UNIVERSIDADE FEDERAL DE SÃO PAULO
INSTITUTO DE CIÊNCIA E TECNOLOGIA
BACHARELADO EM CIÊNCIA E TECNOLOGIA

Projeto 04:

PCA - Principal Component Analysis

Nome: Willian Dihanster Gomes de Oliveira **RA:** 112269

SÃO JOSÉ DOS CAMPOS
2018

Projeto 4:

1. Selecionar pelo menos 3 datasets do repositório da UCI (selecionar datasets com dimensão superior a 4 (>4) [Ou utilizar o dataset MNIST]
2. Utilizando tanto a técnica PCA ou Rede PCA:
    1. Encontrar os componentes principais
    2. Plotar os exemplos utilizando as duas primeiras componentes principais
    3. Estudar qual a máxima redução de dimensionalidade possível preservando 95% da variância original

Base 1: *Digits Dataset*

A base de dados *Digits Dataset* é composta por imagens de dígitos (de 0 a 9) escritos a mão. A base é composta por 1797, sendo aproximadamente 180 de cada classe (10 classes ,ao total). Cada imagem é 8x8. Sendo assim, cada exemplo é 64-dimensional.

Os Componentes Principais podem ser conferidos na Figura 1, a seguir. Na figura 2 é possível visualizar um plot das duas primeiras componentes principais.

```
[[ 4.52876704e-17 -1.73094664e-02 -2.23428842e-01 -1.35913312e-01
  -3.30323134e-02 -9.66340802e-02 -8.32943484e-03  2.26900097e-03
  -3.20516519e-04 -1.19308908e-01 -2.44451675e-01  1.48512739e-01
  -4.67319593e-02 -2.17740750e-01 -1.48136756e-02  4.47779487e-03
  -4.94136550e-05 -7.95419386e-02  8.33951584e-02  2.15915341e-01
  -1.72126794e-01 -1.63712104e-01  2.86444399e-02  4.23251753e-03
   9.85488548e-05  6.42319153e-02  2.54093314e-01 -3.56771079e-02
  -2.09462547e-01 -4.31311515e-02  5.13118559e-02  2.13422712e-04
   0.00000000e+00  1.59950887e-01  3.68690757e-01  1.64406806e-01
   8.52007961e-02  3.72982904e-02  2.15866943e-02  0.00000000e+00
   1.28865599e-03  1.06945298e-01  3.03067471e-01  2.47813045e-01
   2.09637296e-01  1.22325260e-02 -3.69458526e-02  1.61485003e-03
   6.93023632e-04 -8.35143747e-03 -5.58598851e-02  9.30534101e-02
   1.07387711e-01 -1.37734577e-01 -6.32879543e-02  9.61669077e-04
   9.55080317e-06 -1.40786847e-02 -2.35675495e-01 -1.41225592e-01
  -9.15964541e-03 -8.94184740e-02 -3.65977169e-02 -1.14684983e-02]
 [-3.04636720e-17 -1.01064546e-02 -4.90849086e-02 -9.43337660e-03
  -5.36015688e-02 -1.17755314e-01 -6.21281760e-02 -7.93574553e-03
  -1.63216256e-04 -2.10167007e-02  6.03485684e-02 -5.33769735e-03
  -9.19769118e-02 -5.19210490e-02 -5.89354657e-02 -3.33283397e-03
  -4.22872221e-05  3.62458536e-02  1.98257331e-01 -4.86386556e-02
  -2.25574897e-01 -4.50541631e-03  2.67696767e-02 -2.08735626e-04
  -5.66234071e-05  7.71235109e-02  1.88447114e-01 -1.37952514e-01
  -2.61042790e-01  4.98350741e-02  6.51113854e-02  4.03200469e-05
  -0.00000000e+00  8.81559890e-02  8.71737701e-02 -2.70860172e-01
  -2.85291804e-01  1.66461585e-01  1.27860543e-01 -0.00000000e+00
   2.89440144e-04  5.08304808e-02  1.30274454e-01 -2.68906459e-01
  -3.01575535e-01  2.40259068e-01  2.17555550e-01  1.32726072e-03
   2.86742956e-04  1.05548275e-02  1.53370694e-01 -1.19535160e-01
  -9.72507985e-02  2.85869553e-01  1.48776453e-01  5.42292092e-04
  -3.34028111e-05 -1.00791146e-02 -7.02723944e-02  1.71108025e-02
   1.94296389e-01  1.76697126e-01  1.94547126e-02 -6.69693703e-03]]
```

Figura 1: Componentes Principais para a base Digits.

