Universidade de Brasília
Departamento de Estatística

Análise Multivariada por fatores determinantes do desempenho de jogadores do Brasileirão Série A - 2022

Francisco Iago dos Reis Ferreira

Projeto apresentado para o Departamento de Estatística da Universidade de Brasília como parte dos requisitos necessários para obtenção do grau de Bacharel em Estatística.

Brasília
2023

Francisco Iago dos Reis Ferreira

Análise Multivariada por fatores determinantes do desempenho de jogadores do Brasileirão Série A - 2022

Orientador: Prof. Gladston Luiz da Silva

Projeto apresentado para o Departamento de Estatística da Universidade de Brasília como parte dos requisitos necessários para obtenção do grau de Bacharel em Estatística.

Brasília
2023

Agradecimentos

Foi uma longa jornada até aqui, agradeço a todas as pessoas que contribuíram para a realização desse trabalho.

À minha mãe, a Sra. Francislene Maria dos Reis, por ter batalhado para que minha irmã e eu pudessemos ter a melhor educação possível. Todo o seu esforço não foi em vão.

À minha companheira, Evellyn Cristina Lopes Guerbe, cujo o amor e apoio incondicional foram cruciais para superar os momentos de indecisão.

Aos vários amigos que fiz durante toda a graduação, seja no curso de estatística, seja na Casa do Estudante, tenho muita consideração por cada um de vocês.

À meu orientados, Prof. Gladston Luis da Silva, pela empolgação em orientar esse trabalho, sua orientação e conselhos foram fundamentais para aprimorar minha pesquisa e direcionar meus esforços na direção certa.

E por último, à todos os professores do Departamento de Estatística e à Universidade de Brasília, por proporcionarem o ambiente de aprendizado e os recursos necessários para a concretização deste projeto.

Resumo

O Campeonato Brasileiro, também conhecido como Brasileirão, é a principal competição de futebol profissional no Brasil. Ele é disputado todos os anos por clubes que representam diversos estados de todo o país. Adotado em 2003, o formato de pontos corridos é disputado por 20 clubes que se enfrentam em jogos de ida e volta durante 38 rodadas, a equipe que fizer a maior pontuação se sagra como a campeã daquela edição. A competição disputada em 2022 nos forneceu uma vasta quantidade de dados acerca dos jogadores que dela participaram, e tais dados serão base para uma Análise Multivariada detalhada. Esta análise nos permitirá explorar as complexas interações entre as variáveis de desempenho.

Palavras-chaves: Brasileirão, Desempenho, Correlação, Análise Fatorial, Análise de Componentes Principais.

Abstract

The Campeonato Brasileiro, also known as the Brasileirão, is the main professional soccer competition in Brazil. It is contested every year by clubs representing various states across the country. Adopted in 2003, the straight points format is contested by 20 clubs who face each other in back-to-back matches over 38 rounds, with the team that scores the most points becoming the champion of that edition. The 2022 competition provided us with a vast amount of data about the players who took part in it, and this data will be the basis for a detailed Multivariate Analysis. This analysis will allow us to explore the complex interactions between the performance variables.

Keywords: Brasileirão, Performance, Correlation, Factor Analysis, Principal Component Analysis.

Lista de Tabelas

1 Descrição das variáveis dos goleiros 25

2 Descrição das variáveis de Chute 26

3 Descrição das variáveis de Passe 27

4 Descrição das variáveis de Criação de finalizações e gols 28

5 Descrição das variáveis de Ações defensivas 28

6 Descrição das variáveis de Posse 29

7 Descrição das variáveis Tempo de jogo 30

8 Descrição das variáveis Diversas 31

Lista de Figuras

1 Metodologia CRISPY DM 11

2 Tabela após a 38ª Rodada 20

Sumário

1 Introdução 8

2 Objetivos 10

2.1 Objetivo Geral 10

2.2 Objetivos Específicos 10

3 Metodologia 11

3.1 CRISPY DM 11

4 Referencial Teórico 13

4.1 Análise Estatística Descritiva 13

4.2 Análise Multivariada 13

4.3 Análise de Componentes Principais 13

4.3.1 Componentes Principais Populacionais 14

4.4 Análise Fatorial 17

4.4.1 Modelo Fatorial 17

4.4.2 Estrutura da Covariância 19

4.4.3 Metodo de Estimação 19

5 Resultados 20

5.1 Campeonato Brasileiro de Futebol de 2022 - Série A 20

5.2 Banco de Dados 21

5.2.1 Divisão por posição 21

5.3 Goleiros 22

5.4 Defensores 22

5.5 Meio Campistas 22

5.6 Atacantes 22

6 Conclusão 24

7 Anexo 25

7.1 Variáveis específicas dos goleiros 25

7.2 Variáveis para o jogadores de linha 26
7.2.1 Chute 26
7.2.2 Passe 27
7.2.3 Criação de finalizações e gols 28
7.2.4 Ações defensivas 28
7.2.5 Posse 29
7.2.6 Tempo de jogo 30
7.2.7 Diversas 31

1 Introdução

No ano de 2002, Billy Beane, diretor geral da equipe de beisebol *Oakland Athtetics*, desafiou tudo que se acreditava sobre o esporte competitivo até então, ao adotar técnicas estatísticas avançadas para avaliar o desempenho de jogadores. Contrapondo-se ao critérios que eram utilizados pelos olheiros à epoca, como capacidade de correr, de arremessar, receber a bola, de rebater e rebater com potência. Billy e Paul dePodesta (até então um jovem executivo do clube), foram capazes de concluir que nem todas as características importantes num jogador de beisebol tem a mesma importância. Que velocidade na corrida, habilidade na defesa e força bruta tendiam a ser características demasiadamente superestimadas.

