Visualização de Jogadores de Futebol em 2 e 3 dimensões utilizando PCA

Pedro Luis Mello Otero

18.07.2024
Computação Científica e Análise de Dados

OBJETIVO

O objetivo deste trabalho é, a partir das estatísticas individuais dos jogadores de futebol das principais ligas do mundo, representar os jogadores em gráficos 2d e 3d. A disposição dos jogadores nos gráficos deve representar as características de cada jogador em relação aos demais jogadores. Jogadores com características de jogo parecidas, devem ficar próximos um do outro, enquanto jogadores com características diferentes devem ficar mais distantes.

A base de dados utilizada neste trabalho contém mais de 60 estatísticas para cada jogador. É difícil visualizar a informação contida nessas estatísticas, pois é impossível visualizar um espaço geométrico de 60 dimensões. Por isso, foi aplicado o método de PCA (Análise de Componentes Principais) para reduzir a dimensão desses dados, permitindo que eles possam ser dispostos em 2 ou 3 dimensões.

DADOS UTILIZADOS

Os dados utilizados referem-se às estatísticas dos jogadores nas cinco principais ligas europeias na temporada 23/24. Essas são Premier League (Inglaterra), LaLiga (Espanha), Bundesliga (Alemanha), Serie A (Itália) e Ligue 1 (França). A base de dados inclui apenas as estatísticas individuais dos jogadores quando atuaram nessas ligas.

Sendo assim, estatísticas obtidas em torneios de seleções, copas nacionais e torneios continentais não foram consideradas.

Todos os dados foram retirados do Sofascore. As estatísticas contidas na base de dados do Sofascore são divididas nas seguintes categorias: ataque, defesa, passe, goleiro e outras. As estatísticas de goleiro não foram consideradas. As estatísticas utilizadas foram as seguintes:

ATAQUE: gols, gols esperados (xG), grandes chances perdidas, dribles certos, finalizações totais, finalizações no gol, finalizações para fora, chutes bloqueados, conversão de gols %, pênaltis cobrados, gols de pênalti, gols de bola parada, pênaltis sofridos, finalizações de bola parada, gols de dentro da área, gols de fora da área, gols de cabeça, gols com a perna esquerda, gols com a perna direita, finalizações na trave, impedimentos, conversão de pênalti %, conversão de bola parada %

DEFESA: desarmes, interceptações, pênaltis cometidos, cortes, erros que levaram ao gol, erros que levaram ao chute, gols contra, dribles sofridos, clean sheet

PASSE: grandes chances criadas, assistências, passes certos, passes errados, passes totais, acerto no passe %, passes certos no próprio campo, passes certos no campo adversário, passes corretos no terço final, passes decisivos, cruzamentos certos, acerto no cruzamento %, bolas longas certas, acerto na bola longa %, passe para o assistente

OUTROS: cartões amarelos, cartões vermelhos, duelos ganhos no chão, duelos ganhos no chão %, duelos aéreos ganhos, duelos aéreos ganhos % , duelos ganhos, duelos ganhos %, minutos jogados, faltas sofridas, faltas cometidas, desarmes sofridos, perda de posse de bola, jogos, jogos titular

PREPARAÇÃO DOS DADOS

Antes de utilizar o PCA, foram necessários realizar alguns ajustes nos dados para obter resultados melhores.

Primeiro, todos os goleiros foram removidos. Pelo fato dos goleiros terem uma função muito específica no jogo de futebol, que é muito diferente das demais, eles não foram considerados. O objetivo disso é impedir que o PCA apenas separe os goleiros dos jogadores de linha, já que essa é uma informação óbvia e pouco relevante.

Os jogadores com poucos minutos jogados também foram removidos. Jogadores que entraram pouco em campo têm estatísticas muito inconsistentes, por isso foram removidos.

As estatísticas foram substituídas por suas médias a cada 90 minutos jogados. Isso quer dizer, que, em vez de considerar informações como número de gols marcados na temporada, foi considerado o número de gols marcados a cada 90 minutos jogados. O motivo disso, é impedir que o PCA separe os jogadores apenas por minutos jogados. É claro que jogadores que jogaram por mais tempo terão números maiores do que aqueles que jogaram menos tempo.

