UNIVERSIDADE FEDERAL DO RIO DE JANEIRO
INSTITUTO DE COMPUTAÇÃO
BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

GABRIEL MARTINS DE FREIRE (123449127)

RELATÓRIO

CONSTRUINDO RANKING COM MÍNIMOS QUADRADOS E CLUSTERIZANDO ESTILOS DE JOGO DE FUTEBOL

RIO DE JANEIRO
2024

GABRIEL MARTINS DE FREIRE

RELATÓRIO

CONSTRUINDO RANKING COM MÍNIMOS QUADRADOS E CLUSTERIZANDO ESTILOS DE JOGO DE FUTEBOL

Relatório feito em referência à entrega do trabalho da disciplina Computação Científica de Análise de Dados, ofertada pelo Instituto de Computação da Universidade Federal do Rio de Janeiro, que possui como docente João Antonio Recio da Paixão, como parte do método de avaliação discente.

RIO DE JANEIRO
2024

SUMÁRIO

1. INTRODUÇÃO

Este projeto utiliza dados do Campeonato Brasileiro de Futebol de 2023 para desenvolver um modelo quantitativo que analisa o desempenho dos times com base em métricas como gols, finalizações, valor de mercado e outras informações relevantes. A ideia central é aplicar técnicas de estatística e otimização para criar rankings preditivos.

O foco principal está na construção de um sistema baseado no método de Mínimos Quadrados (MMQ), capaz de gerar rankings ajustáveis. O modelo considera fatores como gols marcados dentro e fora de casa, o valor das equipes e o número de finalizações, permitindo uma análise mais precisa dos elementos que influenciam o desempenho das equipes. Para garantir a consistência dos resultados, o modelo foi ajustado com o método de otimização Gradiente Conjugado e os rankings gerados foram comparados com o ranking oficial do campeonato, reduzindo os desvios e melhorando a precisão.

Além disso, foi realizada uma análise de clusterização utilizando o algoritmo de K-Means, com o objetivo de agrupar os times com base em suas características médias, como gols, escanteios, faltas e chutes.

2. DADOS UTILIZADOS

Os dados utilizados neste projeto foram extraídos do Transfermarkt e disponibilizados pela plataforma Base dos Dados, contendo informações detalhadas sobre o Campeonato Brasileiro de Futebol. O dataset abrange diversas métricas e características relevantes para a análise do desempenho dos times ao longo da competição.

Entre as principais colunas, destacam-se:

- Rodada e data: Identificam a rodada do campeonato e o dia em que cada partida foi realizada.
- Times mandante e visitante: Informam as equipes envolvidas em cada confronto.
- Gols e finalizações: Incluem os gols marcados por cada time, tanto no total quanto no primeiro tempo e no segundo tempo, o número de chutes realizados, tanto no total quanto chutes de fora da área e de bola parada.
- Valor das equipes: Apresentam o valor do elenco titular de cada time.
- Métricas táticas: Como escanteios, faltas, defesas e impedimentos.
- Público e estádio: Dados sobre o público presente e a capacidade máxima dos estádios.

Alguns desses campos formam a base para a construção dos rankings e da clusterização, permitindo identificar padrões de comportamento e desempenho das equipes. Caso deseje consultar o dataset, clique aqui.

Vale ressaltar que, durante todo o projeto, foram utilizados exclusivamente os dados do campeonato de 2023.

3. O PODER DOS MÍNIMOS QUADRADOS

3.1 COMO O MÉTODO FUNCIONA

O método de Mínimos Quadrados (MMQ) é uma técnica matemática amplamente utilizada para encontrar a solução de sistemas lineares ou ajustar um modelo aos dados de maneira ótima. Ele é especialmente útil em situações onde há múltiplas variáveis interdependentes, e o objetivo é minimizar os erros entre as previsões do modelo e os valores observados.
O MMQ busca minimizar o somatório dos erros quadráticos entre os valores estimados pelo modelo e os valores reais. Isso significa encontrar uma solução x que minimize a seguinte função de erro:

$$E = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Onde:

y_i : Valores reais observados.
$\hat{y}_i$: Valores estimados pelo modelo.
E : Soma dos erros quadrados.

Quando aplicado a sistemas lineares, o problema pode ser formulado como:

$$Ax = b$$

Onde:

A: Matriz que representa as variáveis independentes.
x: Vetor de coeficientes a ser encontrado.
b: Vetor de resultados observados.

O MMQ encontra a solução x que minimiza $\|\mathbf{Ax} - \mathbf{b}\|^2$, ou seja, encontra a melhor solução possível.

3.2 MÍNIMOS QUADRADOS NA OBTENÇÃO DE RANKINGS

No contexto de rankings, o método dos Mínimos Quadrados (MMQ) é uma ferramenta poderosa para estimar a força relativa de diferentes entidades — sejam equipes, indivíduos ou itens. Sua grande vantagem está na capacidade de lidar com dados inconsistentes, ruidosos ou até incompletos, gerando uma classificação equilibrada e confiável.

Com o MMQ, cada confronto entre entidades é traduzido em números e organizado dentro de um sistema que faz sentido como um todo. Mesmo que os dados estejam fragmentados ou pareçam confusos à primeira vista, o método consegue estabelecer padrões e identificar conexões. Por exemplo, se a entidade A venceu B e B venceu C, o MMQ é capaz de inferir a relação entre A e C, utilizando essas interações indiretas para compor o panorama geral. O coração do método está em minimizar os erros quadráticos entre os valores observados e os estimados. Isso significa que o MMQ busca a solução que melhor se ajusta aos dados disponíveis, reduzindo ao máximo as discrepâncias. O resultado é uma classificação justa e ajustada, que reflete o desempenho relativo de cada entidade.

3.3 DEMONSTRAÇÕES DO PODER DOS MÍNIMOS QUADRADOS

Para ilustrar de forma clara e objetiva como a metodologia funciona na prática, foi elaborada uma série de demonstrações baseadas em jogos fictícios entre os times Vasco, Flamengo e Botafogo. Essas demonstrações têm como objetivo facilitar a compreensão dos conceitos envolvidos, além de destacar como diferentes abordagens — como a escolha de métricas, ajustes de pesos e modelagens alternativas — podem influenciar os resultados finais.

Para aplicar o método dos Mínimos Quadrados (MMQ) na construção de rankings, cada confronto entre times é modelado matematicamente como uma equação em um sistema linear. Esse sistema é representado por uma matriz A, um vetor b, e a incógnita x, que corresponde às pontuações ajustadas de cada time. A partir dessa modelagem, os rankings são ordenados com base nos valores do vetor x.

Matriz A:

- Linhas: Cada linha representa um confronto entre um time mandante e um time visitante
- Colunas: Cada coluna corresponde a um time.
- Valores:
    - O time mandante recebe +1
    - O time visitante recebe -1
    - Todos os demais times (que não participaram do confronto) recebem 0.

A matriz A segue a seguinte lógica:
(Desempenho Time Mandante) - (Desempenho Time Visitante) = Resultado
Onde o “desempenho” pode incluir fatores como: gols, valor de mercado da equipe e número de finalizações

Vetor b:

- Contém os resultados observados de cada confronto a depender das métricas exibidas.
- Exemplo: Gols do mandante - Gols do visitante

Vetor x:

- Representa as pontuações ajustadas ou "força" relativa de cada time.
- É calculado resolvendo o sistema linear Ax=b pelos Mínimos Quadrados, minimizando os erros entre os valores observados (b) e os valores ajustados (Ax).

3.3.1 DEMONSTRAÇÃO 1: CONSIDERANDO APENAS GOLS

Nessa demonstração será considerado apenas a métrica gol para determinar o ranking.

Confrontos:

- Vasco x Botafogo = 2x0
- Flamengo x Botafogo = 1x0

A matriz A, que organiza as relações entre os times, seria:

$$A = \begin{bmatrix} 1 & 0 & -1 \\ 0 & 1 & -1 \end{bmatrix}$$

Sendo a ordem das colunas: Vasco, Flamengo, Botafogo. (Isso serve para todas as demonstrações)

E o vetor b, com (gols feitos pelo mandante) - (gols feitos pelo visitante):

$$b = \begin{bmatrix} 2 \\ 1 \end{bmatrix}$$

Após resolver o sistema, obtemos o vetor x:

$$x = \begin{bmatrix} 1.00 \\ 0.00 \\ -1.00 \end{bmatrix}$$

Ou seja, temos o seguinte ranking:

1. Vasco da Gama: 1.00
2. Flamengo: 0.00
3. Botafogo: -1.00

Esses valores refletem diretamente os desempenhos observados nos confrontos. O Vasco lidera o ranking devido à sua vitória por 2x0 sobre o Botafogo, enquanto o Flamengo

ocupa a segunda posição visto que obteve sua vitória por apenas 1x0. Por outro lado, o Botafogo, que perdeu ambas as partidas, fica na última posição com uma pontuação negativa. Essa análise demonstra o poder do MMQ em distribuir as pontuações de maneira proporcional ao desempenho de cada equipe nos confrontos diretos. O método reconhece que o Vasco teve um desempenho superior ao Flamengo contra o Botafogo, refletindo isso na pontuação atribuída.

3.3.2 DEMONSTRAÇÃO 2: TODOS OS RESULTADOS EMPATADOS

Nesta demonstração, foi considerado um cenário em que todos os times têm uma vitória e uma derrota com o mesmo saldo de gols, criando uma situação simétrica. Este exemplo mostra como o MMQ reflete a igualdade de forças quando os dados observados não permitem distinguir um time do outro.