Eles foram capazes de mostrar que, características como porcentagem de rebatidas, porcentagem de base alcançadas, habilidade de controle da base de strike e etc. Essa abordagem revelou informações valorosas sobre a eficácia de um jogador, permitindo com que talentos muitas vezes subvalorizados por outras equipes fossem identificados.

Com um orçamento limitado, sendo o time com a terceira menor folha salarial da liga na época, mas com uma abordagem estatística autêncica, o time do Oakland Athletics quebrou o recorde de vitórias consecutivas, de uma temporada (LEWIS, 2004). O impacto dessa abordagem abriu caminho para uma nova era no esporte, extrapolou o beisebol e inspirou diversas modalidades. A utilização de técnicas estatísticas promoveu uma mudança na maneira com que o desempenho dos jogadores seriam avaliados e compreendidos.

No âmbito do futebol, é imperativo considerar o avanço relativo à informação contida nos dados. De acordo com Anderson e Sally (2013), os números têm o poder de desafiar conceitos pré-estabelecidos e desmantelar crenças antigas no esporte. Eles oferecem um entendimento do jogo como nunca visto antes, possibilitando uma visão contraposta aos pressupostos tradicionais.

Um exemplo de como os dados podem influenciar na tomada de decisão, é a renovação de contrato feita pelo jogador do Manchester City, Kevin de Bruyne[1]. Na época o jogador conseguiu uma extensão salarial na casa dos 80 milhões de euros, utilizando apenas análise de dados ao invés de um agente[2].

No Brasil, o Brasileirão Série A é uma competição que movimenta milhões de

[1] ⟨https://www.transfermarkt.com.br/kevin-de-bruyne/profil/spieler/88755⟩

[2] ⟨https://www.uol.com.br/esporte/futebol/ultimas-noticias/2021/04/08/big-data-e-sem-agente-como-de-bruyne-renovou-com-o-city.htm⟩

reais, instiga torcedores, investidores e apostadores todos os anos. Com o crescente avanço da tecnologia, o uso de técnicas estatísticas tem sido de imensa importância para se avaliar e entender a performance dos atletas durante a competição. No cenário nacional, por exemplo, times como o Flamengo[3] e Palmeiras[4] utilizam análise de desempenho para tomada de decisões administravias e esportivas.

O presente trabalho, visa analisar, por meio de técnicas de Análise Multivariada, o desempenho de atletas da primeira divisão do Campeonato Brasileiro Série A 2022, pois permite uma avaliação mais precisa e completa, além de identificar padrões e tendências que não seriam encontradas por meio de análises mais simples. Por meio dessa análise, espera-se identificar as características que mais contribuíram para o desempenho dos jogadores durante o campeonato.

Tal abordagem àplicada ao contexto do campeonato brasileiro pode ser útil para os treinadores, apostadores, analistas de desempenho, investidores e patrocinadores, além de entusiastas por futebol, que poderão tomar melhores decisões no âmbito esportivo utilizando-se de informações munidas de análise estatística.

[3] ⟨https://www.flamengo.com.br/noticias/futebol/flamengo-e-o-primeiro-clube-da-america-do-sul-a-implementar-o
[4] ⟨https://www.palmeiras.com.br/noticias/palmeiras-renova-parceria-com-empresa-de-software-de-analise-de-desem
⟩

2 Objetivos

2.1 Objetivo Geral

Utilizar técnicas de Análise mMultivariada para identificar os principais fatores que influenciam o desempenho de atletas de futebol que jogaram o Campeonato Brasileiro da série A em 2022, considerando as posições em que esses jogadores atuaram.

2.2 Objetivos Específicos

Para alcançar o objetivo geral, será necessário realizar os seguintes passos:

