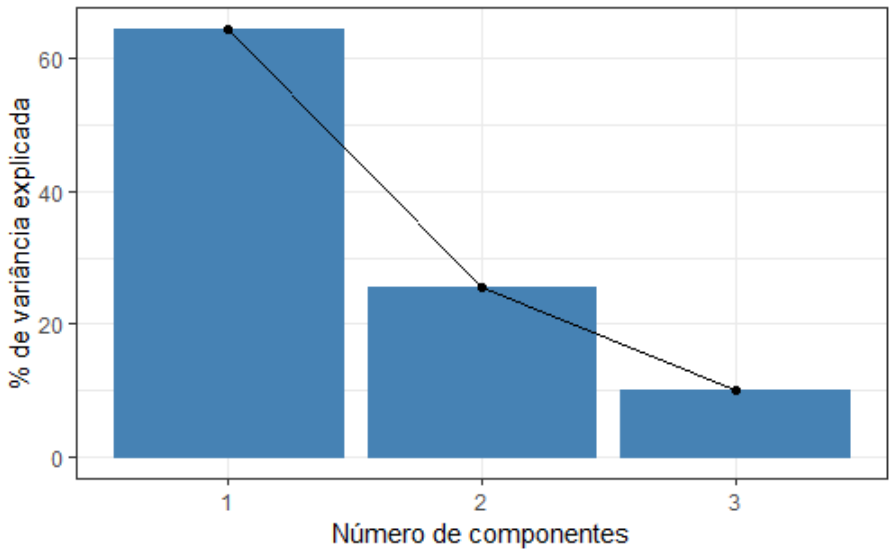


Figura 4.23: Gráfico de cotovelo para os componente principais: goleiro

Na Tabela 4.34 conseguimos verificar o peso de cada variável no componente. O primeiro componente explica praticamente todas as variáveis como se fosse uma média geral das variáveis porém com peso menor para defesa de pênalti que é explicada pelo segundo.

Podemos verificar como os jogadores se distribuem nos componentes através da Figura 4.24 com destaque para os jogadores Vanderlei no primeiro componente e Gatito Fernandez no segundo.