Confrontos:

- Vasco x Botafogo: 2x0.
- Flamengo x Botafogo: 0x2.
- Flamengo x Vasco: 2x0.

A matriz A seria:

$$A = \begin{bmatrix} 1 & 0 & -1 \\ 0 & 1 & -1 \\ -1 & 1 & 0 \end{bmatrix}$$

E o vetor b, com (gols feitos pelo mandante) - (gols feitos pelo visitante), é:

$$b = \begin{bmatrix} 2 \\ 0-2 \\ 2 \end{bmatrix} = \begin{bmatrix} 2 \\ -2 \\ 2 \end{bmatrix}$$

Após resolver o sistema, obtemos o vetor x:

$$x = \begin{bmatrix} 0.00 \\ 0.00 \\ 0.00 \end{bmatrix}$$

Ou seja, temos o seguinte ranking:

1. Vasco da Gama: 0.00
2. Flamengo: 0.00
3. Botafogo: 0.00

Os valores obtidos refletem a simetria completa dos resultados. Todos os times apresentaram o mesmo desempenho nos confrontos diretos, com saldos de gols idênticos, o que levou a um ranking completamente empatado. Nesse caso, o MMQ respeita os dados fornecidos, reconhecendo que não há elementos suficientes para distinguir as forças relativas das equipes de maneira justa.

3.3.3 DEMONSTRAÇÃO 3: VALORIZANDO GOLS FORA DE CASA

Nesta demonstração, introduzimos uma nova variável ao modelo: o peso dos gols fora de casa. Gols marcados fora de casa costumam ser mais difíceis e, por isso, são valorizados com um fator multiplicativo. Essa abordagem ilustra como o MMQ pode ajustar rankings com base em critérios adicionais, destacando diferenças de desempenho.

Confrontos:

- Vasco x Botafogo: 2x0
- Flamengo x Botafogo: 0x2
- Flamengo x Vasco: 2x0

Neste caso, aplicamos um peso de w_{fora} = 1.1 aos gols marcados fora de casa, valorizando esses desempenhos. A matriz A, que organiza as relações entre os times, permanece a mesma:

$$A = \begin{bmatrix} 1 & 0 & -1 \\ 0 & 1 & -1 \\ -1 & 1 & 0 \end{bmatrix}$$

O vetor b, ajustado para considerar o peso dos gols fora, é:

$$b = \begin{bmatrix} 2 \\ 2 \cdot w_{\text{fora}} \\ 2 \end{bmatrix} = \begin{bmatrix} 2 \\ 2.2 \\ 2 \end{bmatrix}$$

Após resolver o sistema, obtemos o vetor x:

$$x = \begin{bmatrix} 0 \\ -0.07 \\ 0.07 \end{bmatrix}$$

Ou seja, temos o seguinte ranking:

- Botafogo: 0.07
- Vasco da Gama: -0.00

- Flamengo: -0.07

Os valores mostram como o ajuste dos pesos altera o ranking final. Apesar de ter perdido para o Vasco em casa, o Botafogo lidera o ranking devido à sua vitória fora de casa contra o Flamengo, que recebeu um peso adicional no cálculo. Essa vitória foi suficiente para colocá-lo à frente do Vasco e do Flamengo, que venceram apenas em jogos realizados em casa.

3.3.4 DEMONSTRAÇÃO 4: INCORPORANDO MÉTRICAS ADICIONAIS

Nesta demonstração, adicionamos métricas adicionais para analisar os confrontos: finalizações e valor de mercado das equipes. O objetivo é mostrar como o método dos Mínimos Quadrados (MMQ) pode integrar múltiplos fatores e ajustar os rankings de forma mais detalhada. Além do saldo de gols, essas métricas oferecem uma visão mais abrangente sobre o desempenho relativo dos times.

Confrontos:

- Vasco x Botafogo: 2x1
  - Finalizações: Vasco = 10, Botafogo = 11
  - Valor dos times: Vasco = 5 milhões, Botafogo = 8 milhões

- Botafogo x Flamengo: 1x2
  - Finalizações: Botafogo = 12, Flamengo = 5
  - Valor dos times: Botafogo = 8 milhões, Flamengo = 7 milhões

- Flamengo x Vasco: 1x2
  - Finalizações: Flamengo = 2, Vasco = 4
  - Valor dos times: Flamengo = 7 milhões, Vasco = 5 milhões

Os pesos atribuídos às métricas foram:

- Peso para gols fora de casa ($w_{gol\ fora}$): 1.1
- Peso para valor de mercado (w_{valor}): 2
- Peso para finalizações ($w_{finalizações}$): 2

$$A = \begin{bmatrix} 1 & 0 & -1 \\ 0 & -1 & 1 \\ -1 & 1 & 0 \end{bmatrix}$$

O vetor b, ajustado para incluir as métricas adicionais, é calculado como:

$$b = \begin{bmatrix} 2 - (5-8) \cdot w_{\text{valor}} + (10-11) \cdot w_{\text{finalizações}} \\ 1 \cdot w_{\text{gol fora}} + (8-7) \cdot w_{\text{valor}} + (12-5) \cdot w_{\text{finalizações}} \\ 2 - (7-5) \cdot w_{\text{valor}} + (4-2) \cdot w_{\text{finalizações}} \end{bmatrix} = \begin{bmatrix} 2-6+(-2) \\ 1.1-2+14 \\ 2-4+4 \end{bmatrix} = \begin{bmatrix} -6 \\ 13.1 \\ 2 \end{bmatrix}$$

Após resolver o sistema, obtemos o vetor x:

$$x = \begin{bmatrix} -1.97 \\ -5.33 \\ 7.30 \end{bmatrix}$$

Ou seja, temos o seguinte ranking:

1. Botafogo: 7.30
2. Vasco da Gama: -1.97
3. Flamengo: -5.33

Os resultados mostram como a inclusão de métricas adicionais e os pesos atribuídos a elas podem ter um impacto significativo no ranking. Nesse caso, o Botafogo terminou na liderança, mesmo tendo perdido ambos os jogos. Isso aconteceu porque os pesos dados às finalizações e ao valor de mercado foram superestimados, o que fez essas métricas compensarem as derrotas no placar. Esse exemplo deixa claro como é fundamental escolher os pesos de forma cuidadosa.

O Vasco, mesmo vencendo o Botafogo em um confronto direto e apresentando um desempenho razoável nas finalizações e no valor de mercado, ficou na segunda posição devido ao peso desproporcional dessas métricas. Por outro lado, o Flamengo, que venceu o Vasco e o Botafogo, acabou em último lugar por ter apresentado menos finalizações e um menor valor de mercado, o que influenciou negativamente sua posição no ranking. Essa análise reforça que, embora o MMQ seja uma ferramenta flexível e permita considerar diversas métricas, a escolha equivocada dos pesos pode levar a resultados distorcidos e pouco intuitivos.

3.3.5 DEMONSTRAÇÃO 5: IMPACTO DOS EMPATES

Nesta demonstração, analisamos o impacto dos empates nos rankings, incorporando métricas adicionais, como finalizações e valor de mercado das equipes. Essa abordagem mostra como diferentes pesos podem influenciar significativamente a interpretação dos resultados, mesmo em partidas sem um vencedor.

Confrontos:

- Vasco x Botafogo: 1x1
    - Finalizações: Vasco = 8, Botafogo = 8
    - Valor dos times: Vasco = 3 milhões, Botafogo = 2 milhões

- Botafogo x Flamengo: 1x1
    - Finalizações: Botafogo = 7, Flamengo = 9
    - Valor dos times: Botafogo = 2 milhões, Flamengo = 4 milhões

- Flamengo x Vasco: 1x1

- Finalizações: Flamengo = 10, Vasco = 6
- Valor dos times: Flamengo = 4 milhões, Vasco = 3 milhões

$$A = \begin{bmatrix} 1 & 0 & -1 \\ 0 & -1 & 1 \\ -1 & 1 & 0 \end{bmatrix}$$

O vetor b, ajustado para incluir as métricas adicionais, é calculado como:

$$b = \begin{bmatrix} 1 - (3-2) \cdot w_{\text{valor}} + (8-8) \cdot w_{\text{finalizações}} \\ 1 \cdot w_{\text{gol fora}} - (2-4) \cdot w_{\text{valor}} + (7-9) \cdot w_{\text{finalizações}} \\ 1 - (4-3) \cdot w_{\text{valor}} + (10-6) \cdot w_{\text{finalizações}} \end{bmatrix} = \begin{bmatrix} 1 - 0.5 + 0 \\ 1.2 - (-1) \cdot 0.5 + (-1.6) \\ 1 - 0.5 + 3.2 \end{bmatrix} = \begin{bmatrix} 0.5 \\ 1.6 \\ 3.7 \end{bmatrix}$$

Após resolver o sistema Ax=b pelos Mínimos Quadrados, obtemos o vetor x:

$$x = \begin{bmatrix} -1.07 \\ -1.03 \\ 2.10 \end{bmatrix}$$

Ranking:
1. Flamengo: 2.10
2. Botafogo: -1.03
3. Vasco da Gama: -1.07

O Flamengo ficou na primeira colocação devido ao seu desempenho superior em finalizações e ao maior valor de mercado entre as equipes, o que foi suficiente para destacá-lo em um cenário de empates. Em seguida, o Botafogo garantiu a segunda posição, apresentando um saldo de finalizações mais favorável em comparação ao Vasco da Gama, mesmo com um valor de mercado inferior ao do Flamengo. Por outro lado, o Vasco da Gama terminou na última posição por ter tido o pior saldo de finalizações, não conseguindo se destacar nas métricas avaliadas. Esse resultado evidencia como a escolha e o peso das métricas podem influenciar de maneira significativa o ranking final, mesmo em partidas equilibradas.