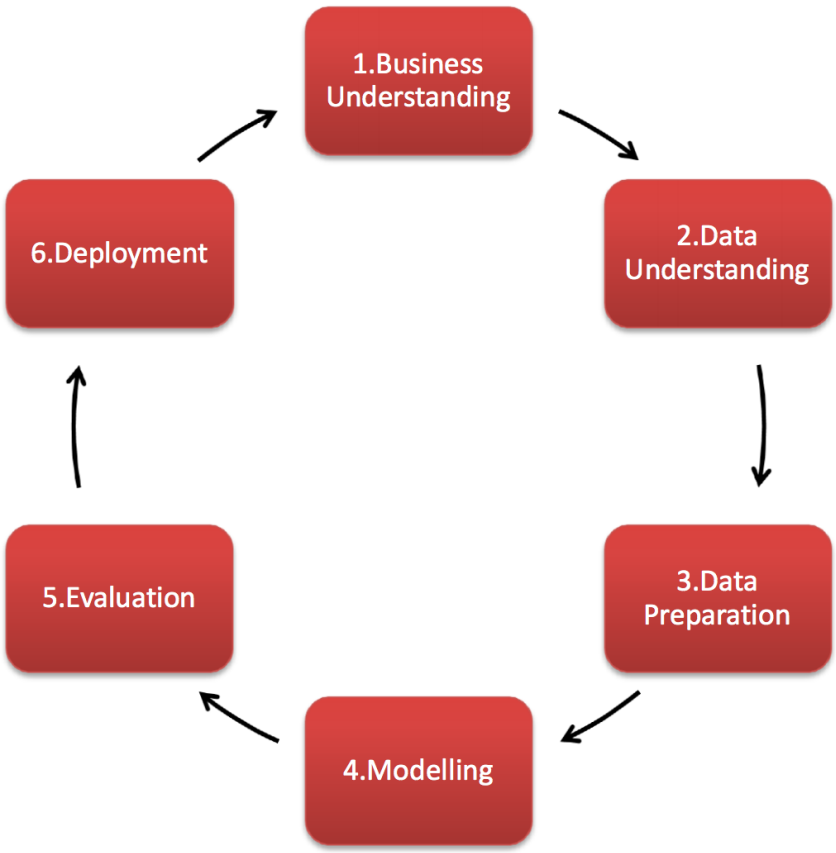
- Coletar e preparar os dados de desempenho dos jogadores que participaram do campeonato;
- Relizar uma análise exploratória dos dados para identificar possíveis valores ausentes e inconsistências;
- Aplicar técnicas de redução de análise de dimensionalidade para selecionar as variáveis mais relevantes para a análise;
- Utilizar as técnicas de Análise Multivariada, para identificar padrões de desempenho entre os jogadores;
- Validar os resultados obtidos;
- Interpretar os resultados obtidos.

3 Metodologia

3.1 CRISPY DM

Nessa fase, serão definidos os métodos para a elaboração do trabalho, que serão baseadas na metodologia CRISP-DM. Esse método irá permitir que a análise seja padronizada em etapas, def modo que o trabalho seja estruturado de forma eficiente.

Figura 1: Metodologia CRISPY DM



Fonte: Elaborada pelo Autor

A partir da imagem acima, podemos organizar o trabalho da seguinte forma:

- **Fase 1 - Entendimento do problema:** Na etapa inicial, será feira uma análise preliminar dos dados para entender o problema. Também será feita uma revisão de literatura para avaliar qual o métoodo mais adequado para a implementação no conjunto de dados.
- **Fase 2 - Entendimento dos dados:** Nesta estapa, será realizada uma análise

exploratória dos dados para entender o seu formato e a qualidade. Serão realizadas técnicas exploratórias para avaliar a distribução e analisar a correlação entre as variáveis.

- **Fase 3 - Preparação dos dados:** Após a etapaa anterior, os dados serão preparados para a análise, isso inclui a seleção das variáveis mais relevantes, a transformação de algumas variáveis, e se necessário, a criação de novas variáveis.

- **Fase 4 - Modelagem:** Será realizada a análise dos dados. Serão utilizados métodos de Análise Multivariada. Além de realizarmos testes para validação dos resultados obtidos.

- **Fase 5 - Avaliação:** Os resultados serão avaliados em relação aos objetivos destre trabalho. Caso necessãrio, serão realizadas análises adicionais para obter resultados mais pertimentes.

- **Fase 6 - Conclusão:** Finalmente, os resultados obtidos durante todo o estudo serãi apresentados.

4 Referencial Teórico

Neste seção, serão descritos os métodos de Análise Estatística que serão empregados na pesquisa. Tais métodos foram selecionados devido à sua relevância dentro do contexto dos dados coletados e contribuirão para alcançar os objetivos descritos na seção anterior.

4.1 Análise Estatística Descritiva

É uma técnica muito utilizada para descrecver as características esscenciais de um conjunto de dados. Nos permite examinar os dados de forma quantitativa, como média, mediana, quartis, gráficos e etc. A intenção de utilizar essa análise é fornecer uma visão resumida das variáveis presentes, com o intuito de obter um entendimento dos dados, estabelecendo um base sólida para a etapa subsequente (MORETTIN; BUSSAB, 2017).

4.2 Análise Multivariada

A Análise Multivariada é um conjunto de métodos estatísticos usados quando múltiplas variáveis são medidas simultaneamente em cada elemento amostral. Geralmente, são variáveis que estão correlacionadas entre si, e quanto maior o número , mais complexa se torna a análise. O objetivo é simplificar a interpretação do fenômeno estudado. Ela pode ser dividida em duas categorias principais: técnicas exploratórias que buscam simplificar a estrutura de variabilidade dos dados, sintetizando as variáveis, e técnicas de inferência. Um dos métodos comuns de simplificação da estrutura de variabilidade é a redução de dimensionalidade, que visa diminuir o número de variáveis originais mantendo as informações essenciais para a análise.

