Relatório: Análise de Componentes Principais (PCA) na Base de Dados sobre o perfil epidemiológico da Tuberculose entre 2020 à 2023

2º Ciclo – Ciência de Dados

Eduardo Miguel Ribeiro Cordeiro

Iris Durante Alvim do Nascimento

Waleska Mayara Silva Reis

Novembro, 2024

Introdução

Este relatório, apresentará uma análise sobre a incidência de casos de tuberculose, usando a Análise de Componentes Principais (PCA). Essa ferramenta é essencial, pois simplifica e interpreta conjuntos de dados multidimensionais, além de facilitar a visualização dos dados. A utilização do PCA, na base de dados, permite não só a identificação dos aspectos recorrentes da doença, como ainda fornece uma visão ampla do perfil epidemiológico da TB.

Descrição da Base de Dados

Os dados foram obtidos pelo site do Departamento de Informática do Sistema Único de Saúde (DataSUS), selecionando nas abas os anos de 2020 a 2023, de situação encerrada e as características de perfil epidemiológico a serem estudadas. Foi feita pesquisa por “tuberculose” e escolhido o arquivo “Dados de Tuberculose”. Esta base de dados traz uma abordagem quantitativa e descritiva, analisando a incidência e os determinantes da tuberculose.

Dimensão dos Dados

- O dataset é formado por 36 entradas e 12 colunas;
- As 12 colunas são nomeadas por: Ano Diagnóstico; Ign/Branco; Cura; Abandono; Óbito por tuberculose; Óbito por outras causas; Transferência; TB-DR; Mudança de Esquema; Falência; Abandono Primário; Total;
- O data type é formado por: 11 colunas float64 e 1 object (que é referente a coluna do Ano Diagnóstico, que vem a ser retirada após o tratamento);
- Usando um total de memória de 3.5+ KB.

Matriz de Covariância

A matriz de covariância resume as relações entre as variáveis do dataset, mostrando como elas covariam juntas. A partir dessa matriz, o PCA identifica os componentes principais, que são combinações lineares das variáveis originais que melhor explicam a variabilidade nos dados, destacando as informações mais importantes. A seguir está a Matriz de Covariância.

```
Matriz de Covariância:
[[1.02857143 0.95613037 0.96580982 0.97132713 0.95373032 0.98671486
  0.96587926 0.96982865 0.94327444 0.97541644 0.97059502]
 [0.95613037 1.02857143 1.02807502 1.02730345 1.02842717 1.02391583
  1.02755671 1.02777521 1.02763706 1.02680797 1.02772383]
 [0.96580982 1.02807502 1.02857143 1.02839344 1.02764665 1.02633693
  1.02839138 1.02848723 1.0263063  1.02794409 1.0284222 ]
 [0.97132713 1.02730345 1.02839344 1.02857143 1.02670545 1.027336
  1.02840111 1.02844826 1.02501377 1.02817338 1.02837966]
 [0.95373032 1.02842717 1.02764665 1.02670545 1.02857143 1.02301453
  1.02702874 1.02732583 1.02807548 1.02601326 1.02729679]
 [0.98671486 1.02391583 1.02633693 1.027336   1.02301453 1.02857143
  1.02639906 1.02696821 1.02001593 1.02757351 1.02701199]
 [0.96587926 1.02755671 1.02839138 1.02840111 1.02702874 1.02639906
  1.02857143 1.02821346 1.02588127 1.02757729 1.02806766]
 [0.96982865 1.02777521 1.02848723 1.02844826 1.02732583 1.02696821
  1.02821346 1.02857143 1.02563279 1.02824066 1.02855875]
 [0.94327444 1.02763706 1.0263063  1.02501377 1.02807548 1.02001593
  1.02588127 1.02563279 1.02857143 1.02356234 1.02550508]
 [0.97541644 1.02680797 1.02794409 1.02817338 1.02601326 1.02757351
  1.02757729 1.02824066 1.02356234 1.02857143 1.02827697]
 [0.97059502 1.02772383 1.0284222  1.02837966 1.02729679 1.02701199
  1.02806766 1.02855875 1.02550508 1.02827697 1.02857143]]
```

Autovalores e Autovetores

Os autovalores medem quanto da variabilidade total dos dados é capturada por cada componente principal, ou seja, classificam os componentes principais em ordem de importância. Já os autovetores são direções específicas no espaço multidimensional dos dados que apontam para onde há maior variação, ou seja, ajudam a criar os componentes principais no espaço dos dados.

Autovalores:

```
Autovalores:
[ 1.11898582e+01  1.21598555e-01  1.86833757e-03  9.60619274e-04
  6.47565620e-15  4.66000919e-15  2.12835770e-15 -1.15286522e-15
 -2.10699543e-15 -4.44835167e-15 -4.81705323e-15]
```