4. MODELAGEM

A modelagem realizada neste projeto segue a abordagem utilizada nas demonstrações anteriores, sendo baseada na representação matricial dos confrontos entre os times. Cada confronto é representado como uma equação em um sistema linear, onde as relações entre os times são organizadas em uma matriz A, e os resultados observados (a depender da modelagem escolhida) são armazenados em um vetor b. A solução do sistema Ax = b pelos Mínimos Quadrados fornece um vetor x, que representa as forças relativas ou pontuações ajustadas de cada time.

A primeira modelagem feita foi considerando a vetor b apenas gols do mandante - gols do visitante.

Para melhor ilustrar, aqui a está a matriz A e vetor b na rodada 1 do Campeonato Brasileiro de 2023. Foram realizados 10 confrontos, envolvendo 20 times.

$$A = \begin{bmatrix} 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 \end{bmatrix} \quad b = \begin{bmatrix} 2 \\ 3 \\ 1 \\ -1 \\ 1 \\ 0 \\ 1 \\ 1 \\ -3 \end{bmatrix}$$

Ao modelar pela primeira vez, considerei apenas os gols marcados sem peso e obtive um ranking final bastante satisfatório, especialmente por refletir diretamente o desempenho das equipes ao longo das rodadas. Esse modelo inicial, apesar de simples, forneceu uma base sólida para análise. No entanto, há sempre a possibilidade de refinar os resultados e trazer mais nuances ao ranking ao introduzir outras métricas relevantes, como finalizações, posse de bola ou valor de mercado das equipes, associadas a pesos específicos para cada uma. Assim como demonstrado nas análises anteriores, a introdução de métricas adicionais pode oferecer uma visão mais detalhada sobre o desempenho relativo das equipes, indo além do simples saldo de gols. No entanto, é importante ter cuidado ao calibrar os pesos das métricas inseridas, para que elas não distorçam o ranking final de forma desproporcional, como observado na demonstração 4.

5. COMPARANDO COM O RANKING OFICIAL

Para avaliar a qualidade do ranking gerado pelo MMQ, considerando apenas os gols sem pesos adicionais, é necessário compará-lo com o ranking oficial do campeonato. Essa análise nos ajuda a entender o quão próximo o modelo está da realidade e a identificar possíveis ajustes para melhorar os resultados. A comparação pode ser feita de diferentes maneiras, cada uma com suas vantagens e limitações.

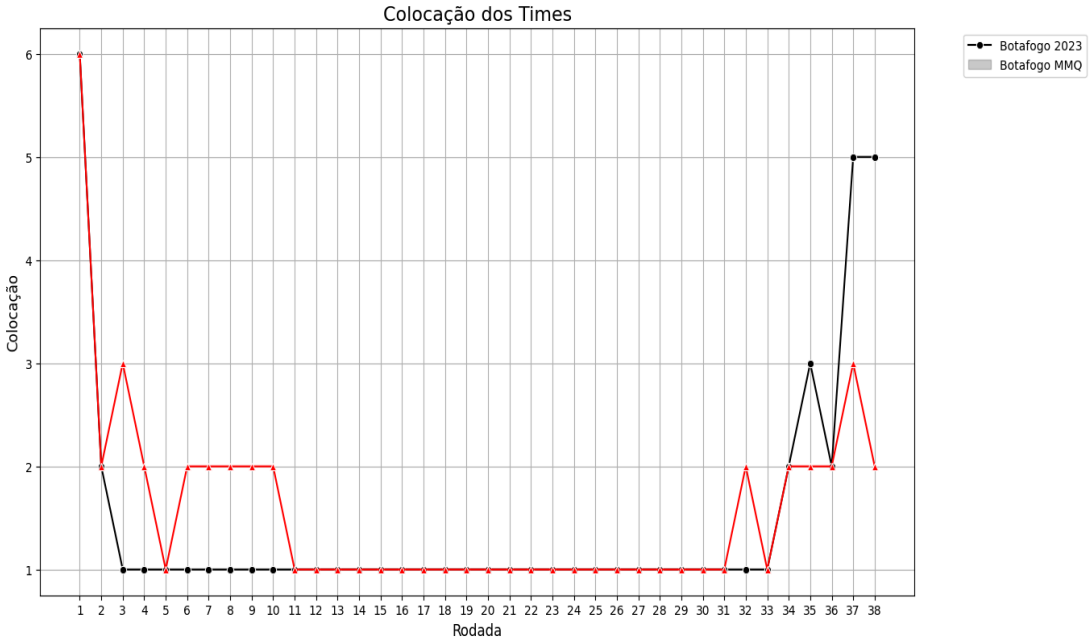
Uma abordagem inicial seria comparar os rankings rodada a rodada, analisando as colocações de cada time. No entanto, essa estratégia se torna pouco prática quando se deseja avaliar múltiplas rodadas, já que a quantidade de informações torna a análise confusa e difícil de interpretar.

A tabela da esquerda abaixo representa o ranking final na última rodada do Brasileirão 2023 de acordo com a modelagem apenas com gols, já a da direita representa o ranking oficial do campeonato.

rodada	time	colocacao
38	Palmeiras	1
38	Botafogo	2
38	Atlético-MG	3
38	RB Bragantino	4
38	Flamengo	5
38	Athletico-PR	6
38	Grêmio	7
38	Fluminense	8
38	Cruzeiro	9
38	São Paulo	10
38	Cuiabá-MT	11
38	Internacional	12
38	Fortaleza	13
38	Corinthians	14
38	EC Bahia	15
38	Vasco da Gama	16
38	Goiás	17
38	Santos	18
38	Coritiba FC	19
38	América-MG	20

rodada	time	colocacao
38	Palmeiras	1.0
38	Grêmio	2.0
38	Atlético-MG	3.0
38	Flamengo	4.0
38	Botafogo	5.0
38	RB Bragantino	6.0
38	Fluminense	7.0
38	Athletico-PR	8.0
38	Internacional	9.0
38	Fortaleza	10.0
38	São Paulo	11.0
38	Cuiabá-MT	12.0
38	Corinthians	13.0
38	Cruzeiro	14.0
38	Vasco da Gama	15.0
38	EC Bahia	16.0
38	Santos	17.0
38	Goiás	18.0
38	Coritiba FC	19.0
38	América-MG	20.0

Para tornar a comparação mais clara, podemos utilizar gráficos, selecionando alguns times específicos para plotar suas colocações ao longo das rodadas. Essa visualização permite observar o quão próximo o ranking gerado está do oficial, mas com a limitação de não incluir todos os times ao mesmo tempo. Na imagem abaixo, as linhas de vermelho representam a colocação obtida pelo Botafogo de acordo com a modelagem por MMQ utilizando apenas os gols. As linhas em preto são as colocações oficiais.



Para quantificar a proximidade entre os rankings de forma mais objetiva, foram desenvolvidas duas funções principais. A primeira calcula o erro total ao longo de todas as rodadas, somando as diferenças absolutas entre as colocações reais e as geradas pelo MMQ para todos os times e rodadas. Essa métrica fornece uma visão geral do desempenho do modelo em todo o campeonato, permitindo identificar padrões de erro mais amplos. Já a segunda função calcula o erro para uma rodada específica, como a última, que corresponde à classificação final. Essa métrica é útil para analisar com mais detalhe o momento mais relevante do campeonato, verificando a precisão do ranking no encerramento da competição.

```
print(f"Erro utilizando o ranking dado por MMQ apenas com gols: {erro_ranking(df_ranking,df_ranking_mmq)}")
print(f"Erro na rodada 38 utilizando o ranking dado por MMQ apenas com gols: {erro_ranking_rodada(df_ranking,
df_ranking_mmq, 38)}")
```

```
Erro utilizando o ranking dado por MMQ apenas com gols: 1216.0
Erro na rodada 38 utilizando o ranking dado por MMQ apenas com gols: 32.0
```

Essas métricas tornam possível não apenas verificar a proximidade entre os rankings, mas também medir o impacto de ajustes no modelo, como a inclusão de novas métricas e a aplicação de diferentes pesos. O objetivo é claro: reduzir tanto o erro total quanto o erro por rodada, ajustando o modelo para que ele reflita com maior precisão a realidade observada no ranking oficial. Por meio dessa análise, conseguimos validar se as modificações introduzidas, como pesos específicos para métricas adicionais, contribuem para melhorar a qualidade do modelo. Ao transformar a comparação em um processo quantificável e objetivo, é possível avaliar a eficácia do modelo inicial e refinar os resultados, garantindo que o ranking gerado seja uma representação mais fiel do desempenho real das equipes ao longo do campeonato.

6. OBTENÇÃO DOS PESOS

A definição dos pesos é uma etapa fundamental para ajustar a influência de diferentes métricas no modelo gerado pelos Mínimos Quadrados. Esses pesos calibram a importância de variáveis como saldo de gols, finalizações e valor de mercado, garantindo que o ranking reflita de forma equilibrada o desempenho real das equipes. Ao ajustar os pesos, é possível valorizar aspectos mais relevantes, como gols fora de casa, e minimizar a influência de métricas secundárias, reduzindo o erro total em relação ao ranking oficial. Neste tópico, será abordado como os pesos são escolhidos e ajustados para aprimorar a precisão do modelo.