4.3 Análise de Componentes Principais

Segundo (JOLLIFFE, 2002), a essência da técnica de Análise de Componentes Principais é a redução da complexidade de um conjunto de dados que contém inúmeras variáveis(p). Essa redução busca preservar ao máximo a variabilidade dos dados originais. Esse tipo de simplificação pode ser alcançada por meio da transformação dos dados em um novo conjundo de variáveis chamadas de componentes principais(k). Tais componentes não possuem uma correlação entre si e são organizados de modo que as primeiras

componentes possam reter a maior parte da variação presente no conjunto original dos dados.

O desenvolvimento matemático expresso na sessão 4.3.1 foi extraído da obra de (JOHNSON; WICHERN et al., 2002).

4.3.1 Componentes Principais Populacionais

Esse método tem por objetivo principal explicar a estrutura da variância/covariância dos dados através de combinações lineares das variáveis, onde as p variáveis são retidas em k componentes. Podemos escrever esses componentes como combinações lineares das p variáveis $X_1, X_2, ..., X_p$. Geometricamente, essas combinações representam a seleção de um novo sistema de coordenadas obtido por meio da rotação do istema original com $X_1, X_2, ..., X_p$ como o eixo de coordenadas, tais eixos representam as direções com a variabilidade máxima e nos fornecem uma descrição mais parcimoniosa da estrutura de covariância.

Seja o vetor aleatório **X'** $= [X_1, X_2, ..., X_p]$ com matrix de covariância Σ com autovetores próprios $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p = 0$

Considere as combinações lineares:

$$Y_1 = a_1'X = a_{11}X_1 + a_{12}X_2 + ... + a_{1p}X_p \tag{4.3.1}$$
$$Y_2 = a_2'X = a_{21}X_1 + a_{22}X_2 + ... + a_{2p}X_p \tag{4.3.2}$$
$$\vdots \tag{4.3.3}$$
$$Y_p = a_p'X = a_{p1}X_1 + a_{p2}X_2 + ... + a_{pp}X_p \tag{4.3.4}$$

que pode ser escrita como

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_p \end{bmatrix}_{p\times 1} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1p} \\ a_{21} & a_{22} & \cdots & a_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ a_{p1} & a_{p2} & \cdots & a_{pp} \end{bmatrix}_{p\times p} \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{bmatrix}_{p\times 1}$$

Onde

$$E(Y_i) = E(a_i'\underline{X}) \tag{4.3.5}$$
$$E(Y_i) = a_i' E(\underline{X}) \tag{4.3.6}$$
$$E(Y_i) = a_i'\underline{\mu} \tag{4.3.7}$$

$$Var(Y_i) = V(a_i'\underline{X}) \tag{4.3.8}$$
$$Var(Y_i) = a_i' V(\underline{X}) a_i \tag{4.3.9}$$
$$Var(Y_i) = a_i'\Sigma a_i, \quad i = 1, 2, ..., p \tag{4.3.10}$$

e

$$Cov(Y_i, Y_k) = V(a_i'\underline{X}, a_k'\underline{X}) \tag{4.3.11}$$
$$Cov(Y_i, Y_k) = a_i'\Sigma a_k, \quad i, k = 1, 2, ..., p \tag{4.3.12}$$

São combinações lineares não correlacionadas de $Y_1, Y_2, ..., Y_p$ no qual a variância é máxima. Onde, podemos definir a partir disso:

- A 1ª componente principal será a combinação linear de $a_1'X$ que maximizará a $V(a_1'X)$ sujeita à restrição $a_1'a_1 = 1$
- A 2ª componente principal será a combinação linear de $a_2'X$ que maximizará a $V(a_2'X)$ sujeita à restrição $a_2'a_2 = 1$ e $Cov(a_1'X, a_2'X) = 0$

$$\vdots$$

- A i-ésima ª componente principal será a combinação linear de $a_i'X$ que maximizará a $V(a_i'X)$ sujeita à restrição $a_i'a_k = 1$ e $Cov(a_i'X, a_k'X) = 0$ para $k < i$

Seja $\boldsymbol{\Sigma}$ a matriz de covariância associada ao vetor aleatório $\boldsymbol{X' = [X_1, X_2, ..., X_p]}$, com pares de autovalores-autovetores $(\lambda_1, e_1), (\lambda_2, e_2), ..., (\lambda_p, e_p)$, onde $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p \geq 0$, então a i-ésima ª componente principal é dada por

$$Y_i \approx e_i'X = e_i1X + e_i2X + ... + e_ipX \quad i = 1, 2, ..., p \tag{4.3.13}$$

Com

$$Var(Y_i) = e_i'\Sigma e_i = \lambda_i \quad i = 1, 2, ..., p \tag{4.3.14}$$

$$Cov(Y_i, Y_k) = e_i'\Sigma e_i = 0, \quad i \neq k \tag{4.3.15}$$

E seja, $\boldsymbol{X' = [X_1, X_2, ..., X_p]}$ com matriz de covariância Σ, e com pares de autovalores-autovetores $(\lambda_1, e_1), (\lambda_2, e_2), ..., (\lambda_p, e_p)$, onde $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p \geq 0$, tomemos $Y_i \approx e_i'X = e_i1X + e_i2X + ... + e_ipX \quad i = 1, 2, ..., p$, como os componentes principais. Então

$$\sigma_{11} + \sigma_{22} + ... + \sigma_{pp} = \sum_{i=1}^{p} Var(X_i) = \lambda_1 + \lambda_2 + ... + \lambda_p = \sum_{i=1}^{p} Var(Y_i) \tag{4.3.16}$$