Autovetores (a matriz dos autovetores ficou muito grande, por isso será mostrado apenas parte dela):

```
Autovetores (primeiros 5 vetores):
[[-0.28786686  0.91225648  0.22637927 -0.08074324  0.0730605 ]
 [-0.30264642 -0.16947318  0.17099471  0.30763954  0.47495894]
 [-0.30302511 -0.09223085 -0.14060751  0.01141097  0.08139174]
 [-0.30311719 -0.04461777 -0.31384428 -0.14716024 -0.17117867]
 [-0.30248473 -0.18779322  0.48621666  0.05406277 -0.73423078]
 [-0.30298354  0.09790119 -0.29555779 -0.16023499 -0.22495541]
 [-0.30295964 -0.08874929 -0.465967   -0.39771406  0.02074849]
 [-0.30295964 -0.08874929 -0.465967   -0.39771406  0.02074849]
 [-0.30295964 -0.08874929 -0.465967   -0.39771406  0.02074849]
 [-0.30295964 -0.08874929 -0.465967   -0.39771406  0.02074849]
 [-0.30311904 -0.05914428 -0.01676051  0.09379053  0.13081371]
 [-0.30295964 -0.08874929 -0.465967   -0.39771406  0.02074849]
 [-0.30311904 -0.05914428 -0.01676051  0.09379053  0.13081371]
 [-0.30295964 -0.08874929 -0.465967   -0.39771406  0.02074849]
```

Cinco maiores autovetores:

```
Dados transformados (primeiras 5 amostras):
[[ -2.83325585  -0.628443  ]
 [ -3.24536472  -0.70142911]
 [ -3.81388889  -0.51134413]
 [ -4.50169169   1.75974313]
 [-17.47867283  -0.09893164]]
```

Resultados Obtidos

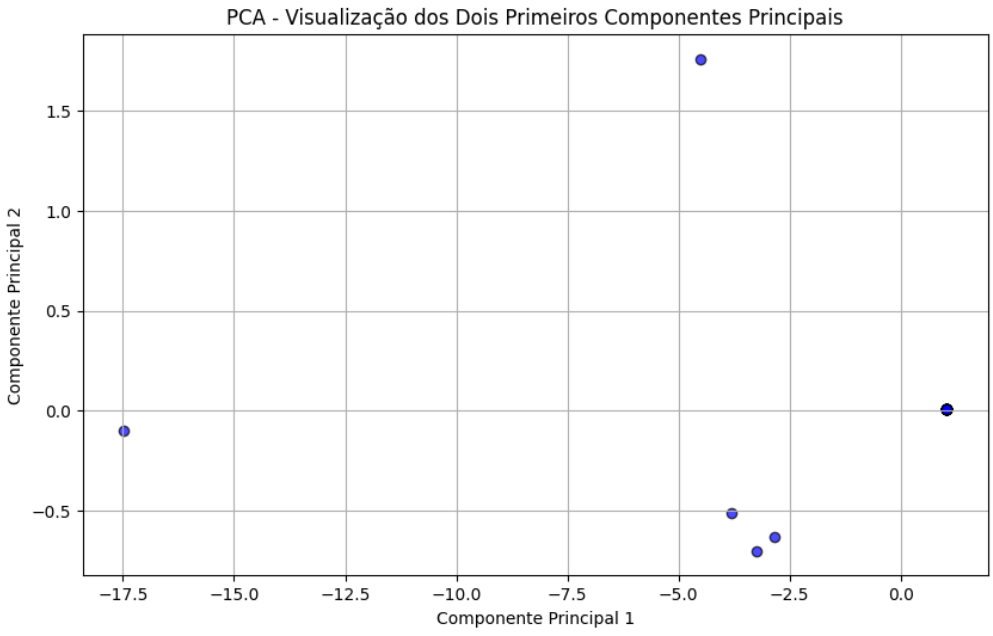

Figura 1: Elaborada no Visual Studio Code pelos autores

Através da análise desse PCA é possível perceber que o primeiro componente principal (PC1) possui um maior foco na variância deixando o segundo componente principal (PC2) com pouca ênfase. A alta concentração de variância no PC1 indica que ele captura quase toda a estrutura dos dados. O PC2 representa variações secundárias ou menos relevantes.

Os pesos de cada variável para os dois componentes principais ajudam a interpretar o que cada um representa. Com base nos resultados, é possível inferir que o PC1 deve estar relacionado a variáveis que afetam diretamente o total de casos de tuberculose, como:

- "Cura", que reflete desfechos positivos;
- "Óbito por tuberculose" e "Abandono", que representam desfechos negativos;
- "Total", que tende a ser altamente correlacionado com outras variáveis.

Já o PC2 tende a estar relacionado a variáveis com menor impacto nos dados gerais, mas que ainda fornecem informações específicas, como:

- "Mudança de Esquema" e "Falência", que são categorias menos frequentes.

Em suma, é possível, observando o gráfico bidimensional baseado nos dois primeiros componentes principais, entender que os registros com valores mais altos no PC1 indicam anos com maior impacto em termos de casos totais, curas ou mortes e que registros separados ao longo do PC2 podem sugerir diferenças menores em variáveis secundárias.

Conclusão

Concluindo este relatório, é possível afirmar que a aplicação da Análise de Componentes Principais (PCA), em conjunto com a redução de dimensionalidade e a representação visual, contribui diretamente para a interpretação dos dados e para a tomada de decisões no contexto do perfil

epidemiológico da doença analisada. Dessa forma, com os dados analisados, suponha-se que agrupamentos identificados no gráfico podem indicar padrões regionais, mudanças em políticas públicas ou impactos de programas específicos e que anos com valores extremos no PC1 devem ser investigados mais profundamente para identificar os fatores que contribuíram para o aumento ou redução de casos.

Referências

Link do código-fonte no GitHub:
https://github.com/EduMiguel013/PCA

Link para base de dados disponível no site do DataSUS:
http://tabnet.datasus.gov.br/cgi/tabcgi.exe?sinannet/cnv/tubercbr.def