Tabela 4.34: Autovetores e pesos dos componentes: goleiro

Variáveis	CP1	CP2
TDP	0.666	0.731
TSG	0.903	-0.119
PDEF	0.820	-0.463

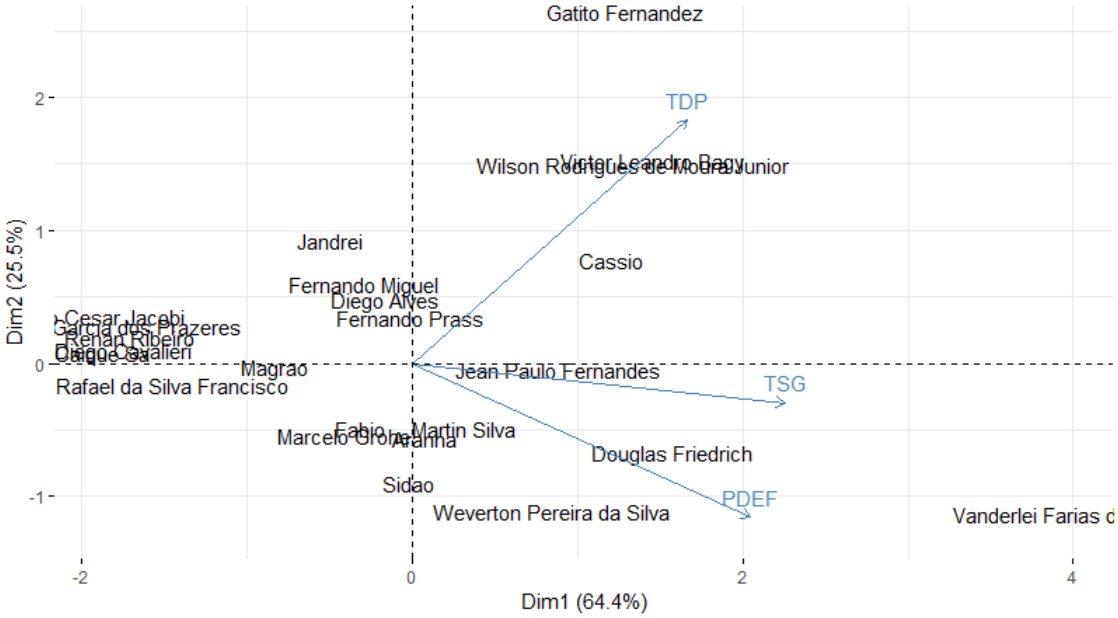


Figura 4.24: Biplot dos goleiros com relação aos dois componentes analisados

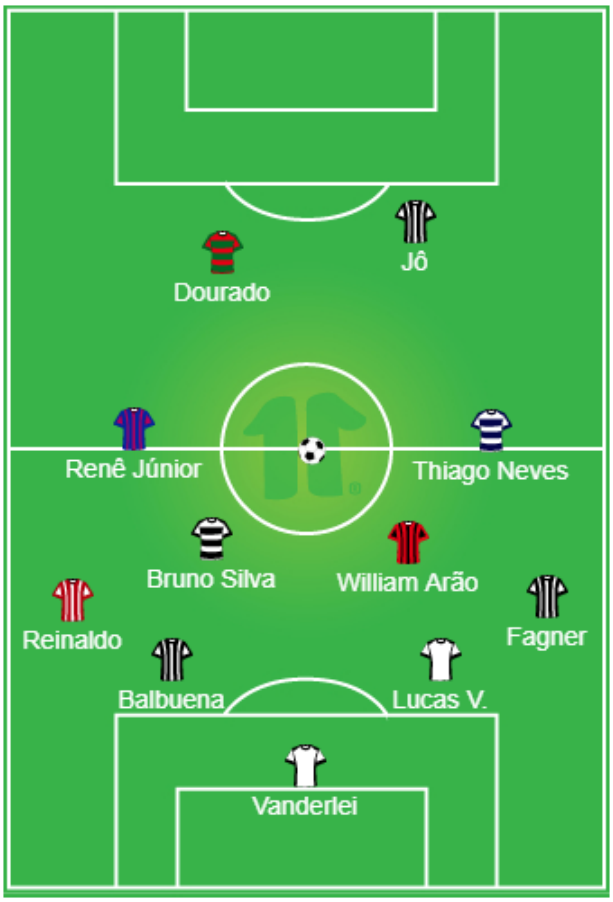
Usando a variância de cada componente como seu respectivo peso, uma média ponderada foi calculada entre os dois para geração do *rating*. A Tabela 4.35 nos mostra os 5 melhores goleiros de acordo com nosso indicador.

Tabela 4.35: *Rating* e informações dos goleiros

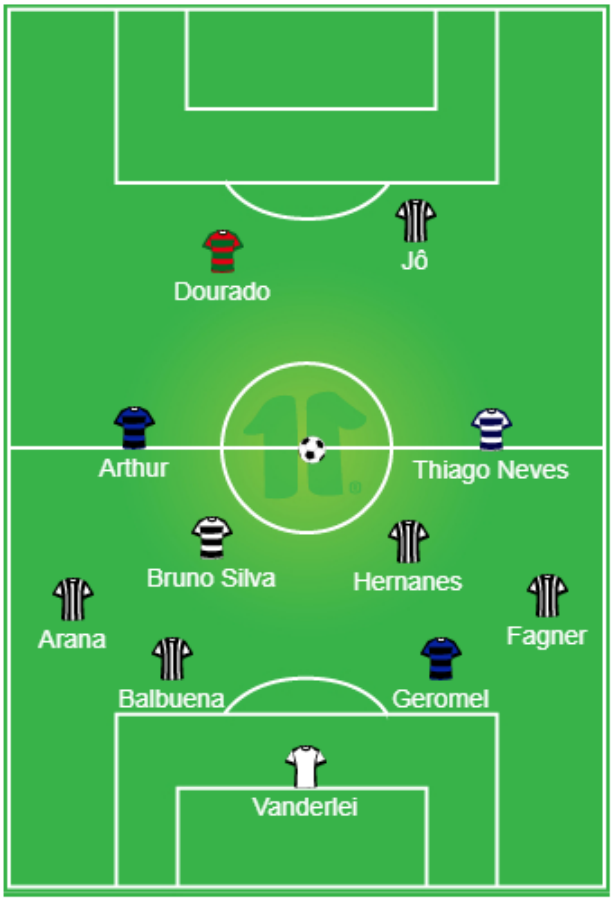
Jogadores	TDP	TSG	PDEF	*Rating*
Vanderlei	2	13	2.826	10.000
Gatito	4	7	1.172	7.809
Victor	3	8	1.454	7.301
Wilson	3	7	1.512	7.052
Cassio	2	9	1.428	6.278