Se $Y_1 = e_1'X = Y_2 = e_2'X, ..., Y_p = e_p'X$ são os componentes principais obtidos através da matriz de covariância Σ, então

$$\rho_{Y_iX_k} = \frac{e_{ik}\sqrt{\lambda_i}}{\sqrt{\sigma_{kk}}} \quad i, k = 1, 2, ..., p \tag{4.3.17}$$

são os coeficientes de correlação entre as componentes(Y_i) e as variáveis (X_k). E $(\lambda_1, e_1), (\lambda_2, e_2), ..., (\lambda_p, e_p)$ são os pares de autovalores-autovetores para a matriz Σ. Alternativamente, também é possivel obter a proporção total da k-ésima[a] componente principal por meio da relação:

$$\frac{\lambda_k}{\lambda_1 + \lambda_2 + ... + \lambda_p} \quad i, k = 1, 2, ..., p \tag{4.3.18}$$

4.4 Análise Fatorial

A Análise Fatorial, segundo (HAIR et al., 2009), é uma técnica que fornece as ferramentas necessárias para examinar a estrutura das relações entre as variáveis, identificando o conjunto que apresentam correlações fortes, conhecidas como fatores.

Em outras palavras, essa técnica é utilizada para explorar a extrutura subjacente de um conjunto de dados, com a missão de identificar fatores latentes que explicam a correlação entre as variáveis observadas.

O desenvolvimento matemático expresso na sessão 4.4.1,4.4.2 foi extraído da obra de (JOHNSON; WICHERN et al., 2002).

4.4.1 Modelo Fatorial

O vetor aleatório $\boldsymbol{X}$, com p componentes, tem média $\boldsymbol{\mu}$ e matriz de covariância $\boldsymbol{\Sigma}$ No modelo fatorial $\boldsymbol{X}$, é linearmente dependente em relação à algumas variáveis aleatórias não observáveis F_1, F_2, ..., F_m, chamadas de fatores comuns, e p fontes de adicionais de variação ε_1, ε_2, ..., ε_p, chamados de erros ou de fatores específicos. Em particular o modelo pode ser descrito como:

$$X_1 - \mu_1 = \ell_{11}F_1 + \ell_{12}F_2 + ... + \ell_{1m}F_m + \varepsilon_1 \tag{4.4.1}$$

$$X_2 - \mu_2 = \ell_{21}F_1 + \ell_{22}F_2 + ... + \ell_{2m}F_m + \varepsilon_2 \tag{4.4.2}$$

$$\vdots \tag{4.4.3}$$

$$X_p - \mu_p = \ell_{p1}F_1 + \ell_{p2}F_2 + ... + \ell_{pm}F_m + \varepsilon_p \tag{4.4.4}$$

Que pode ser expressa em notação matricial

$$(X - \mu)_{p\times 1} = L_{p\times m}F_{m\times 1} + \varepsilon_{p\times 1} \tag{4.4.5}$$

- μ média da variável i
- ℓ_{ij} são as cargas fatoriais da i-ésima variável no j-ésimo fator;
- L é a matriz das cargas fatoriais;
- F_j são as variáveis latentes; e

- ε_i são os erros não observáveis

Ainda do modelo geral, assume-se que:

- $E(F) = 0_{m\times 1}$;
- $Cov(F) = E(FF') = I_{m\times m}$
- $E(\varepsilon) = 0_{p\times 1}$
- $Cov(\varepsilon) = \Psi_{p\times p}$

Onde, as seguintes condições sao satisfeitas:

- $\boldsymbol{F}$ and ϵ são indepentendes
- E($\boldsymbol{F}$) = 0, Cov($\boldsymbol{F}$) = $\boldsymbol{I}$
- E($\boldsymbol{\epsilon}$) = 0, Cov(ϵ) = Ψ, tal que Ψ é a matriz ortogonal

Ainda no modelo fatorial ortogonal, a covariância pode ser representada como:

$$\begin{aligned}(X-\mu)(X-\mu)' &= (LF+\epsilon)(LF+\epsilon)\\ &= (LF+\epsilon)((LF)'+\epsilon')\\ &= LF(LF') + \epsilon(LF)' + LF\epsilon' + \epsilon\epsilon'\end{aligned}$$

Então, aplicando a esperança na expressão, temos

$$\begin{aligned}&E(X-\mu)(X-\mu)'=\\ &= LE(FF')L' + E(\epsilon F')L' + LE(F\epsilon') + E(\epsilon\epsilon')\\ &= LL' + \Psi\end{aligned}$$

$$\Sigma = LL' + \Psi \tag{4.4.6}$$

4.4.2 Estrutura da Covariância

1. $Cov(\boldsymbol{X}) = \boldsymbol{LL'} + \boldsymbol{\Psi}$

ou

$$Var(X_i) = \ell_{i1}^2 + ... + \ell_{im}^2 + \Psi_i$$
$$Cov(X_i, X_k) = \ell_{i1}\ell_{k1} + ... + \ell_{im}\ell_{km}$$