Todos os dados foram normalizados. O objetivo é impedir que o PCA leve mais em conta estatísticas que são naturalmente maiores, em detrimento das que são naturalmente menores. Um jogador pode facilmente completar mais de 1000 passes em uma temporada, enquanto é muito difícil um jogador marcar mais de 30 gols. O modelo deve considerar todas as estatísticas igualmente.

PCA

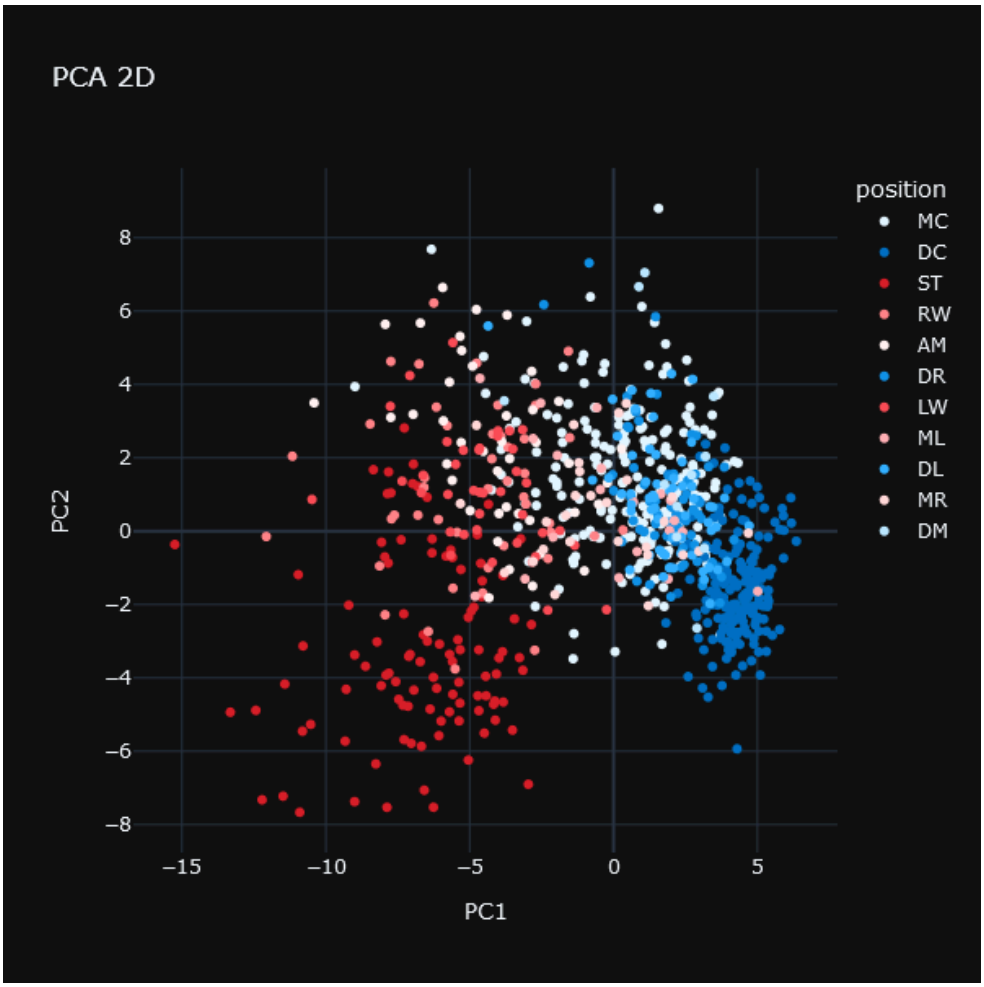
O PCA (Análise de Componentes Principais) é um método utilizado para redução de dimensões de dados com a menor perda de informação possível. A forma como o PCA funciona é encontrando os vetores ortogonais que minimizam a soma das distâncias entre os vetores e os dados. A partir disso, a projeção dos dados no espaço formado pelas componentes principais, é a que melhor aproxima os dados na sua dimensão original.

No projeto o PCA foi aplicado a uma matriz 836 x 62, contendo 836 linhas que representam todos os jogadores que satisfizeram os critérios estabelecidos e 62 colunas que representam cada uma das estatísticas mencionadas.

Foi calculada a matriz de covariância, a partir dessa matriz foram obtidos os autovalores, e a partir dos três maiores autovalores foram obtidos os três autovetores correspondentes. Esses autovetores são os componentes principais. Os jogadores foram, então, projetados em cada uma dessas componentes.

