Fatec Baixada Santista - Rubens Lara
Ciência de Dados - 2° Ciclo

Principal Component Analysis (PCA)
Atividade - Álgebra Linear

Josiel Santana de Medeiros
Stefany de Oliveira Fernandes

Santos - SP
Dezembro/2022

1 Introdução

Conceito de PCA

A Análise de Componentes Principais, ou Principal Component Analysis (PCA), consiste em uma técnica que transforma dados com várias dimensões em dimensões menores, tentando manter o máximo de informação possivel. Essa técnica é útil em vários casos, principalmente em que há um dataset com várias features, como, por exemplo, em processamento de imagens ou pesquisa de genomas.

Escolha do Dataset

Para fazer esse trabalho, foi escolhido um dataset da Copa do Mundo da Fifa disponibilizado no Kaggle, que contém dados das edições de 1930 até 2014. Nele haviam três tabelas, sendo elas: todos os jogadores que já participaram de alguma edição, todas as partidas e todos os vencedores do evento. Nesse trabalho será usada apenas a tabela com os vencedores da Copa do Mundo, que nesse caso é o arquivo *WorldCups.csv*.

	GoalsScored	QualifiedTeams	MatchesPlayed
0	70	13	18
1	70	16	17
2	84	15	18
3	88	13	22
4	140	16	26
5	126	16	35
6	89	16	32
7	89	16	32
8	95	16	32
9	97	16	38
10	102	16	38
11	146	24	52
12	132	24	52
13	115	24	52
14	141	24	52
15	171	32	64
16	161	32	64
17	147	32	64
18	145	32	64
19	171	32	64

Figure 1: Dataset utilizado para o trabalho.

2 Processo

Tratamento do Dataset

Usando da biblioteca **Pandas**, o dataset foi importado e foram definidos as colunas que serão **inputs e outputs**. Selecionamos as colunas de 'GoalsScored', 'QualifiedTeams' e 'MatchesPlayed' para serem as três variáveis de inputs (X) e o nome das classes dos vencedores para como output (Y), essa que contém 20 países: 'Uruguay', 'Italy', 'Argetina', 'France', entre outros.
Outras colunas que estavam nesse dataset e continham nomes, como 'Country' ou 'Runners' foram desprezadas, pois a análise se aplica apenas para analisar dados quantitativos.
Com isso, temos os inputs com a dimensão de 20 linhas e 3 colunas (20,3) e os outputs com dimensão de 20 linhas (20,).

Definido o número de Principal Components

Nessa etapa foram usados dois módulos da biblioteca **Sklearn**, no caso a **Preprocessing** e a **Decomposition**. Com isso, já pode ser usada a função PCA do módulo Decomposition, onde foi definido como **dois** a quantidade de Principal Components.
Vale considerar que o número de componentes principais é sempre menor ou igual ao número de variáveis originais, então também seria possível definir o número de componentes como 3.

Valores de Scores

Tendo essas colunas, era necessário fazer uma transformação nesses dados para todos terem o mesmo peso, para isso foi calculado o scores dessas variáveis:

	PC1	PC2	Teams
0	-2.367875	-0.231737	Uruguay
1	-2.156237	-0.396303	Italy
2	-1.957052	0.002819	Italy
3	-1.911713	0.143488	Uruguay
4	-0.610394	1.203500	Germany FR
5	-0.543743	0.651918	Brazil
6	-1.299687	-0.231376	Brazil
7	-1.299687	-0.231376	England
8	-1.194030	-0.077705	Brazil
9	-0.950023	-0.155160	Germany FR

10	-0.861976	-0.027102	Argentina
11	1.057175	0.303533	Italy
12	0.810643	-0.055031	Argentina
13	0.511283	-0.490430	Germany FR
14	0.969128	0.175474	Brazil
15	2.572151	0.190438	France
16	2.396057	-0.065679	Brazil
17	2.149525	-0.424243	Italy
18	2.114306	-0.475467	Spain
19	2.572151	0.190438	Germany

Figure 2: Dataset após o cálculo dos scores junto com a coluna de vencedores.

Valores de Loadings

Após ter os scores foi feito também uma tabela com os valores de loading, sendo uma linha para cada um dos 3 descritores, ou seja, 'GoalsScored', 'QualifiedTeams' e 'MatchesPlayed'

Variância Explicada

Para explicar a contribuição de cada um dos 2 componentes principais na variância do dataset foi feito o calculo da variância explicada. Os resultados para o primeiro e segundo componente foram de aproximadamente 93% e 5%.

Variância Cumulativa

Com os valores dos dois componentes, 93% e 5%, foi feita uma soma cumulativa, resultando numa matriz de aproximadamente 93% e 98%. Com isso é possível concluir que o **PC1 e o PC2 representam cerca de 98% da variância**.

