Atividade Prática 3: Apliacação do PCA

Hector José Rodrigues Salgueiros

[1]Sistemas de Informação – Universidade Federal do Piauí (UFPI)
Caixa Postal 64.600-000 – Picos – PI – Brazil

hectorsalg@ufpi.edu.br

Abstract. *This report presents a detailed analysis of the Iris database, a valuable source of information about plants in the genus Iris. Understanding botanical diversity is essential in a variety of scientific fields, and the Iris database plays a key role in providing taxonomic, morphological and genetic information about these plants.*

Resumo. *Este relatório apresenta uma análise detalhada da base de dados Iris, uma fonte valiosa de informações sobre as plantas do gênero Iris. A compreensão da diversidade botânica é essencial em uma variedade de campos científicos, e a base de dados Iris desempenha um papel fundamental ao fornecer informações taxonômicas, morfológicas e genéticas sobre essas plantas.*

1. Introdução

A compreensão da diversidade botânica é crucial para uma série de aplicações científicas e práticas, desde a taxonomia até a ecologia e a biotecnologia. Uma ferramenta essencial nesse domínio é a base de dados IRIS (Internet Resource for Iris and Plant Systematic), uma compilação abrangente de informações sobre as plantas do gênero Iris. A base de dados IRIS não só fornece detalhes taxonômicos e morfológicos sobre várias espécies de Iris, mas também serve como um recurso valioso para explorar a evolução, distribuição geográfica e características genéticas dessas plantas fascinantes. Desde sua criação, a base de dados IRIS tem sido uma fonte fundamental para pesquisadores, botânicos e entusiastas de plantas em todo o mundo, fornecendo dados acessíveis e atualizados que impulsionam uma variedade de estudos e descobertas científicas.

O conjunto de dados Iris é reconhecido como um dos conjuntos de dados mais emblemáticos no campo da estatística e aprendizado de máquina. Sua popularidade decorre da sua simplicidade e relevância para uma variedade de problemas de classificação. Neste relatório, exploraremos o conjunto de dados Iris utilizando a técnica de Análise de Componentes Principais (PCA), com o objetivo de visualizar as relações entre suas características e interpretar essas relações.

2. Metodologia

Nesta seção, será descrito detalhadamente os procedimentos e abordagens adotados para conduzir o presente estudo. A metodologia empregada é fundamental para garantir a validade e confiabilidade dos resultados obtidos.

2.1. Análise Descritiva

Após carregar os dados, é essencial realizar uma análise descritiva para compreender as principais características das variáveis. O conjunto de dados Iris é composto por quatro variáveis (também conhecidas como características ou atributos) que descrevem as medidas das sépalas e pétalas de três espécies de íris. Aqui está uma descrição de cada uma dessas variáveis:

Comprimento da Sépala (sepal length): Esta variável representa o comprimento da sépala, a parte exterior e protetora da flor, medida em centímetros. O comprimento da sépala é uma das características morfológicas importantes usadas para distinguir entre diferentes espécies de íris.

Largura da Sépala (sepal width): Esta variável descreve a largura da sépala, medida em centímetros. Assim como o comprimento da sépala, a largura da sépala é uma medida morfológica fundamental que pode ajudar a diferenciar entre as espécies de íris.

Comprimento da Pétala (petal length): Esta variável indica o comprimento da pétala, a parte interna e colorida da flor, medida em centímetros. O comprimento da pétala é uma característica chave na determinação da forma e tamanho das flores de íris e é útil na classificação das diferentes espécies.

Largura da Pétala (petal width): Por fim, esta variável descreve a largura da pétala, medida em centímetros. A largura da pétala, juntamente com seu comprimento, é uma característica importante na caracterização da forma das flores de íris e é utilizada para diferenciar entre as espécies.

A análise descritiva também inclui o cálculo de medidas de tendência central (média e mediana) e medidas de dispersão (desvio padrão e variância) para as principais variáveis. A seguir na Tabela 1 será detalhado:

Table 1. Medidas de tendência central e medidas de dispersão

index	sepal length (cm)	sepal width (cm)	petal length (cm)	petal width (cm)
count	150.0	150.0	150.0	150.0
mean	5.84	3.05	3.75	1.19
std	0.82	0.43	1.76	0.76
min	4.3	2.0	1.0	0.1
25%	5.1	2.8	1.6	0.3
50%	5.8	3.0	4.35	1.3
75%	6.4	3.3	5.1	1.8
max	7.9	4.4	6.9	2.5

2.2. Normalização

A normalização de dados é uma etapa fundamental no pré-processamento de dados antes de aplicar técnicas de análise ou modelagem. Neste artigo, o processo de normalização de dados na base de dados Iris utilizado foi classe StandardScaler da biblioteca Python scikit-learn [Pedregosa et al. 2011].

2.3. PCA

A Análise de Componentes Principais (PCA) é uma técnica poderosa de redução de dimensionalidade que é frequentemente utilizada para explorar a estrutura subjacente de conjuntos de dados complexos. Neste artigo, o processo de aplicação do PCA foi aplicado base de dados Iris, utilizando a implementação da biblioteca scikit-learn em Python [Pedregosa et al. 2011].