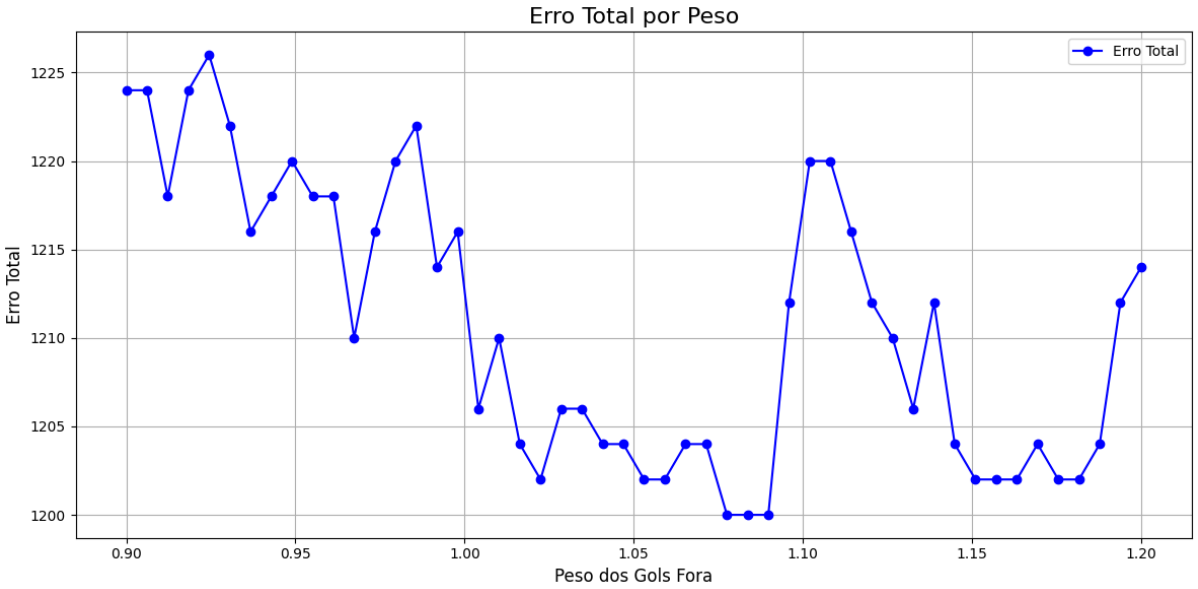
6.1 MANUALMENTE

A obtenção dos pesos manualmente foi realizada como um experimento para analisar a influência de métricas específicas no ranking final gerado pelo modelo de Mínimos Quadrados (MMQ). Inicialmente, foi escolhido um peso inicial "intuitivo" para cada métrica (gols como visitante, valor do time e finalizações) e, em seguida, ajustado gradualmente com base nos erros observados ao comparar o ranking gerado com o ranking oficial.

6.1.1 PESO PARA GOLS FORA DE CASA

A escolha do peso para gols marcados fora de casa foi realizada manualmente, começando com um valor inicial de 1.112. Após isso, foi ajustado iterativamente partindo de um outro intervalo para minimizar o erro total em relação ao ranking oficial. Esse processo experimental teve como objetivo avaliar a influência dos gols feitos como visitante (gols fora) no modelo gerado. Inicialmente, com o peso 1.112, o erro total foi 1218.0, enquanto o erro na rodada final (rodada 38) permaneceu em 26.0. Embora não seja perfeito, esse valor inicial já permitiu observar como a métrica de gols fora afeta a proximidade do ranking gerado em relação ao oficial. Para refinar o peso, foi criada uma faixa de valores entre 0.9 e 1.2, dividida em 50 incrementos iguais. O modelo foi recalculado para cada peso, e os erros foram registrados e plotados para identificar o valor que minimiza o erro total. O menor erro encontrado foi 1200.0, com um peso ideal de aproximadamente 1.078. Apesar da redução do erro total, o erro na rodada final permaneceu o mesmo (26.0).

Essa análise demonstrou que os gols fora de casa têm um impacto significativo no modelo e que o ajuste do peso pode melhorar o ranking gerado. No entanto, como as mudanças foram mínimas em algumas rodadas, os resultados sugerem que ajustes adicionais em outras métricas podem ser necessários para um refinamento mais completo.

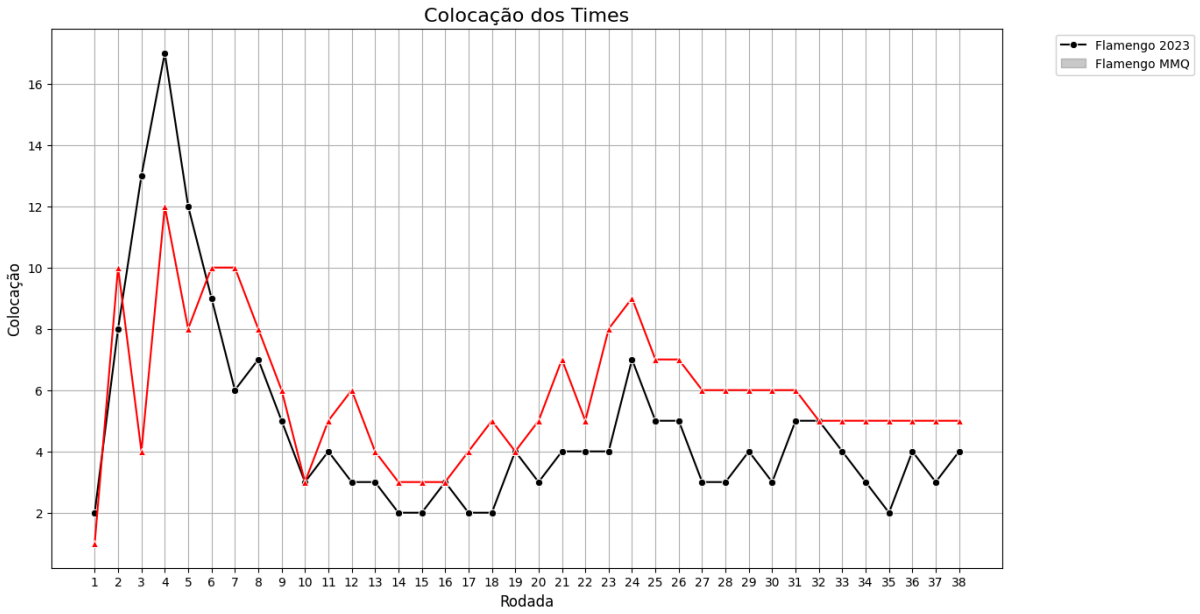


6.1.2 PESO PARA O VALOR DA EQUIPE

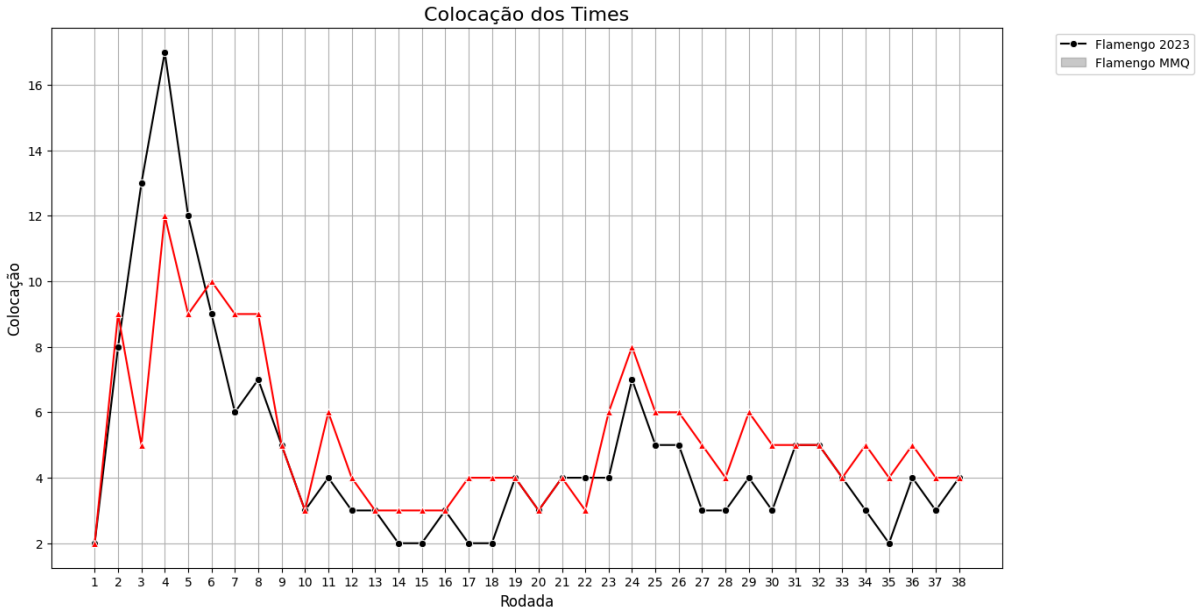
Após definir o peso ideal para gols fora de casa, o próximo passo foi ajustar manualmente o peso atribuído ao valor de mercado das equipes. Essa métrica representa a qualidade teórica do elenco e pode impactar significativamente o ranking, especialmente em times com investimentos mais elevados, como o Flamengo. Inicialmente, foi atribuído um peso experimental de 0.00001, que apresentou uma leve melhora no ranking em comparação com a configuração sem essa métrica. Esse resultado destacou o potencial da métrica, principalmente para times que possuem elencos mais caros. No entanto, para obter um peso mais preciso, foi criada uma faixa de valores entre −0.0000025 e 0.00001, dividida em 50 incrementos. Cada peso foi testado, e os erros correspondentes foram registrados e analisados. O menor erro total encontrado foi 1190.0, com um peso ideal de 1.327×10^{-6}, enquanto o erro na rodada 38 permaneceu em 26.0. Esse ajuste mostrou que o valor do elenco tem influência no ranking, mas com um impacto mais limitado quando comparado aos gols fora de casa.