4.4 Comparação dos resultados

Nesse capítulo será comparado a escalação da seleção do campeonato brasileiro 2017 de acordo com os indicadores criados nesse estudo Figura (a) e a seleção eleita por jornalistas e especialistas do futebol Figura (b), levando em consideração o esquema tático 4-4-2 para as duas escalações ficarem o mais comparável possível. As pranchetas de esquemas táticos foram criadas afim de facilitar a visualização [8]. Podemos notar algumas diferenças na escalação.



(a) Seleção sugerida 4-4-2



(b) Seleção Jornalistas 4-4-2

A primeira delas é em relação ao segundo volante na qual os jornalistas e profissionais elegeram o jogador Arthur do grêmio e nosso indicador selecionou Renê Júnior do Bahia. Renê Júnior foi destaque do time do Bahia e contratado pelo Corinthians, campeão de 2017, como reforço para a temporada de 2018 [21].

A Tabela 4.36 mostra que Renê Junior teve números superiores ao de Arthur em quase todos os indicadores que consideramos importantes para essa posição com exceção ao número de passes para oportunidade de gol. Observamos também desempenho abaixo

da média apenas no número total de assistência.

Tabela 4.36: Comparação dos segundos volantes

Jogadores	TASS	POG	DRC	EINT	EDES	EDAV	ETRB	GOL
Renê Junior	0.000	17.000	26.000	1.229	2.479	1.104	1.958	4.000
Arthur Henrique	0.000	22.000	16.000	0.805	1.388	0.250	1.361	1.000

A segunda diferença é com relação ao primeiro volante na qual Hernanes do São Paulo foi eleito pelos jornalistas e profissionais e nosso *rating* elegeu William Arão, um dos maiores destaques do Flamengo na temporada de 2017 com números expressivos nas variáveis analisadas, sendo todos acima da média e um dos responsáveis por levar o time para a taça Libertadores da América de 2018.

Na Tabela 4.37 os números de William Arão são superiores ao de Hernanes em todas as váriaveis analisadas e além disso estão muito acima da média dos jogadores dessa posição.

Tabela 4.37: Comparação dos primeiros volantes

Jogadores	EINT	EDES	EDAV	EREB	EBLO	ETRB
Willian Arão	1.157	1.526	1.157	0.789	0.289	1.631
Hernanes	0.306	0.632	0.408	0.387	0.102	0.591

A terceira mudança é na lateral esquerda com Reinaldo no lugar de Guilherme Arana. Com números expressivos na Chapecoense, Reinaldo passou de desacredito no São Paulo à essencial no time da Chapeconse sendo apelidado pelos torcedores de "Kingnaldo". Com um forte apoio no ataque e variáveis acima da média em todas as características analisadas nesse estudo, Reinaldo ajudou a classificar a Chepecoense para a taça Libertadores da América. Seu desempenho o fez voltar ao São Paulo e ser titular absoluto na temporada 2018 [2].

Podemos observar na Tabela 4.38 que os números de Reinaldo superam os de Guilherme Arana em totais de roubada de bola, total de assistência, passe para oportunidade de gol e cruzamentos precisos, que na maioria são variáveis de auxílio ao ataque. Guilherme Arana tem números superiores em total de jogos sem sofrer gols e valores muito próximos em eficiência na interceptação e eficiência no desarme.

Tabela 4.38: Comparação dos laterais

Jogadores	TRB	TSG	TASS	POG	EINT	EDES	CRP
Reinaldo	50.000	8.000	8.000	67.000	0.857	1.265	51.000
Guilherme Arana	18.000	15.000	3.500	28.000	1.066	1.266	21.000

A última diferença é com relação aos zagueiros na qual o indicador criado elegeu Lucas Veríssimo como melhor ao invés de Geromel do Grêmio. Lucas Veríssimo fez parte da segunda zaga menos vazada do campeonato com apenas 32 gols sofridos e ajudou o Santos a chegar na terceira posição no campeonato com classificação direta para taça Libertadores da América de 2018 [12].