3. Resultados

Nesta seção, serão apresentados os resultados obtidos após uma análise minuciosa dos dados coletados da base de dados Iris. As métricas e gráficos aqui exibidos visam fornecer uma compreensão detalhada da distribuição e dispersão dos dados em cada variável após aplicar o PCA.

O gráfico de cotovelo é comumente usado na Análise de Componentes Principais (PCA) para determinar o número ótimo de componentes principais a serem retidos na análise de dados. A Figura 1 contem o gráfico de cotovelo das variáveis da base Iris.

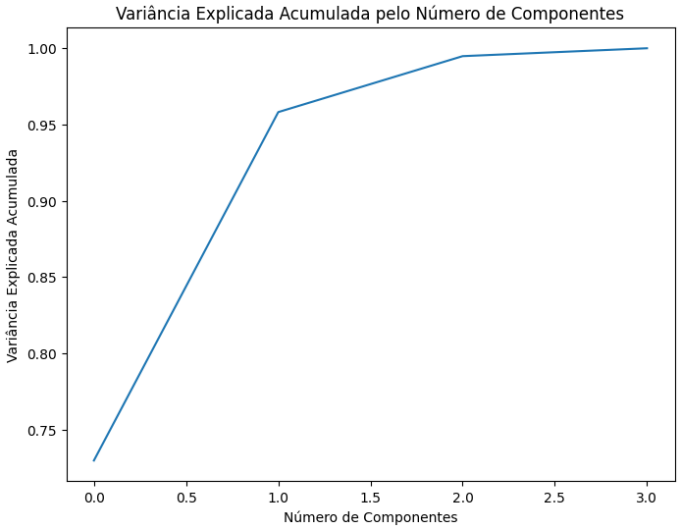

Figure 1. Gráfico de cotovelo

O ponto de cotovelo é crucial para decidir quantos componentes principais manter. Nesse gráfico o ponto fica entre 1 e 2 no eixo x, depois disso a linha começa a se estabilizar. Isso sugere que reter mais de 1 componentes principais não contribui significativamente para explicar mais variância.

Na Figura 2 mostra um gráfico de dispersão com a visualização dos dois primeiros componentes principais após a aplicação da técnica de Análise de Componentes Principais (PCA).

O PCA ajuda a simplificar a complexidade dos dados ao reduzir o número de variáveis, mas ainda assim preserva as relações significativas entre os pontos de dados. No gráfico da Figura 2, podemos ver claramente a separação entre as três espécies de flores, o que indica que a PCA foi eficaz na distinção entre elas.

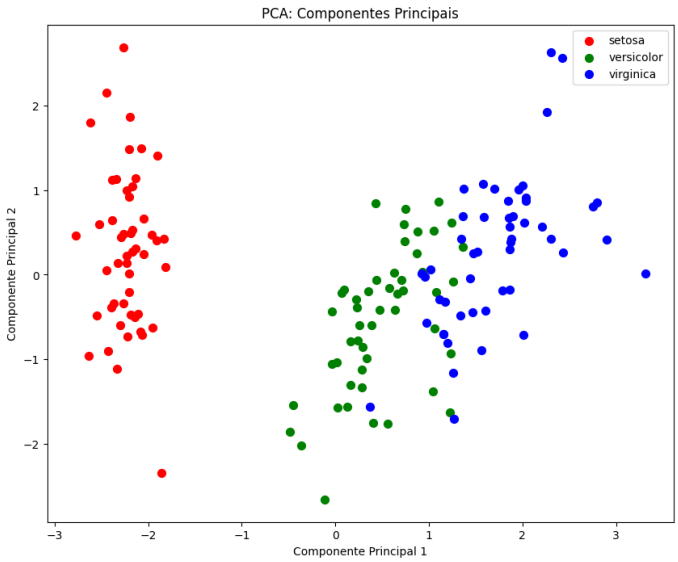

Figure 2. Gráfico de dispersão

4. Conclusão

A análise minuciosa dos dados da base de dados Iris revelou insights valiosos sobre a distribuição e dispersão das medidas das sépalas e pétalas das três espécies de íris. Através do uso do gráfico de cotovelo na Análise de Componentes Principais (PCA), identificamos o ponto ideal de retenção de componentes principais, sugerindo que manter mais de duas ou três componentes principais não contribuiria significativamente para explicar a variância dos dados. Este resultado enfatiza a importância de encontrar um equilíbrio entre a simplificação da dimensionalidade dos dados e a preservação da variabilidade significativa.

A visualização dos dois primeiros componentes principais através do gráfico de dispersão Figura 2 destacou a eficácia da PCA na distinção entre as espécies de flores de íris. A clara separação entre as três espécies demonstra como o PCA preservou as relações significativas entre os pontos de dados, mesmo após a redução da dimensionalidade. Esses resultados corroboram a utilidade do PCA como uma ferramenta poderosa para a análise exploratória de conjuntos de dados complexos, fornecendo uma compreensão mais profunda da estrutura subjacente dos dados e dos padrões que neles podem surgir.

References

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.