A análise manual permitiu observar que a métrica de valor de mercado pode complementar as demais, mas precisa ser cuidadosamente calibrada para evitar que times com altos investimentos sejam supervalorizados, especialmente quando o desempenho em campo não corresponde ao valor do elenco. Apesar da melhora, ajustes adicionais em outras métricas podem ser necessários para alcançar um modelo ainda mais preciso.

Plot das colocações obtidas pela modelagem apenas com peso para gols fora:



Plot das colocações obtidas pela modelagem com peso para gols fora e peso para valor de mercado.



Percebe-se uma pequena melhora no modelo. Isso principalmente porque o Flamengo é um time com alto valor de mercado, o que melhora a colocação dele com a atribuição desse peso.

6.1.3 PESO PARA O FINALIZAÇÕES

A métrica de finalizações foi ajustada manualmente com o objetivo de identificar um peso ideal que reduzisse o erro total do modelo. Com base nos testes realizados, foi possível encontrar um peso eficiente que melhorou a proximidade do ranking gerado com o oficial. Para realizar o ajuste, foi definido um intervalo de pesos entre −0.04 e −0.01, dividido

em 50 valores igualmente espaçados. Cada peso foi avaliado calculando o erro total do modelo em todas as rodadas e o erro específico na última rodada (rodada 38). Os resultados foram registrados e plotados em um gráfico para identificar o peso com o menor erro. O peso ideal encontrado foi −0.0278, com um erro total de 1126.0 e erro na rodada 38 de 26.0. O gráfico demonstra claramente que o peso escolhido minimizou o erro total, indicando que as finalizações, quando calibradas corretamente, podem contribuir positivamente para o modelo.

Esse resultado mostra que a métrica de finalizações, combinada com os pesos ajustados para gols fora de casa e valor da equipe, pode desempenhar um papel significativo no refinamento do ranking. A abordagem manual funcionou bem nesse caso, permitindo um ajuste efetivo que resultou em uma melhora considerável no modelo.

6.2 REGRESSÃO LINEAR

Para automatizar o processo de ajuste de pesos das métricas no modelo, foi aplicada uma tentativa de regressão linear. O objetivo era calcular os coeficientes (pesos) associados às métricas principais do modelo: gols como mandante, gols como visitante, valor da equipe titular e finalizações, com o intuito de minimizar o erro no ranking gerado.

A modelagem do sistema feita foi:

a*(gols_mandantes) + b*(gols_visitante) + c*(finalizacao_mandante) + d*(finalizacao_visitante) + e*(valor_mandante) + f*(valor_visitante) = b

Sendo b = 3, se o time mandante tiver ganho a partida, b = 1 em casos de empate e b = 0 em casos de derrota do mandante.

Embora a regressão tenha identificado coeficientes para as métricas, os resultados não foram satisfatórios, já que o erro aumentou significativamente ao incluir todas as métricas no modelo. Isso sugere que as relações entre as métricas e a colocação final podem não ser puramente lineares. A regressão linear, embora útil para sistematizar a escolha dos pesos, apresentou limitações neste caso. A relação não linear entre algumas métricas e a colocação final pode ter prejudicado a eficácia do modelo.

6.3 GRADIENTE

Para otimizar os pesos e melhorar ainda mais o modelo de ranking, foi utilizado o método de Gradiente Conjugado. Esse método é uma evolução do Gradiente Descendente, ajustando os valores dos pesos considerando não apenas a direção do erro atual, mas também as direções das iterações anteriores. Isso permite que o modelo encontre de forma mais eficiente o ponto de mínimo local para os pesos. Inicialmente, foi testada uma configuração de pesos iniciais arbitrários [4,4,0,0] para avaliar como o Gradiente Conjugado ajustaria os valores relacionados às métricas de gols como mandante e visitante, valor da equipe e finalizações. Apesar do ajuste realizado pelo algoritmo, o erro final obtido foi maior que o erro dos pesos definidos manualmente:

- Erro com pesos otimizados via Gradiente Conjugado: 1184.0
- Erro com pesos manuais: 1122.0

Essa diferença inicial demonstrou que os pesos escolhidos manualmente já estavam próximos de um ponto de mínimo, sugerindo a necessidade de testar o Gradiente Conjugado com esses pesos como valores iniciais. Com os pesos manuais como ponto de partida, o Gradiente Conjugado foi aplicado novamente. Dessa vez, o método convergiu para os mesmos pesos manuais, confirmando que esses valores já representavam um ponto de mínimo local. Os resultados finais foram:

Pesos Otimizados via Gradiente Conjugado: [1, 1.053, 2.755×10^{-6} ,−0.0265]

- Erro Total: 1122.0
- Erro na Rodada 38: 24.0

O uso do Gradiente Conjugado validou os pesos obtidos manualmente, demonstrando que a escolha inicial já era eficiente para minimizar o erro total do modelo. A ausência de uma melhora significativa ao aplicar o método indica que os pesos encontrados já em um ponto de mínimo local, estavam adequadamente ajustados e não havia ganhos adicionais a serem explorados na proximidade desse ponto.

7. REFLEXÕES FINAIS SOBRE OS RANKINGS

O ranking final obtido pelo modelo foi:

rodada	time	colocacao
38	Palmeiras	1
38	Atlético-MG	2
38	Botafogo	3
38	Flamengo	4
38	Grêmio	5
38	RB Bragantino	6
38	Athletico-PR	7
38	Fluminense	8
38	Corinthians	9
38	Internacional	10
38	Cuiabá-MT	11
38	São Paulo	12
38	Cruzeiro	13
38	Fortaleza	14
38	EC Bahia	15
38	Vasco da Gama	16
38	Goiás	17
38	Santos	18
38	Coritiba FC	19
38	América-MG	20

Ao comparar com o ranking oficial abaixo, percebemos um resultado positivo.

rodada	time	colocacao
38	Palmeiras	1.0
38	Grêmio	2.0
38	Atlético-MG	3.0
38	Flamengo	4.0
38	Botafogo	5.0
38	RB Bragantino	6.0
38	Fluminense	7.0
38	Athletico-PR	8.0
38	Internacional	9.0
38	Fortaleza	10.0
38	São Paulo	11.0
38	Cuiabá-MT	12.0
38	Corinthians	13.0
38	Cruzeiro	14.0
38	Vasco da Gama	15.0
38	EC Bahia	16.0
38	Santos	17.0
38	Goiás	18.0
38	Coritiba FC	19.0
38	América-MG	20.0

Esse resultado reflete uma boa aproximação do desempenho observado no ranking oficial, evidenciando que o modelo, com os pesos otimizados, conseguiu gerar uma classificação coerente com os resultados reais. O Gradiente Conjugado se mostrou útil para validar as escolhas feitas manualmente, garantindo que o modelo atingisse seu potencial máximo de precisão.

Uma das limitações observadas ao ajustar pesos para métricas do Brasileirão 2023 foi a especificidade dos resultados. Os pesos otimizados para um ano podem não generalizar bem para outros anos, como demonstrado ao aplicar os pesos encontrados no modelo do Brasileirão 2022. Esse teste mostrou que o erro total foi maior ao utilizar os pesos otimizados em comparação com um modelo simples baseado apenas em gols.

Resultados do Brasileirão 2022

- Erro utilizando os pesos otimizados (2023):
    - Erro total: 2762
    - Erro na rodada 38: 72
- Erro utilizando apenas gols como métrica:
    - Erro total: 1862
    - Erro na rodada 38: 38

Esses resultados evidenciam que os gols são a métrica mais robusta e consistente para modelar rankings em diferentes anos. Como os pesos otimizados foram ajustados para os dados específicos de 2023, eles capturam características particulares desse ano, mas não se adaptam bem ao comportamento dos times e das métricas em 2022. Para minimizar

essas inconsistências e criar um modelo mais geral, seria necessário realizar o mesmo processo de otimização utilizando dados de vários anos.