A Tabela 4.39 nos mostra que os números de Lucas Veríssimo são superiores ao de Geromel em quase todas as váriaveis com excessão a eficiência na roubada de bola, onde o desempenho dos dois jogadores estão bem próximos.

Tabela 4.39: Comparação dos zagueiros

Jogadores	TSG	EINT	EDES	EREB	EDAV	EBLO	ETRB	PPC
Lucas Verissimo	15.000	2.031	1.156	9.187	3.218	1.187	1.125	88.454
Geromel	9.000	1.750	0.972	2.527	1.111	0.388	1.277	81.861

Na Figura 4.25 temos uma sugestão de seleção com outra formação no formato 4-3-3 que inclui os pontas, dessa maneira, temos algumas alterações na escalação. O atacante Jô deixa de fazer parte da seleção, com apenas Henrique Dourado de centroavante e William Arão também deixa o time, ficando apenas com Bruno Silva de primeiro volante.

Para as pontas, segundo nosso indicador, o primeiro jogador eleito é Dudu do Palmeiras, principal responsável por garantir o segundo lugar do time no campeonato brasileiro. Em 2017, foi o ano que ele obteve os melhores números com a camisa do clube igualando a temporada de 2015, ano em que ele disputou menos jogos [4].

O outro ponta é Bruno Henrique do Santos, que ajudou o clube a chegar na terceira colocação do campeonato brasileiro de 2017. Ao final da temporada, o atacante recebeu diversas sondagens do futebol Chinês. Com números expressivos nos nossos indicadores principalmente em gols, assistências, dribles desconcertantes e boa finalização [7].

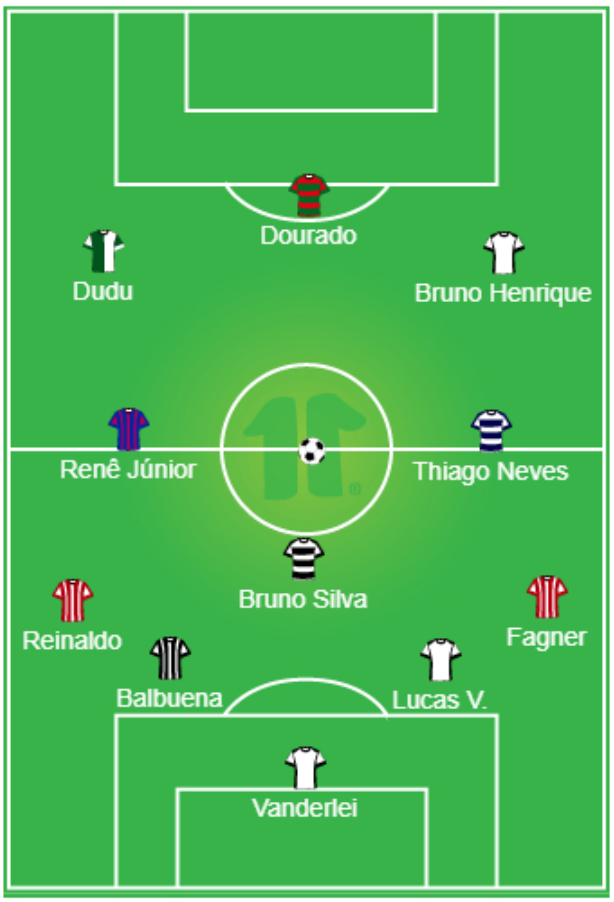


Figura 4.25: Seleção sugerida 4-3-3.

Capítulo 5

Considerações finais

O presente estudo permitiu identificar as variáveis mais correlacionadas com cada posição em campo sendo elas segmentadas em centroavante, ponta, meia ofensivo, segundo volante, primeiro volante, lateral, zagueiro e goleiro.

Uma escoragem foi criada para cada posição com base em suas variáveis de modo a identificar os melhores jogadores por posição. Também foi sugerida duas escalações com sistemas táticos diferentes (4-4-2 e 4-3-3) a fim de abrangir a maior diversidade de posições no campo.

Para a formação 4-4-2, os jogadores escolhidos para centroavante foram: Jô (Corinthians) e Dourado (Fluminense), para segundo volante: Renê Júnior (Bahia), como meia ofensivo: Thiago Neves (Cruzeiro), como primeiro volante: Bruno Silva (Botafogo) e William Arão (Flamengo), nas laterais: Reinaldo (São Paulo) e Fagner (Corinthians), os zagueiros foram: Balbuena (Corinthians) e Lucas Veríssimo (Santos) e no gol: Vanderlei (Santos).