2. $Cov(\boldsymbol{X}, \boldsymbol{F}) = \boldsymbol{L}$

ou

$$Cov(X_i, F_j) = \ell_{ij}$$

A parte da variância da i-ésima variável contribuída pelos m fatores comuns é chamada de comunalidade. A parte da $Var(X_i) = \sigma_{ii}$ devida ao fator específico é chamada de singularidade ou variância específica. Denotando-se a i-ésima comunalidade por h_i^2:

$$\sigma_{ii} = \ell_{i1}^2 + ... + \ell_{im}^2 + \Psi_i$$

Onde

$$h_i^2 = \ell_{i1}^2 + ... + \ell_{im}^2$$
$$\sigma_{ii} = h_i^2 + \Psi_i \quad i = 1, 2, ..., p$$

Em outras palavras, a comunalidade medirá a proporção da variância total de uma variável observada que pode ou não ser explicada pelos fatores identificados no modelo. Quanto mais próximo de 1, a variável será completamente explicada pelos fatores, e quanto mais próxima de 0, a variável é única e não é explicada por nenhum dos fatores.

4.4.3 Metodo de Estimação

Descrever o que for utilizar na análise...

5 Resultados

5.1 Campeonato Brasileiro de Futebol de 2022 - Série A

Popularmente conhecido como Brasileirão, foi a 67ª edição da principal divisão de futebol no Brasil. Foi disputada entre os dias 9 de abril e 13 de novembro. Com a disputa da Copa do Mundo 2022 no final do ano, a competição teve inicio e fim antes do habitual. O Palmeiras se sagrou como o campeão dessa edição, enquanto Ceará, Atlético Goianiense, Avaí e Juventude foram rebaixados para a Série B.

Figura 2: Tabela após a 38ª Rodada

Pos	Equipe V·D·E	Pts	J	V	E	D	GP	GC	SG
1	Palmeiras (C)	81	38	23	12	3	66	27	+39
2	Internacional	73	38	20	13	5	58	31	+27
3	Fluminense	70	38	21	7	10	63	41	+22
4	Corinthians	65	38	18	11	9	44	36	+8
5	Flamengo	62	38	18	8	12	60	39	+21
6	Athletico Paranaense	58	38	16	10	12	48	48	0
7	Atlético Mineiro	58	38	15	13	10	45	37	+8
8	Fortaleza	55	38	15	10	13	46	39	+7
9	São Paulo	54	38	13	15	10	55	42	+13
10	América Mineiro	53	38	15	8	15	40	40	0
11	Botafogo	53	38	15	8	15	41	43	−2
12	Santos	47	38	12	11	15	44	41	+3
13	Goiás	46	38	11	13	14	40	53	−13
14	Red Bull Bragantino	44	38	11	11	16	49	59	−10
15	Coritiba	42	38	12	6	20	39	60	−21
16	Cuiabá	41	38	10	11	17	31	42	−11
17	Ceará	37	38	7	16	15	34	41	−7
18	Atlético Goianiense	36	38	8	12	18	39	57	−18
19	Avaí	35	38	9	8	21	34	60	−26
20	Juventude	22	38	3	13	22	29	69	−40

Fonte: Elaborada pelo Autor

Acima, apresentamos a classificação final do campeonato. Os clubes classificados até a 14ª posição participaram das competições internacionais da América do Sul. Desde

o Palmeiras até o Athlético Paranaense, garantiram suas vagas na fase de grupos da Libertadores da América, enquanto o Atlético Mineiro e o Fortaleza se qualificaram para a fase pré-classificatória da Libertadores. Entre o 9º e o 14º lugares, os clubes asseguraram suas posições na Copa Sul-Americana. Notavelmente, Coritiba e Cuiabá foram os únicos clubes que não tiveram a oportunidade de disputar qualquer competição internacional no ano seguinte.

5.2 Banco de Dados

Os dados que serão analisados foram coletados e organizados em tabelas Excel. Eles foram obtidos da seguinte fonte: `FBref`[5]. Cada equipe possui estatísticas subdivididas em **Chute**, **Passe**, **Criação de finalizações e gols**, **Ações Defensivas**, **Posse**, **Tempo de jogo**, **Diversas** e **Goleiro**[6], onde cada jogador possui seus próprios dados. As variáveis de cada umas delas pode ser vistas em detalhe no Anexo 7. uma observação deve ser feita, no campo das tabelas existem algumas variáveis denominadas com o prefixo "Esperado". Essas variáveis levam em consideração um valor, que é fornecido por outro site[7], e calculado tomando o valor real obtido pelo jogador para aquela variável. Como não obtive a informação sobre a forma como os valores esperados eram calculados, optei por utilizar apenas as variáveis que tomam os valores reais. O conjunto consiste em registros dos jogadores referentes aos 20 clubes que disputaram o campeonato brasileiro do ano de 2022. Serão considerados dados de 752 jogadores dos clubes que participaram das 38 rodadas da competição, cujas análises irão considerar a área de atuação dos atletas em campo, ao ínves da posição propriamente dita do jogador.