8. CLUSTERIZAÇÃO

8.1 MÉTRICAS PARA CLUSTERIZAÇÃO

Para realizar a análise de clusterização, foi definido um dataframe consolidado contendo a média de todas as métricas relevantes para cada time ao longo da temporada. A abordagem permite agrupar as equipes com base em características semelhantes, considerando tanto os jogos como mandantes quanto como visitantes. Esse dataframe consolidado forma a base para a aplicação de técnicas de clusterização, permitindo identificar grupos de times com características semelhantes. A inclusão de métricas como gols, escanteios, faltas e chutes fornece uma visão abrangente sobre o estilo de jogo e o desempenho das equipes, servindo como um ponto de partida para análises mais avançadas.

time	gols	escanteios	faltas	chutes_bola_parada	defesas	impedimentos	chutes	chutes_fora
América-MG	1.105263	4.459459	14.540541	12.189189	3.540541	1.540541	13.513514	5.810811
Athletico-PR	1.342105	5.710526	15.263158	13.263158	3.368421	1.763158	14.578947	6.394737
Atlético-MG	1.368421	4.947368	15.763158	12.526316	2.894737	1.605263	12.815789	4.921053
Botafogo	1.526316	4.578947	13.552632	12.868421	3.526316	1.078947	13.763158	5.236842
Corinthians	1.236842	4.394737	12.500000	13.394737	3.684211	1.447368	12.157895	4.947368
Coritiba FC	1.078947	4.342105	15.421053	13.973684	3.947368	1.631579	11.289474	4.763158
Cruzeiro	0.921053	6.315789	15.000000	12.394737	2.894737	1.842105	13.894737	6.263158
Cuiabá-MT	1.052632	4.921053	12.421053	14.789474	3.000000	2.026316	11.815789	4.842105
EC Bahia	1.315789	5.500000	14.447368	13.473684	3.684211	1.421053	14.157895	6.000000
Flamengo	1.473684	5.270270	13.108108	13.702703	3.594595	1.243243	13.027027	4.648649
Fluminense	1.342105	5.526316	13.289474	14.736842	3.052632	1.973684	14.552632	5.605263
Fortaleza	1.184211	5.054054	13.270270	14.000000	3.162162	2.243243	14.540541	5.756757
Goiás	0.947368	5.289474	16.763158	15.184211	3.236842	2.263158	13.052632	5.526316
Grêmio	1.657895	5.315789	12.605263	14.131579	3.500000	1.552632	13.394737	4.947368
Internacional	1.210526	4.131579	13.631579	13.105263	3.184211	1.578947	12.157895	5.078947
Palmeiras	1.684211	6.432432	15.324324	15.108108	3.027027	2.675676	16.135135	6.702703
RB Bragantino	1.289474	7.500000	15.868421	12.342105	2.526316	1.368421	16.789474	6.710526
Santos	1.026316	4.894737	15.078947	14.447368	3.736842	1.763158	12.342105	5.236842
São Paulo	1.052632	5.631579	13.947368	14.026316	2.605263	2.026316	13.657895	5.368421
Vasco da Gama	1.078947	5.342105	14.631579	13.447368	3.500000	1.789474	12.578947	5.500000

8.2 NORMALIZAÇÃO DOS DADOS

A normalização dos dados é uma etapa essencial quando lidamos com métricas de diferentes escalas, como gols, escanteios, faltas, chutes e outras. Sem a normalização, as métricas com maiores valores absolutos podem dominar a análise, tornando difícil a comparação entre as variáveis e a identificação de padrões. Ao normalizar os dados, garantimos que todas as métricas tenham a mesma importância, ajustando suas escalas para que possam ser comparadas de maneira justa.

Sem normalização, as diferenças entre os times podem ser pequenas demais para analisar efetivamente os clusters. Por exemplo, ao observar a métrica de defesas, vemos que o Vasco da Gama tem 3.5 defesas, enquanto o Athletico-PR tem 3.36. No data frame não normalizado, essa diferença é de apenas 0.14 defesas, o que parece uma variação pequena. No entanto, após normalização, essa diferença se torna mais evidente, pois a métrica do Vasco da Gama passa a ser 0.573 e a do Athletico-PR 0.225, o que representa uma diferença de 154%. Ou seja, a normalização revela muito mais sobre as diferenças entre os times, tornando a análise de clusters mais eficiente e precisa.

Para normalizar os dados, foi utilizado o StandardScaler da biblioteca scikit-learn, que padroniza as variáveis, ajustando-as para que tenham média 0 e desvio padrão 1. Isso preserva a distribuição original dos dados, mas a coloca em uma escala comum, facilitando comparações entre diferentes métricas.

time	gols	escanteios	faltas	chutes_bola_parada	defesas	impedimentos	chutes	chutes_fora
América-MG	-0.646486	-1.042026	0.180604	-1.635252	0.681156	-0.540094	0.002008	0.480027
Athletico-PR	0.451320	0.550782	0.776071	-0.437352	0.225357	0.057571	0.793745	1.421392
Atlético-MG	0.573299	-0.420840	1.188091	-1.259222	-1.029032	-0.366332	-0.516479	-0.954379
Botafogo	1.305169	-0.889898	-0.633475	-0.877640	0.643487	-1.779339	0.187522	-0.445285
Corinthians	-0.036594	-1.124428	-1.500887	-0.290590	1.061616	-0.790234	-1.005368	-0.911954
Coritiba FC	-0.768464	-1.191436	0.906182	0.355165	1.758499	-0.295681	-1.650702	-1.208926
Cruzeiro	-1.500335	1.321378	0.559217	-1.405985	-1.029032	0.269522	0.285300	1.209270
Cuiabá-MT	-0.890443	-0.454344	-1.565943	1.265093	-0.750279	0.764074	-1.259591	-1.081652
EC Bahia	0.329342	0.282748	0.103826	-0.202532	1.061616	-0.860884	0.480856	0.785025
Flamengo	1.061212	-0.009734	-0.999781	0.052914	0.824299	-1.338252	-0.359506	-1.393530
Fluminense	0.451320	0.316252	-0.850328	1.206388	-0.610903	0.622774	0.774190	0.148658
Fortaleza	-0.280550	-0.285012	-0.866152	0.384518	-0.320849	1.346463	0.765205	0.392885
Goiás	-1.378356	0.014715	2.012133	1.705380	-0.123085	1.399928	-0.340479	0.021384
Grêmio	1.915061	0.048219	-1.414146	0.531280	0.573798	-0.507633	-0.086256	-0.911954
Internacional	-0.158572	-1.459470	-0.568419	-0.613467	-0.262461	-0.436982	-1.005368	-0.699832
Palmeiras	2.037040	1.469883	0.826474	1.620496	-0.678707	2.507420	1.950167	1.917873
RB Bragantino	0.207363	2.829066	1.274833	-1.464690	-2.004668	-1.002185	2.436414	1.930486
Santos	-1.012421	-0.487848	0.624273	0.883510	1.200993	0.057571	-0.868479	-0.445285
São Paulo	-0.890443	0.450269	-0.308195	0.413870	-1.795603	0.764074	0.109300	-0.233163
Vasco da Gama	-0.768464	0.081723	0.255623	-0.231885	0.573798	0.128221	-0.692479	-0.021040

8.3 CLUSTERIZANDO ESTILO DE JOGO OFENSIVO

A clusterização foi realizada com o objetivo de identificar grupos de times com estilo de jogo ofensivo e não ofensivo. Para isso, selecionei métricas que refletem diretamente o desempenho ofensivo das equipes, como gols, escanteios, chutes de bola parada e chutes totais. A ideia é que times que marcam mais gols, têm maior número de escanteios e tentativas de finalizações, podem ser caracterizados como mais ofensivos.

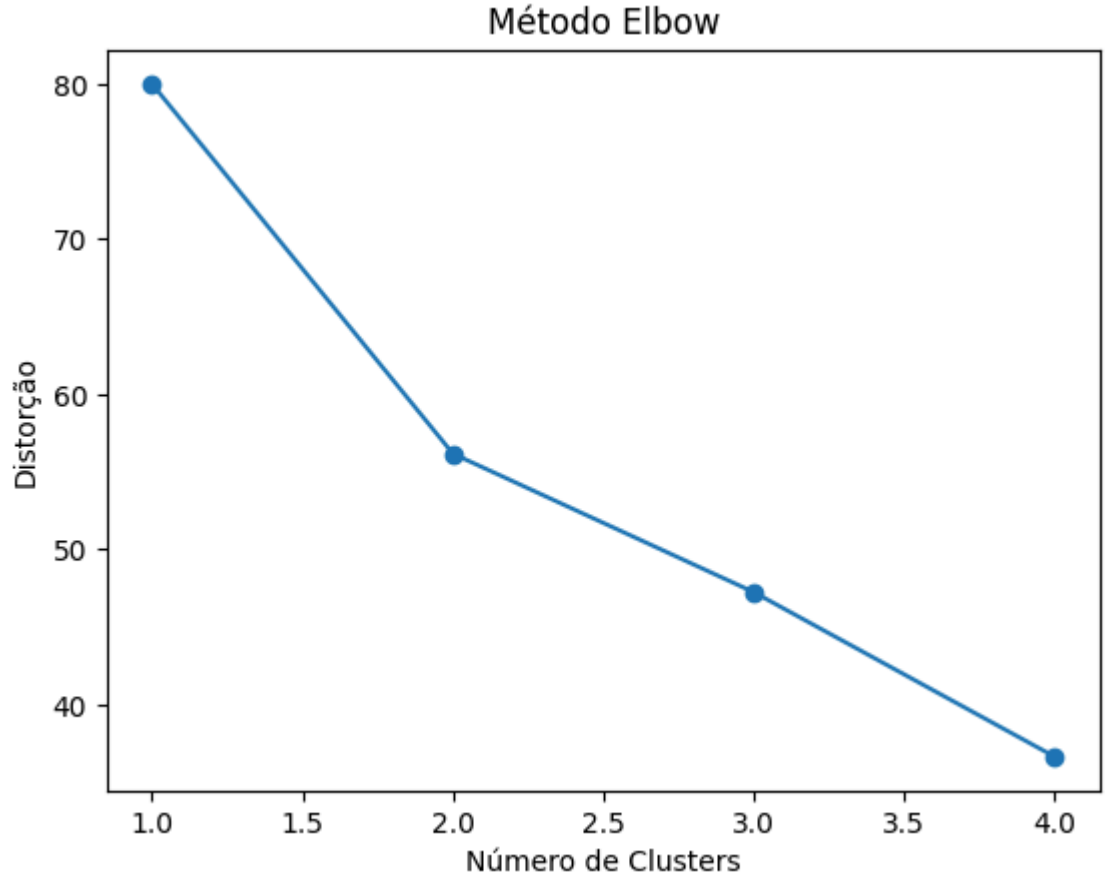
As métricas escolhidas para realizar a clusterização são:

- Gols: Indicador direto de ofensividade, já que times que marcam mais gols são geralmente mais ofensivos.
- Escanteios: O número de escanteios pode indicar pressão ofensiva, já que quanto mais escanteios, mais a equipe está forçando o adversário a se defender.
- Chutes de bola parada: Pode refletir tanto um aspecto de versatilidade no ataque, quanto uma limitação, dependendo da eficiência.
- Chutes totais: Mostra a capacidade de finalização de um time, onde times com maior número de chutes tendem a ser mais agressivos ofensivamente.