RESULTADOS

GRÁFICO 1

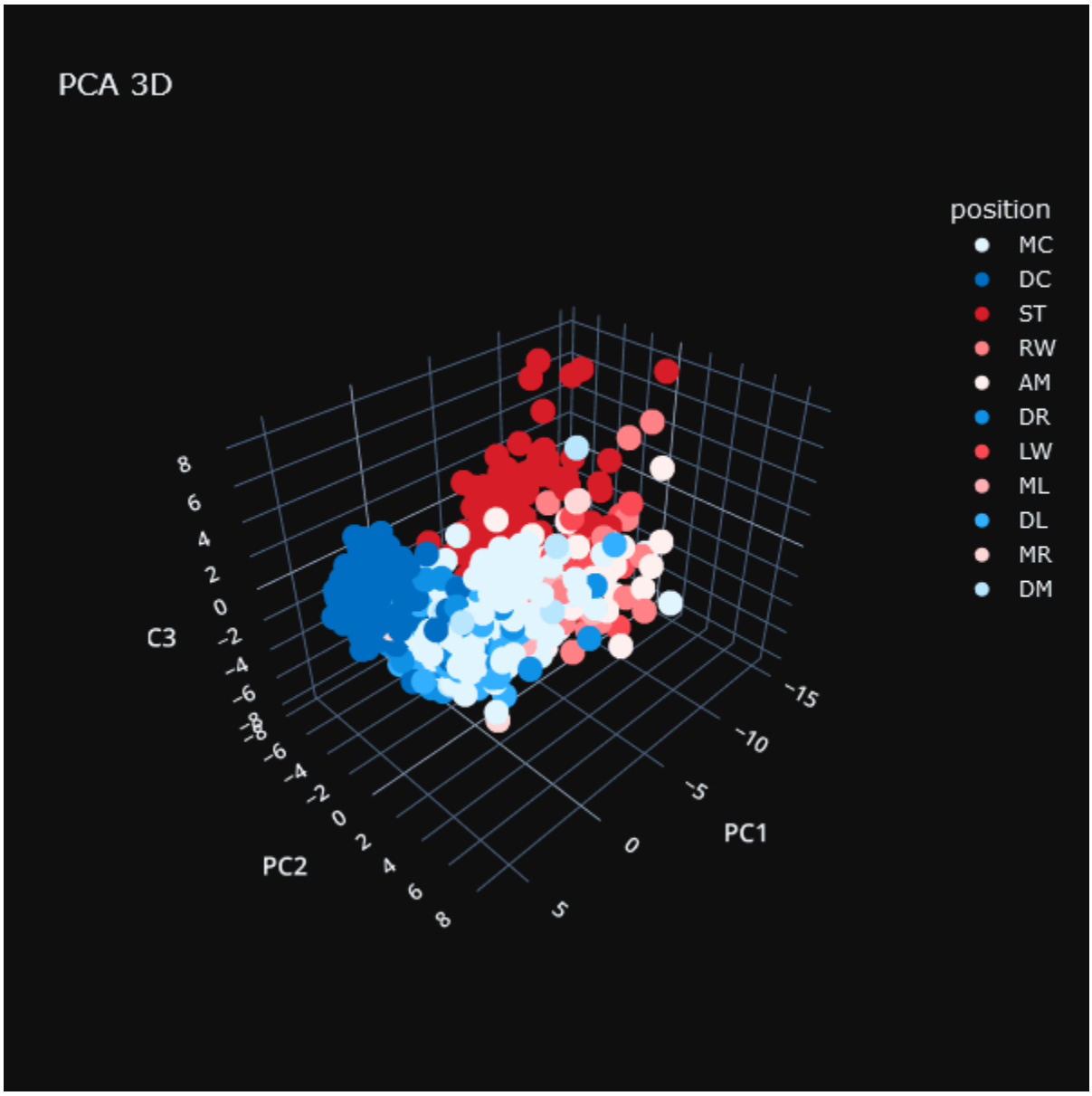


A legenda do gráfico, que indica as cores, se refere a posição dos jogadores. Jogadores em vermelho representam atacantes, jogadores em azul representam defensores e jogadores em branco são meio campistas.