5.2.1 Divisão por posição

Após organizar os dados e com a finalidade de melhorar os resultados da análise, os jogadores foram divididos em 4 grupos. Os **Goleiros**, **Defensores**, **Meio Campistas** e **Atacantes**. Naturalmente, na análise dos goleiros utilizaremos a tabela de mesmo nome, os demais jogadores de linha utilizarão todas as outras. Poderiamos optar por subdividir os dados por posição, porém o que é fornecido no site, em sua versão em inglês, são as áreas de atuação. Formas alternativas de buscar a posição de cada jogador foram consideradas por meio de uma busca em outras fontes, porém não havia uma unanimidade quanto a

[5] ⟨https://fbref.com/en/comps/24/2022/2022-Serie-A-Stats⟩
[6] Existem duas tabelas com estatísticas dos goleiros, ambas foram organizadas em uma só.
[7] ⟨https://www.statsperform.com/opta/⟩

posição dos jogadores. Exemplificando, um jogador poderia ser considerado como um volante na fonte A e meio-campo na fonte B.

5.3 Goleiros

Uma das posições mais importantes dentro do esporte concerteza é a de goleiro. Sua principal função dentro de campo é impedir que o time adversário marque gols. Além de serem ágeis, eles também precisam ter boa cordenação motora para serem capazes de se movimentar rapidamente entre as traves.

Continua...

5.4 Defensores

Atuam na linha de defesa e podem ser divididos em zagueiros e laterais. Os zagueiros são responsáveis por marcar os atacantes adversários, desarmando-os e evitando que eles cheguem ao gol. Os laterais são um pouco mais versáteis, pois além de atuar na defesa auxiliando os zagueiros, eles também apoiam o ataque ajudando a criar jogadas ofensivas.

Continua...

5.5 Meio Campistas

São os jogadores que atuam entre as linha de defesa e ataque, sendo responsáveis por recuperar a posse de bola, distribuir passes, controlar o ritmo de jogo e armar jogadas ofensivas. Podem ser divididos em volantes, que têm uma função mais defensiva, ajudando na marcação e no desarme do time adversário e os meio-ofensivos que ajudam na criação de jogadas, dando assistências e marcando gols.

Continua...

5.6 Atacantes

São os jogadores responsáveis por marcar gols. Atuam na linha de frente tendo como principal função finalizar as jogadas. Podem ser divididos em centroavantes, que são os jogadores que atuam mais centralizados, buscam sempre se posicionar bem para

finalizar as jogadas, e os pontas, que atuam principalmente nas laterias do campo, geralmente são jogadores rápidos e habilidosos, sendo responsáveis por driblar os adversário e criar situações de gols.

Continua...

6 Conclusão

7 Anexo

7.1 Variáveis específicas dos goleiros

Tabela 1: Descrição das variáveis dos goleiros

Codificação	Resumo
G01	Número de partidas que o goleiro realizou no campeonato
G02	Minutos em que o goleiro esteve em campo
G03	Soma dos gols sofridos durante a competição
G04	Média de gols por partida
G05	Chutes no alvo
G06	Chutes defendidos
G07	Defesas realizadas, em porcentagem
G08	Jogos sem sofrer gols
G09	Gols de pênalti sofridos
G10	Pênaltis defendidos
G11	Passes longos certos
	(São considerados os passes acima de 36 metros)
G12	Tentativas de passe longo
G13	Total de passes tentados
	(inclui todo tipo de passe)
G14	Lançamentos feitos
	(Considerados aqueles feitos com as mãos)
G15	Aproveitamento de passe, em porcentagem
G16	Distância média, em metros, dos passes tentados
	(Não incluem os tiros de meta)
G17	Número de lançamentos adversários para a pequena área
G18	Interceptações de lançamentos a pequena área
G19	Número de ações defensivas fora da pequena área
G20	Distância média, em metros, das ações defensivas fora da pequena área

7.2 Variáveis para o jogadores de linha

7.2.1 Chute

Tabela 2: Descrição das variáveis de Chute

Codificação	**Resumo**
A01	Número de gols marcados
A02	Total de finalizações
A03	Total de finzalizações ao alvo
A04	Total de finalizações ao alvo, em porcentagem
A05	Média de finalizações por jogo
A06	Média de finalizações certas por jogo
A07	Média de gols por finalização
A08	Média de gols por finalizações certas
A09	Distância média, em metros, de todas as finalizações (Não incluem os chutes de pênalti)
A10	Gols de Falta
A11	Pênaltis cobrados
A12	Gols de pênalti

7.2.2 Passe

Tabela 3: Descrição das variáveis de Passe

Codificação	Resumo
P01	Total de passes certos
P02	Total de passes tentados
P03	Aproveitamento dos passes, em porcentagem
P04	Distância total, em metros, de todos os passes realizados
P05	Distância progressiva dos passes, em metros (Considerados os passes em direção ao gol adversário)
P06	Passes curtos certos (Entre 4,5m e 13,7m)
P07	Tentativas de passes curtos
P08	Aproveitamento nos passes curtos, em porcentagem
P09	Passes médios certos (Entre 13,7m e 27,4m)
P10	Tentativas de passes médios
P11	Aproveitamento nos passes médios, em porcentagem
P12	Passes longos certos (Acima de 27,4m)
P13	Tentativas de passes longos
P14	Aproveitamento nos passes longos, em porcentagem
P15	Assistências para gol
P16	Passes que geraram um chute ao gol
P17	Passes no último terço do campo (Terço defensivo do time adversário)
P18	Passes completos para a grande área
P19	Cruzamentos completos para a grande área
P20	Passes que movem a bola em qualquer direção a linha de gol à pelo menos 9,1m dos seus ou qualquer passe concluído para a áre adversária