Essas variáveis foram selecionadas para representar o estilo de jogo ofensivo das equipes.

8.3.1 MÉTODO DO COTOVELO

O primeiro passo foi utilizar o método do cotovelo para determinar o número ideal de clusters. Esse método mede a "distorção" ou soma das distâncias quadráticas entre os pontos e o centróide mais próximo em cada cluster.



A partir do gráfico, ficou claro que dois clusters são suficientes para separar os times de acordo com seu desempenho ofensivo, sendo um grupo de times mais ofensivos e outro de times menos ofensivos.

8.3.2 CLUSTERIZAÇÃO COM K-MEANS

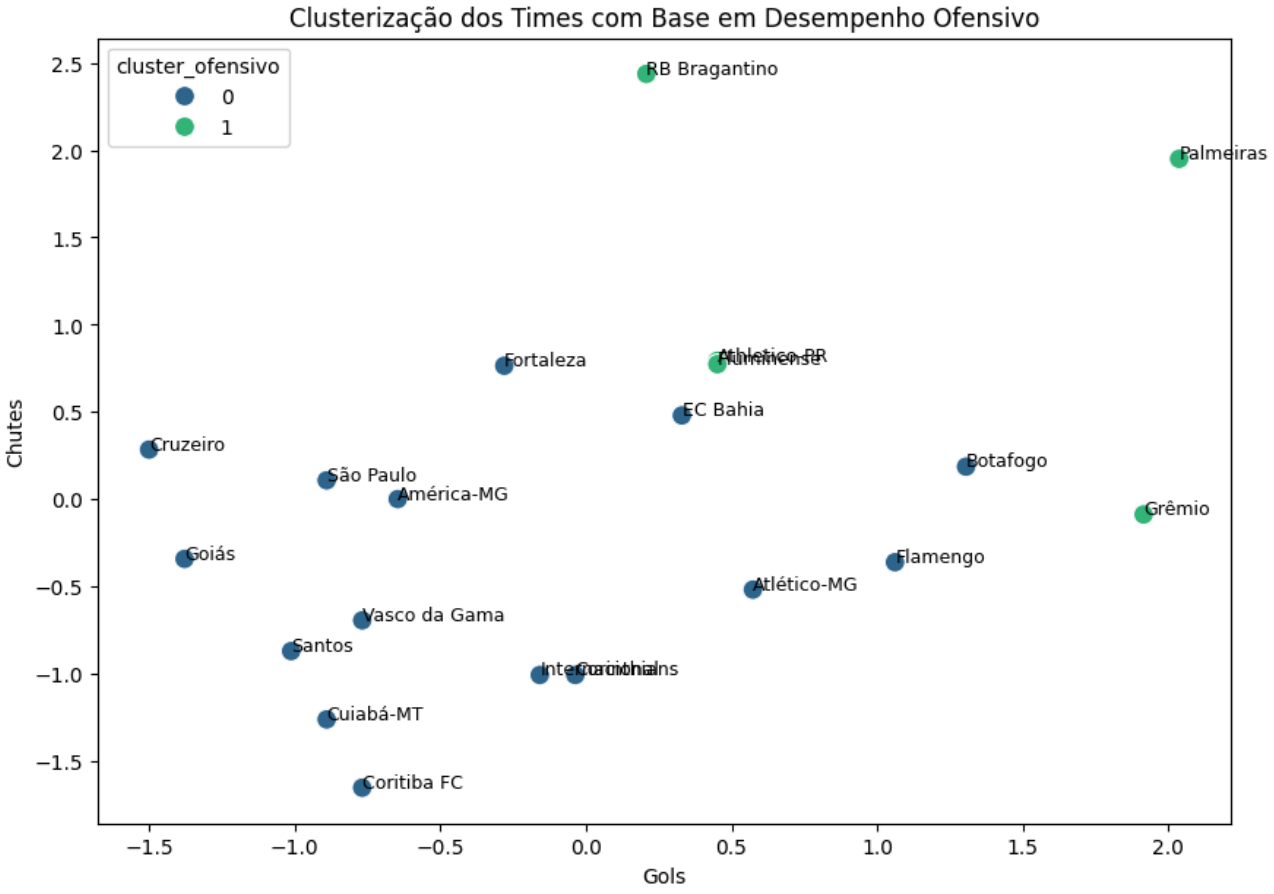
K-Means é um algoritmo de clusterização amplamente utilizado em aprendizado de máquina não supervisionado. Ele visa agrupar dados em k clusters (ou grupos) com base em características comuns entre as observações. O algoritmo divide os dados em grupos, de tal forma que os elementos dentro de um grupo são mais semelhantes entre si do que com os elementos de outros grupos.

Após definir o número de clusters como 2, apliquei o algoritmo K-Means para realizar a clusterização. O modelo foi treinado utilizando os dados normalizados das métricas selecionadas.

Para analisar o que cada grupo significa, a tabela a seguir corresponde a médias das métricas em cada grupo.

cluster_ofensivo	gols	escanteios	chutes_bola_parada	chutes
0	-0.337474	-0.347613	-0.097075	-0.391217
1	1.012421	1.042840	0.291224	1.173652

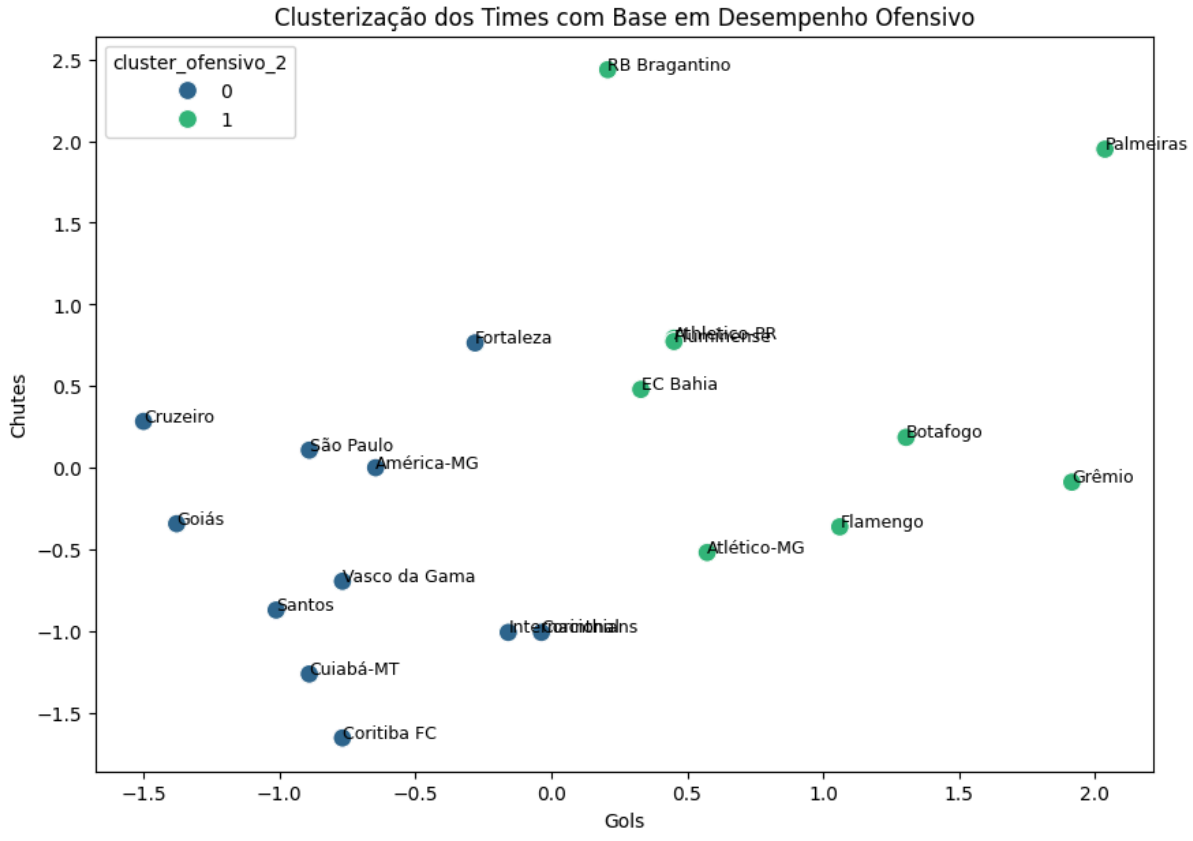
Os resultados mostraram que o Grupo 0 apresentou médias baixas em todas as métricas analisadas, indicando que este grupo consiste em times menos ofensivos. Por outro lado, o Grupo 1 apresentou médias mais altas em todas as métricas, caracterizando-o como times mais ofensivos. Para entender melhor os clusters, visualizamos a distribuição dos times com base nos gols e chutes, utilizando a biblioteca seaborn para um gráfico de dispersão. A análise visual confirmou que os times ofensivos e não ofensivos estavam bem diferenciados com base nos critérios escolhidos.