Mais precisamente: (DC = Zagueiro central, DR = Lateral direito, DL = Lateral Esquerdo, DM = Volante, MC = Meia central, AM = Meia ofensivo, ML = Meia esquerda, MR = Meia direita, RW = Ponta direita, LW = Ponta esquerda, ST = Centro-avante)

Apenas olhando para o gráfico, é perceptível que o PCA foi capaz de afastar os jogadores de posições ofensivas dos jogadores de posições defensivas. Principalmente no eixo da primeira componente principal. Os atacantes ficaram dispostos mais à esquerda, enquanto os defensores ficaram mais à esquerda. Os meio-campistas ficaram entre atacantes e os defensores, já que o meio-campo envolve tanto atributos defensivos, quanto ofensivos.

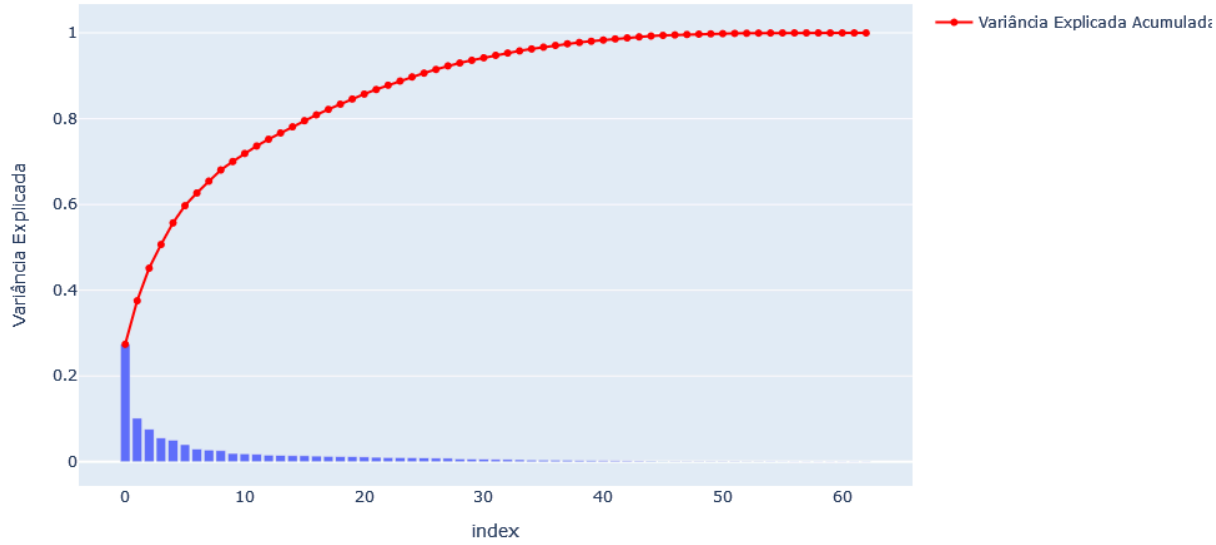
GRÁFICO 2



Esse gráfico representa os jogadores em três dimensões.

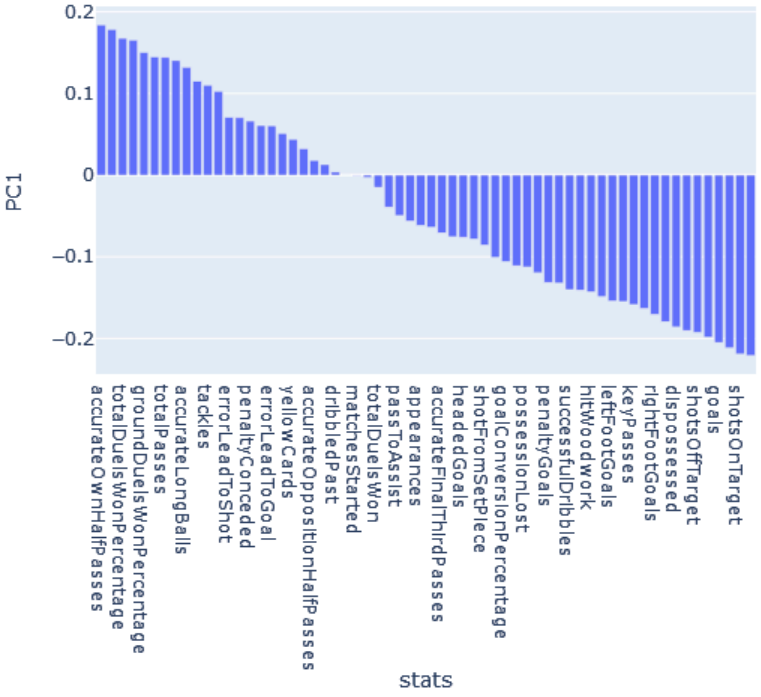
Os gráficos 1 e 2 podem ser visualizados de maneira interativa acessando os arquivos html ou executando o notebook. Na versão interativa, é possível ver o nome dos jogadores passando o mouse por cada ponto e filtrar os jogadores por posição. Também é possível girar o gráfico 3d para uma melhor visualização.