3 Visualização

Nessa etapa foram usadas as bibliotecas **Numpy**, para manipular as matrizes e para usar as funções matemáticas, e o **Plotly Express**, com o objetivo de plotar os gráficos.

Variância Explicada e Cumulativa

Foi feito uma junção de dois gráficos, o gráfico de barras (trace 1) representa a variância explicada dos dois componentes e o gráfico de linha (trace 2) representa a variância acumulada.

Figure 3: Multi plot graph com a variância explicada e cumulativa em decimal.

Mapa de Calor

Também foi feito um gráfico do tipo Heatmap para entender quanto os componentes principais representam em cada coluna.

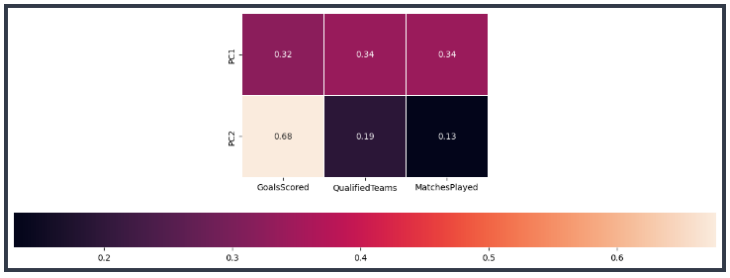

Figure 4: Heatmap dos componentes principais.

É possível concluir que o PC1 consiste em 32% da coluna GoalsScored e 34% das colunas QualifiedTeams e MatchesPlayed. O PC2 é composto por 68% dos GoalsScored, 19% dos QualifiedTeams e 13% dos MatchesPlayed.

Dados com os Componentes Principais

Por último, esses dados foram plotados em um gráfico de pontos, com o eixo X sendo o primeiro componente principal e o eixo Y sendo o segundo componente principal. Com isso, é demostrado como as dimensões desse dataset foram transformados, passando de três para apenas duas.

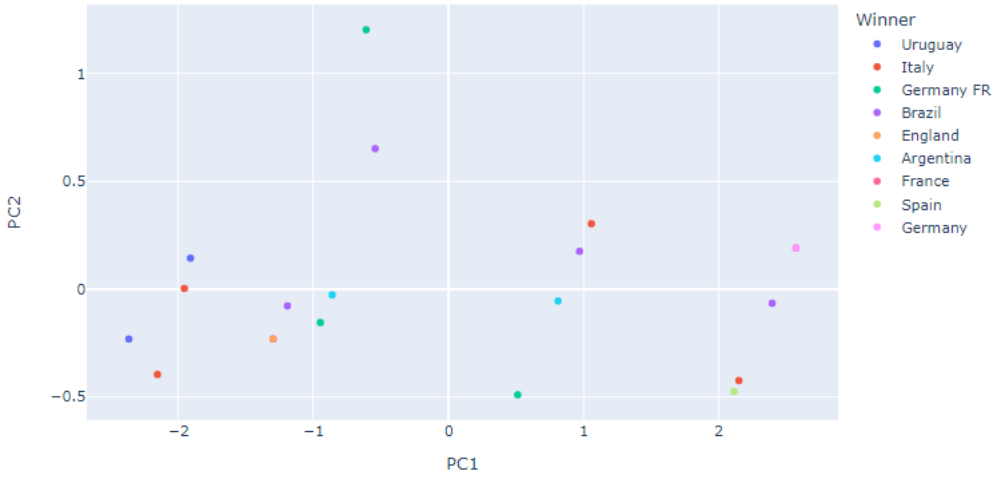

Figure 5: Gráfico final com os componentes principais após o tratamento com o PCA.

4 Referências

- Data Professor - Machine Learning in Python: Principal Component Analysis (PCA) for Handling High-Dimensional Data:
www.youtube.com/watch?v=oiusrJ0btwA

- Casey Cheng - Principal Component Analysis (PCA) Explained Visually with Zero Math:
towardsdatascience.com/principal-component-analysis-pca-explained-visually-with-zero-math-1cbf392b9e7d

- Ecologia Descomplicada - Aula PCA simplificada:
www.youtube.com/watch?v=KqZAC4jyJKc

- Eduardo - Ciência dos Dados - ENTENDENDO DE VEZ O QUE É PCA - PRINCIPAL COMPONENT ANALYSIS:
www.youtube.com/watch?v=p4bvCFygfW0

- Gayathri Siva - PCA — Principal Component Analysis Explained with Python Example:
gayathri-siva.medium.com/pca-principal-component-analysis-explained-with-python-example-e403f9fef52b

Repositório do Github:
https://github.com/StefanyFernandes675/PCA-worldCup
Dataset utilizado:
https://www.kaggle.com/datasets/abecklas/fifa-world-cup.