Embora a separação dos clusters tenha sido bem-sucedida, os times como Flamengo e Botafogo, que são conhecidos por seu estilo de jogo ofensivo, não foram adequadamente classificados como ofensivos. Para melhorar essa distinção, atribuímos um peso de 1.5 aos gols, que é a métrica mais importante para caracterizar um time como ofensivo. Após a modificação, os resultados mostraram uma separação mais clara entre os grupos, com o Grupo 0 ainda sendo composto por times menos ofensivos, mas com o Grupo 1 agora refletindo melhor os times de estilo ofensivo, como Flamengo e Botafogo.

cluster_ofensivo_2	gols	escanteios	chutes_bola_parada	chutes
0	-1.136063	-0.379680	0.075487	-0.514605
1	1.388521	0.464053	-0.092262	0.628962

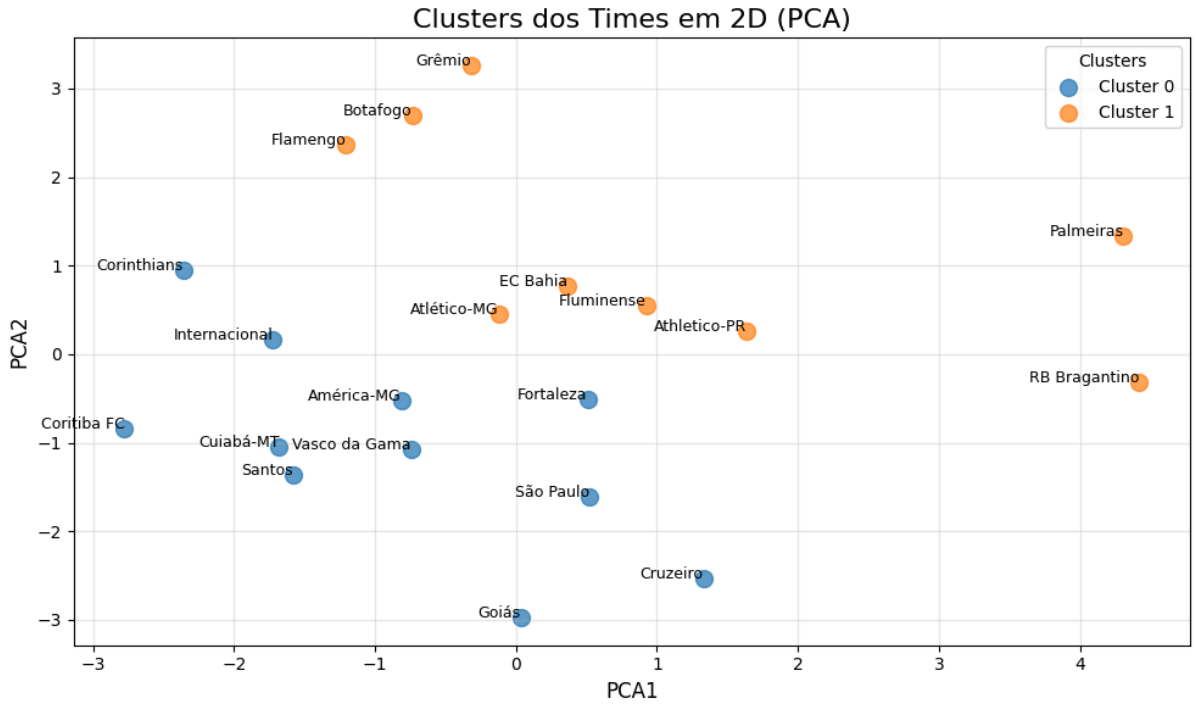
Com o peso para gols, temos times como Flamengo e Botafogo inseridos na categoria de times ofensivos, o que faz mais sentido.



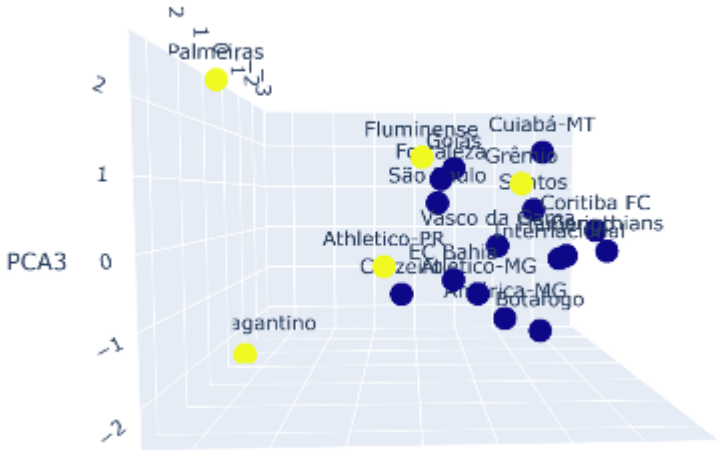
8.3.3 PCA

PCA (Principal Component Analysis) é uma técnica usada para reduzir a dimensionalidade dos dados, ou seja, diminuir o número de variáveis (ou "atributos") de um conjunto de dados enquanto mantém a maior parte da informação possível. Ela faz isso transformando os dados originais em um novo conjunto de variáveis, chamadas de componentes principais, que são combinações lineares das variáveis originais. Para

melhorar ainda mais a análise, utilizamos PCA (Análise de Componentes Principais) para visualizar os clusters em um espaço bidimensional e tridimensional.



Para uma visualização mais detalhada, o gráfico 3D foi criado utilizando três componentes principais, permitindo uma análise ainda mais precisa da distribuição dos times.



Para uma melhor visualização do gráfico 3D, acesse por meio do notebook.

8.3.4 CONCLUSÃO

A clusterização foi bem-sucedida em separar os times com base no estilo de jogo ofensivo. A técnica de PCA, combinada com o peso adicional dado aos gols, ajudou a

melhorar a definição dos grupos. No final, os times ofensivos, como Flamengo e Botafogo, foram corretamente identificados, resultando em uma clusterização mais precisa e significativa.

9. IMPLEMENTAÇÃO

A aplicação desenvolvida para este trabalho, amplamente documentada, está disponível no seguinte repositório do GitHub:
https://github.com/gbmartins9/Ranking-Brasileirao-CoCADa

10. CONCLUSÕES E RESULTADOS

Este projeto teve como objetivo aplicar métodos quantitativos para analisar o desempenho das equipes no Campeonato Brasileiro de 2023, utilizando técnicas como Mínimos Quadrados (MMQ) para gerar rankings preditivos e análise de clusterização para identificar padrões no estilo de jogo das equipes.

10.1 RANKING PREDITIVO

O modelo de Mínimos Quadrados (MMQ) demonstrou ser eficaz na criação de rankings preditivos para as equipes do Campeonato Brasileiro. Ao comparar o ranking gerado pelo modelo com o ranking oficial da competição, foi possível observar uma boa aproximação entre os resultados, indicando que o modelo consegue refletir de forma razoável o desempenho das equipes com base em métricas como gols, finalizações e valor de mercado. Apesar de bons resultados iniciais, ajustes adicionais nas métricas e pesos utilizados podem ser necessários para aumentar a precisão do modelo, especialmente ao se considerar outras variáveis que podem influenciar o desempenho das equipes. O modelo também mostrou limitações ao ser testado em edições passadas do campeonato, sugerindo que a generalização para outras temporadas deve ser cuidadosamente analisada.

10.2 ANÁLISE DE CLUSTERIZAÇÃO

A análise de clusterização, realizada com base em métricas como gols, escanteios, finalizações e chutes, foi eficaz em agrupar os times de acordo com seu estilo de jogo ofensivo e defensivo. A aplicação do algoritmo K-Means permitiu identificar dois grupos principais de times: um com um estilo de jogo mais ofensivo e outro com um estilo mais defensivo. A técnica de PCA (Análise de Componentes Principais) também foi utilizada para reduzir a dimensionalidade dos dados e proporcionar uma visualização mais clara dos clusters formados. Embora a clusterização tenha sido bem-sucedida em separar os times com base nas características ofensivas, ajustes adicionais nas métricas e pesos, especialmente em relação aos gols, foram necessários para melhorar a precisão da classificação de alguns times, como o Flamengo e o Botafogo.

11. AGRADECIMENTOS

Quero agradecer ao professor João Paixão pelo apoio e pelas ideias que me ajudaram ao longo do trabalho. Também sou muito grato ao monitor Matheus do Ó, que esteve sempre por perto para me ajudar nessa jornada.

12. REFERÊNCIAS

PAIXÃO, João. Notas de aula. 2024. Disponível em: <https://drive.google.com/drive/folders/1AhXSBGpbzY3WsybiZmCQ_xVMX6rCF6He?usp=drive_link>. Acesso em: 04 dez. 2024.