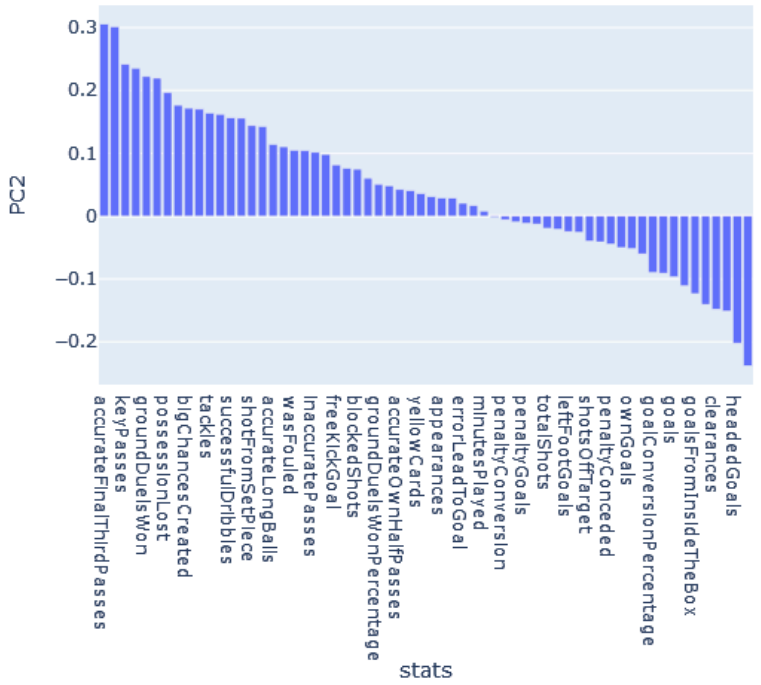
GRÁFICO 3

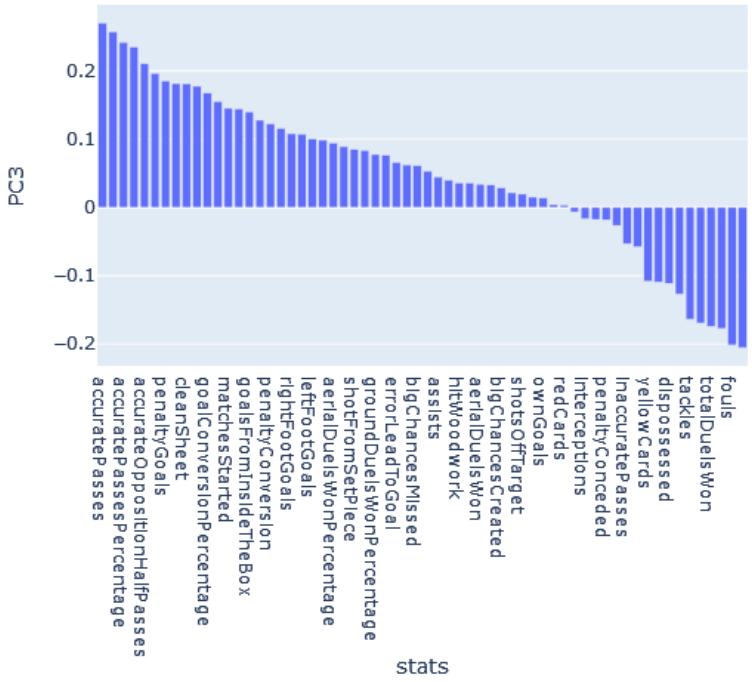


A variância explicada mostra a porcentagem da variância dos dados que pode ser explicada por cada componente. Nesse caso, temos que 37,5% da informação contida nos dados utilizados é representada com duas componentes principais (primeiro gráfico) e 45% da informação é representada com três componentes principais (segundo gráfico).