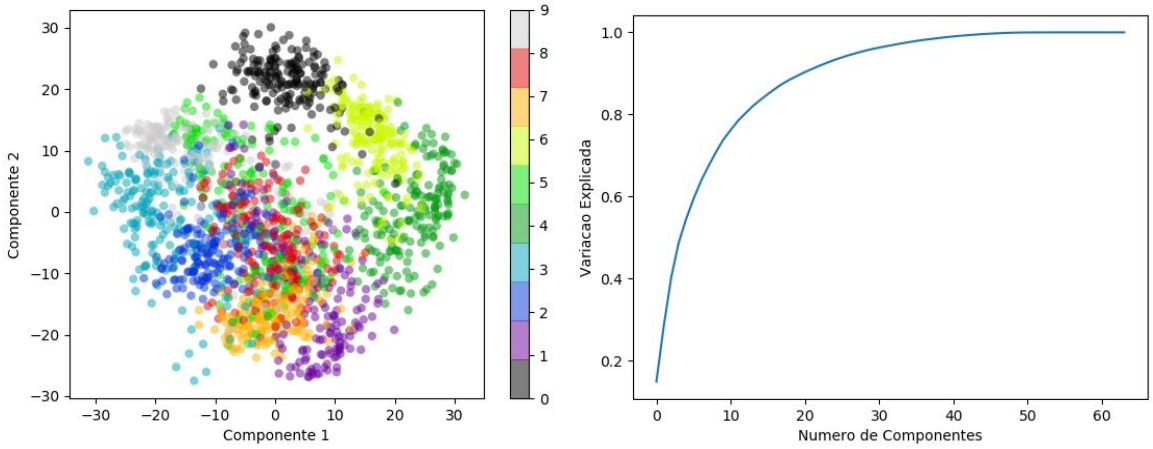

Figura 2: Plot dos 2 Componentes Principais e do Número de Componentes pela Variação Explicada

Observando o gráfico da Figura 2, pode se notar uma separação dos dados, com uma certa sobreposição de classes, como alguns dados mais centrais.
Pelo gráfico da Variação Explicada podemos ver que com 2 componentes acabamos perdendo um parte da informação do dado original, pois com 2 componentes, representamos apenas cerca de 30% da variação.
Para conseguirmos algo em torno de 95% de variação explicada, seria necessário trabalhar com mais ou menos, 30 componentes, o que é metade da dimensão original.

Base 2: *Wine Dataset*
É uma base dados com características do vinhos, com a intenção de predizer a origem dos vinhos. A base é composta por 178 exemplos, com 3 classes (0, 1, 2), sendo cada exemplo representado por 13 atributos (13-dimensionalidade).
Na Figura 3, temos os componentes principais encontrados. E na Figura 4, o plot das 2 componentes principais e da variação explicada.

```
[[ 1.65926472e-03 -6.81015556e-04  1.94905742e-04 -4.67130058e-03
   1.78680075e-02  9.89829680e-04  1.56728830e-03 -1.23086662e-04
   6.00607792e-04  2.32714319e-03  1.71380037e-04  7.04931645e-04
   9.99822937e-01]
 [ 1.20340617e-03  2.15498184e-03  4.59369254e-03  2.64503930e-02
   9.99344186e-01  8.77962152e-04 -5.18507284e-05 -1.35447892e-03
   5.00440040e-03  1.51003530e-02 -7.62673115e-04 -3.49536431e-03
  -1.77738095e-02]]
```

Figura 3: Componentes Principais para a base de dados *Wine Dataset*.

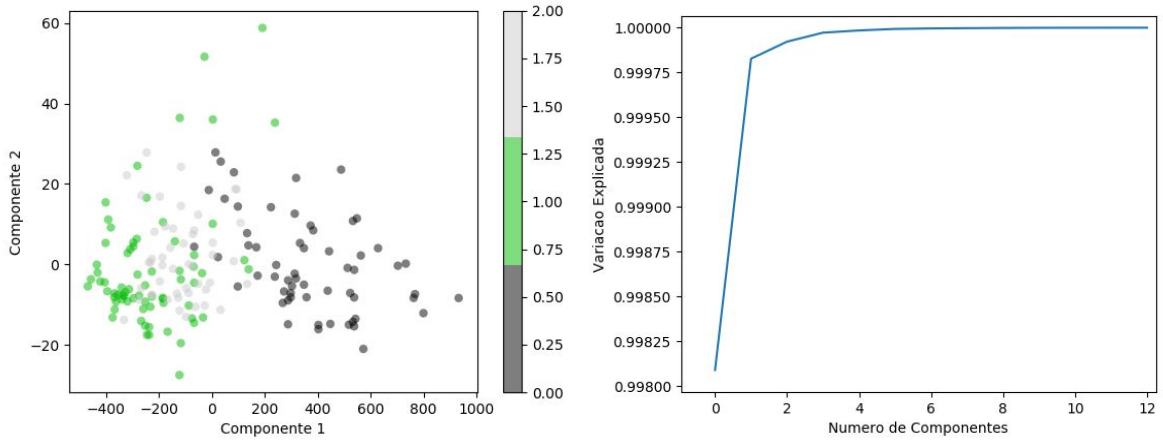

Figura 4: Plot das 2 Componentes Principais e do Número de Componentes pela Variação Explicada.