Para a formação 4-3-3 os escolhidos foram para centroavante: Dourado (Fluminense), nas pontas: Dudu (Palmeiras) e Bruno Henrique (Santos), como segundo volante: Renê Júnior (Bahia), para meia ofensivo: Thiago Neves (Cruzeiro), para primeiro volante: Bruno Silva (Botafogo), nas laterais: Reinaldo (São Paulo) e Fagner (Corinthians), os zagueiros: Balbuena (Corinthians) e Lucas Veríssimo (Santos) e no gol: Vanderlei do Santos).

Após a criação do *rating* e escolha dos melhores jogadores comparou-se o resultado com a seleção eleita por jornalistas e profissionais do esporte. Com isso, verificamos quatro jogadores diferentes nas escalações (Renê Júnior, William Arão, Reinaldo e Lucas veríssimo) e sete indicados iguais. Dessa forma, o *rating* criado pode auxiliar nas indicações dos me-

lhores jogadores do campeonato com base nos números de seu desempenho ao longo de toda temporada.

Recomenda-se para trabalhos futuros considerar outras variáveis explicativas, de forma a criar indicadores de efetividade mais precisos daqueles apresentados nesse estudo.

Além disso, sugere-se a utilização de mais componentes e outra forma para o cálculo do *rating*, por exemplo, utilizando a distância entre os valores dos jogadores nos componentes até a sua média (origem dos eixos no sistema cartesiano).

Apêndice A

Códigos *WebScraping*

```
########################################.#
####            Dados G1              ####
########################################.#

#install.packages("rvest")
library(rvest)
library(dplyr)
library(tidyr)

## Extraindo dados de um site através de uma tabela
tabela <- read_html("http://g1.globo.com/politica/noticia/veja
como-deputados-votaram-no-impeachment-de-dilma-na-pec-241-na-r
forma-trabalhista-e-na-denuncia-contra-temer.ghtml")
tab_votos <- tabela %>%
  html_node("table") %>%
  html_table(header=TRUE) %>%

##Mudar os nomes das variaveis
  setNames(c('deputados',
             'impdilma',
             'pectegas',
             'reftrab',
```

```
             'rejdentemer')) %>%
  tbl_df()

getwd()
write.csv(tab_votos,"tab_votos.csv")

voto_sim <- rep(NA,nrow(tab_votos))
for(i in 1:nrow(tab_votos)){
  voto_sim[i] <- ifelse(tab_votos$rejdentemer[i]=="SIM",
                        tab_votos$deputados[i],NA)
}

voto_sim <- data.frame(voto_sim)

## Comando para excluir os NA do dataframe
voto_sim <- voto_sim %>% drop_na()
View(voto_sim)
write.csv(voto_sim,"voto_sim.csv")
```

Referências Bibliográficas

[1] Analytics for everyone. https://www.kinanalytics.com. Accessed: 2018-03-18.

[2] Após sair do são paulo, ele virou 'kingnaldo' e desperta cobiça do corinthians: 'ano muito especial'. http://www.espn.com.br/noticia/741427_apos-sair-do-sao-paulo-ele-virou-kingnaldo-e-desperta-cobica-\ \do-corinthians-ano-muito-especial. Accessed: 2018-11-11.

[3] Big data é o novo petróleo, afirma executiva da ibm. https://olhardigital.com.br/noticia/big-data-e-onovo-petroleo,-afirma-executiva-da-ibm/34986. Accessed: 2018-03-18.

[4] Dudu joga para terminar 2017 com os melhores números da carreira. https://www.lance.com.br/palmeiras/dudu-joga-par-terminar-2017-com-melhores-numeros-carreira.html. Accessed: 2018-11-11.

[5] Estatística no esporte: Moneyball - o homem que mudou o jogo. http://www.estatisti.co/2013/02/estatistica-no-eporte-moneyball-o.html. Accessed: 2018-03-18.

[6] Estatísticas do campeonato brasileiro. https://www.whoscored.com. Accessed: 2018-03-20.

[7] Exclusivo! dono de números invejáveis em 2017, bruno henrique valoriza identificação com santos e espera títulos em 2018. https://www.goal.com/br/not%C3%ADcias/exclusivo-ono-de-numeros-invejaveis-em-2017-bruno-henrique/jpr5lqkgsgc1aomw0bak2abn. Accessed: 2018-11-11.