GRÁFICOS 4, 5, 6


0.2
0.1
0
−0.1
−0.2
PC1
shotsOnTarget
goals
shotsOffTarget
dispossessed
rightFootGoals
keyPasses
leftFootGoals
hitWoodwork
successfulDribbles
penaltyGoals
possessionLost
goalConversionPercentage
shotFromSetPiece
headedGoals
accurateFinalThirdPasses
appearances
passToAssist
totalDuelsWon
matchesStarted
dribbledPast
accurateOppositionHalfPasses
yellowCards
errorLeadToGoal
penaltyConceded
errorLeadToShot
tackles
accurateLongBalls
totalPasses
groundDuelsWonPercentage
totalDuelsWonPercentage
accurateOwnHalfPasses
stats


0.3
0.2
0.1
0
−0.1
−0.2
PC2
headedGoals
clearances
goalsFromInsideTheBox
goals
goalConversionPercentage
ownGoals
penaltyConceded
shotsOffTarget
leftFootGoals
totalShots
penaltyGoals
penaltyConversion
minutesPlayed
errorLeadToGoal
appearances
yellowCards
accurateOwnHalfPasses
groundDuelsWonPercentage
blockedShots
freeKickGoal
inaccuratePasses
wasFouled
accurateLongBalls
shotFromSetPiece
successfulDribbles
tackles
bigChancesCreated
possessionLost
groundDuelsWon
keyPasses
accurateFinalThirdPasses
stats



Esses gráficos mostram as direções de cada componente principal em relação às estatísticas. A partir desses três gráficos, podemos saber quais estatísticas mais foram levadas em consideração para distribuir os jogadores em cada eixo dos gráficos. As estatísticas com valores mais distantes de zero, são as mais levadas em consideração.

RECOMENDAÇÃO DE JOGADORES

Utilizando PCA, também é possível fazer um sistema de recomendação de jogadores, o que pode ser útil em scouting. Basta calcular a distância euclidiana entre os jogadores no espaço gerado pelas componentes principais, e a partir de um jogador, encontrar os jogadores mais próximos a ele.

Segue os 5 jogadores recomendados pelo algoritmo implementado para substituir jogadores que foram transferidos ou se aposentaram essa temporada:

- Kyllian Mbappé - ['Mohamed Salah', 'Deniz Undav', 'Cole Palmer', 'Vinícius Júnior', 'Harry Kane']
- Toni Kroos - ['Rodri', 'Joshua Kimmich', 'Exequiel Palacios', 'Trent Alexander-Arnold', 'Kieran Trippier']

- Douglas Luiz - ['Piotr Zieliński', 'Felipe Anderson', 'Robert Navarro', 'James Ward-Prowse', 'Julien Ponceau']
- Savinho - ['Adrián Embarba', 'Lorenzo Pellegrini', 'Ernest Nuamah', 'Gabriel Martinelli', 'Takefusa Kubo']
- João Palhinha - ['Christian Nørgaard', 'Juan Miranda', 'Marc Roca', 'Joakim Mæhle','Fabian Holland']

CONCLUSÃO

Há algumas limitações, só 45% da informação é preservada no gráfico 3d e 37,5% da informação no gráfico 2d. Também, pelo fato da base de dados utilizada não incluir informações sobre a posição dos jogadores em campo, como mapas de calor, os gráficos não mostraram muito a diferença entre jogadores de lado de campo. Pontas direita e esquerda se encontram misturados no gráfico, o mesmo acontece com meio-campistas e laterais.

No entanto, a aplicação do PCA em jogadores de futebol forneceu uma maneira muito interessante de visualizar os jogadores. É uma maneira muito mais fácil de identificar similaridades em estilo de jogo de jogadores do que apenas olhar para tabelas. Também permitiu identificar quais estatísticas são melhores em diferenciar os jogadores olhando para as componentes principais.

REFERÊNCIAS

http://www2.ic.uff.br/~aconci/PCA-ACP.pdf

https://exploratory.io/note/kanaugust/Introduction-to-PCA-Principal-Component-Analysis-with-FIFA-Soccer-Data-POb2BMx6ap

https://datalesdatales.medium.com/visualising-football-players-in-two-dimensions-with-pca-92c7bb005ab4