Pode-se observar também uma certa mistura entre as classes no primeiro plot. Um outro fato, é que com 1 componente já possui 99,8% de variância explicada. Com 2 componentes quase 99,9%.

Base 3: *Breast Cancer*

A *Breast Cancer* é uma base dados sobre Câncer de Mama. É composta por 569 exemplos, tendo 2 classes e 30 atributos (30-dimensionalidade).

Na Figura 5, temos os componentes principais encontrados. E na Figura 6, o plot das 2 componentes principais e da variação explicada, para a base.

```
[[ 5.08623202e-03  2.19657026e-03  3.50763298e-02  5.16826469e-01
   4.23694535e-06  4.05260047e-05  8.19399539e-05  4.77807775e-05
   7.07804332e-06 -2.62155251e-06  3.13742507e-04 -6.50984008e-05
   2.23634150e-03  5.57271669e-02 -8.05646029e-07  5.51918197e-06
   8.87094462e-06  3.27915009e-06 -1.24101836e-06 -8.54530832e-08
   7.15473257e-03  3.06736622e-03  4.94576447e-02  8.52063392e-01
   6.42005481e-06  1.01275937e-04  1.68928625e-04  7.36658178e-05
   1.78986262e-05  1.61356159e-06]
 [ 9.28705650e-03 -2.88160658e-03  6.27480827e-02  8.51823720e-01
  -1.48194356e-05 -2.68862249e-06  7.51419574e-05  4.63501038e-05
  -2.52430431e-05 -1.61197148e-05 -5.38692831e-05  3.48370414e-04
   8.19640791e-04  7.51112451e-03  1.49438131e-06  1.27357957e-05
   2.86921009e-05  9.36007477e-06  1.22647432e-05  2.89683790e-07
  -5.68673345e-04 -1.32152605e-02 -1.85961117e-04 -5.19742358e-01
  -7.68565692e-05 -2.56104144e-04 -1.75471479e-04 -3.05051743e-05
  -1.57042845e-04 -5.53071662e-05]]
```

Figura 5: Componentes Principais para a base *Breast Cancer*.

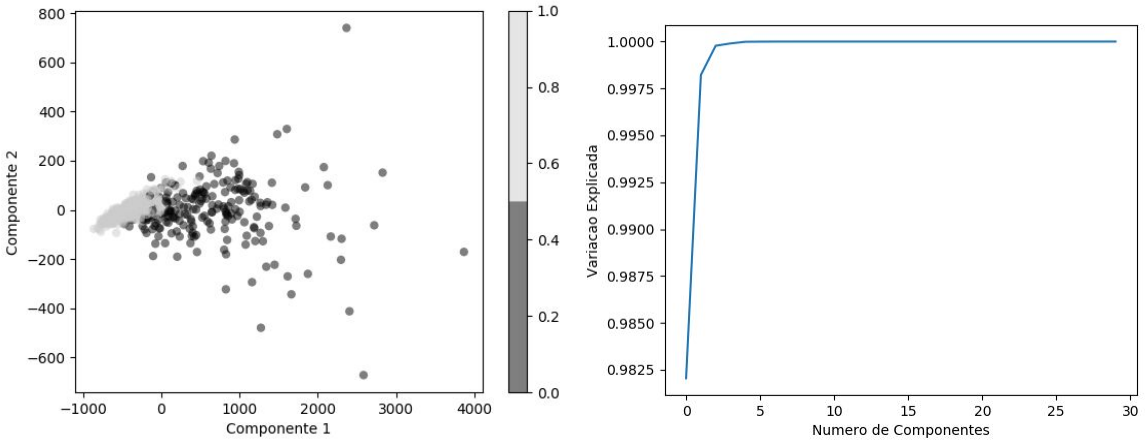

Figura 6: Plot das 2 Componentes Principais e do Número de Componentes pela Variação Explicada.

Nesse exemplo, com o plot das componentes principais, é possível observar uma grande sobreposição das classes diferentes. Além disso, pelo gráfico do número de componentes pela variação explicada, é possível notar que com 1-2 já se representa quase 100% da variância.

Conclusões

Sendo assim, pode-se concluir a eficácia do método/rede PCA na redução de dimensionalidade e visualização. Pois assim, é possível observar e ter uma visão (ou tentar) dos exemplos e sua separação no espaço, gerando possíveis interpretações dos dados.