7.2.3 Criação de finalizações e gols

Tabela 4: Descrição das variáveis de Criação de finalizações e gols

Codificação	Resumo
GC01	Total de ações que geraram finalizações durante o campeonato
GC02	Total de ações médias que geraram finalizações (durante 90 minutos)
GC03	Total de ações que geraram gols durante o campeonato
GC04	Total de ações médias que geraram gols (durante 90 minutos)

7.2.4 Ações defensivas

Tabela 5: Descrição das variáveis de Ações defensivas

Codificação	Resumo
DEF01	Número de divididas
DEF02	Número de vezes em que o time ganhou a posse de bola após uma dividida
DEF03	Divididas no terço defensivo do campo
DEF04	Divididas no terço central do campo
DEF05	Divididas no terço de atque do campo
DEF06	Duelos individuais ganhos (Considera-se as vezes em que o jogador frustou um drible adversário)
DEF07	Total de duelos individuais
DEF08	Porcentagem de duelos ganhos
DEF09	Duelos individuais perdidos
DEF10	Total de vezes em que bloqueou a bola
DEF11	Número de vezes em que bloqueou um chute
DEF12	Número de vezes em que bloqueou um passe
DEF13	Número de interceptações (Vezes que, conscientemente, o jogador impediu uma ação ofensiva)
DEF14	Número de divididas mais número de interceptações
DEF15	Erros individuais que levaram a uma finalizaçao do time adversário

7.2.5 Posse

Tabela 6: Descrição das variáveis de Posse

Codificação	Resumo
PO01	Número de vezes que o jogador tocou na bola
PO02	Toques na grande área de defesa
PO03	Toques no terço defensivo
PO04	Toques no terço central
PO05	Toques no terço de ataque
PO06	Toques na grande área de ataque
PO07	Toques com a bola em jogo
PO08	Número de tentativas de drible
PO09	Tentativas de drible bem sucedidas
PO10	Porcentagem de dribles bem sucedidos
PO11	***
PO12	***
PO13	Número de vezes que o jogador dominou a bola
PO14	Distância total, em metros, que a bola foi movida após o domínio (Até o toque seguinte na bola)
PO15	Distância progressiva que a bola foi movida após o domínio (Até o toque seguinte na bola)
PO16	Vezes que o jogador se moveu na direção do gol adversário (Pelo menos 9 metros do ponto inicial)
PO17	Domínios no terço ofensivo
PO18	Domínios na área adversária
PO19	Vezes que o jogador não conseguiu dominar a bola
PO20	Número de vezes que o jogador foi desarmado após o domínio da bola
PO21	Número de vezes que o jogador recebeu um passe
PO22	Passes que moveram a bola em direção ao gol adversário

7.2.6 Tempo de jogo

Tabela 7: Descrição das variáveis Tempo de jogo

Codificação	**Resumo**
T01	Jogos disputados
T02	Minutos que esteve em campo (Total)
T03	Média de minutos por partida
T04	Porcentagem de minutos jogados *
T05	Jogos como titular
T06	Média de minutos como titular
T07	Partidas completas (O jogador não foi substituido)
T08	Jogos como reserva
T09	Minutos médios em campo após sair do banco de reservas
T10	Jogos que o jogador esteve no banco de reservas mas não atuou

7.2.7 Diversas

Tabela 8: Descrição das variáveis Diversas

Codificação	**Resumo**
M01	Cartões Amarelos
M02	Cartões Vermelhos
M03	2ª Cartão amarelo (No mesmo jogo)
M04	Faltas cometidas
M05	Faltas sofridas
M06	Impedimentos
M07	Cruzamentos*
M08	Cortes
M09	Divididas em que o próprio jogador ficou com a bola
M10	Pênaltis convertidos*
M11	Pênaltis concedidos*
M12	Gols contra
M13	Bolas recuperadas
M14	Disputas aéras ganhas
M15	Disputas aéreas perdidas
M16	Porcentagem de disputas aéreas ganhas

Referências

ANDERSON, C.; SALLY, D. *The numbers game: Why everything you know about soccer is wrong*. [S.l.]: Penguin, 2013.

HAIR, J. F. et al. *Análise multivariada de dados*. [S.l.]: Bookman editora, 2009.

JOHNSON, R. A.; WICHERN, D. W. et al. Applied multivariate statistical analysis. Prentice hall Upper Saddle River, NJ, 2002.

JOLLIFFE, I. T. *Principal component analysis for special types of data*. [S.l.]: Springer, 2002.

LEWIS, M. *Moneyball: The art of winning an unfair game*. [S.l.]: WW Norton & Company, 2004.

MORETTIN, P. A.; BUSSAB, W. O. *Estatística Básica*. [S.l.]: Saraiva Educação SA, 2017.