[8] Ferramentas de esquemas táticos. https://this11.com/#!/editor/. Accessed: 2018-11-11.

[9] Google chrome. https://www.google.com/chrome/. Accessed: 2018-11-28.

[10] Json for r. https://cran.r-project.org/web/packages/rjson/index.html. Accessed: 2018-11-28.

[11] Linkedin ordered to allow scraping of public profile data. https://www.infoq.com/news/2017/08/linkedin-rulig-scraping. Accessed: 2018-05-28.

[12] O ano da redenção: Lucas veríssimo vai de última opção a intocável no santos. https://globoesporte.globo.com/sp/santos-e-regiao/futebol/times/santos/noticia/o-ano-da-redencao-lucas-verissimo-vai-de-ultima-opcao-a-intocavel-no-santos.ghtml. Accessed: 2018-11-11.

[13] O homem que quer mudar o jogo. https://epocanegocios.globo.com/Informacao/Visaonoticia/2015/05/o-homem-que-quer-mudar-o-jogo.html. Accessed: 2018-03-18.

[14] O que é webscraping? http://blogbrasil.westcon.com/o-que-e-web-scrapig. Accessed: 2018-03-18.

[15] Open source and enterprise-ready professional software for r. https://www.rstudio.com/. Accessed: 2018-11-11.

[16] Os limites da garimpagem de dados na internet. http://observatoriodaimprensa.com.br/etica-jornaistica/os-limites-da-garimpagem-de-dados-na-internet/ , note = Accessed: 2018-03-18.

[17] Package rvest. https://cran.r-project.org/web/packages/rvest/rvest.pdf. Accessed: 2018-11-28.

[18] Parse xml. https://cran.r-project.org/web/packages/xml2/index.html. Accessed: 2018-11-28.

[19] Projeto pedagógico do curso de bacharelado em estatística - ufscar. http://www.ufscar.br/~des/Cat.htm. Accessed: 2018-03-18.

[20] R bindings for selenium webdriver. https://cran.r-project.org/web/packages/RSelenium/index.html. Accessed: 2018-11-28.

[21] Renê júnior, tréllez, andré... os destaques dos times do ne que agitam o mercado. https://globoesporte.globo.com/futebol/central-d-mercado/noticia/rene-junior-trellez-andre-os-destaque-dos-clubes-do-ne-que-agitam-o-mercado.ghtml. Accessed: 2018-11-11.

[22] A robust, high performance json parser and generator for r. https://cran.r-project.org/web/packages/jsonlite/index.html. Accessed: 2018-11-28.

[23] Software, algoritmos e inspiração no beisebol: a empresa que encanta roger e chegou ao palmeiras. http://www.espn.com.br/futebol/artigo/_/id/397710/software-algoritmos-e-inspiracao-no-beisebol-a-empresa-que-encanta\-roger-e-chegou-ao-palmeiras. Accessed: 2018-03-18.

[24] Tools for working with urls and http. https://cran.r-project.org/web/packages/httr/index.html. Accessed: 2018-11-28.

[25] Votação da rejeição da denúncia contra temer. http://especiais.g1.globo.com/politica/2017/votacao-da-denuncia-contra-temer-na-camara/. Accessed: 2018-11-28.

[26] HONGYU, K.; SANDANIELO V. L. M.; DE OLIVEIRA, G. J. J. Análise de Componentes Principais: Resumo Teórico, Aplicação e Interpretação. *ES Engineering and Science* (2016), 83–90.

[27] HOTELLING, H. *Analysis of a Complex of Statistical Variables Into Principal Components*, vol. 24. Journal of Educational Psychology, 1933.

[28] JOHNSON, R. A.; WICHERN, D. W. *Applied Multivariate Statistical Analysis*, 6nd ed. Pearson, New Jersey, 2007.

[29] MITCHELL, R. *Web Scraping with Python Collecting Data from the Modern Web*. O'Reilly Media, United States of America, 2015.

[30] PEARSON, K. *On lines and planes of closest fit to systems of points in space, Philosophical Magazine*, 6nd ed., vol. 2. 1901.

[31] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2016.

[32] REGAZZI, A. Análise multivariada, notas de aula INF 766, Departamento de Informática da Universidade Federal de Viçosa.