Relatório: Multivariada 2

Adriel Martins[†*]
[†] UFPE - Departamento de Estatística

Sumário

1. Introdução

Este trabalho consiste em estudar 3 famílias de técnincas estatísticas: decomposição, agrupamento e classificacação/análise de discriminante.

1.1. **Decomposição.**

1.1.1. *PCA*. PCA consiste em uma decomposição de tal forma que nosso objetivo seja comprimir a *informação* de um conjunto de dados. Queremos, a partir de p variáveis, gerar outras p variáveis de tal forma que as primeiras dessas variáveis transformadas possam conter muita informação, e assim, usaremos apenas $q < p$ variáveis. Em geral, $q = 2$.

Além disso, iremos definir a informação que queremos preservar como a variabilidade dos dados. Além disso, desejamos que essas novas variáveis, as quais chamaremos de *Principal Components* (PC), sejam independentes uma das outras.

[*]adriel.martins@ufpe.br

1.1.2. *Análise Fatorial.* Essa técnica tem o foco específico de encontrar índices mediante muitas varíaveis. A técnica é composta de misturas de modelos ou distribuições. Assumimos que existe *fatores* latentes (ou desconhecidos) que estão linearmente relacionados com as variáveis atuais. Esses fatores são menores em quantidade do que as variáveis atuais.

As variáveis observadas são modeladas como uma combinação linear de fatores e termos de erro. O fator está associado à múltiplas variáveis observadas, que têm padrões comuns de respostas. Cada fator explica uma determinada quantidade de variância nas variáveis observadas.

Além disso, também podemos diminuir a quantidade de variáveis.

1.2. **Agrupamento.** Adentrando agora em agrupamento. Iremos realizar 3 técnicas: K-Means, PAM, Hierárquico. Em outro conjunto de dados.

1.2.1. *K-Means.* Consiste em declarar uma quantidade k de grupos em um plano p-dimensional. Com cada criação do grupo, irá se obter um chute inicial para o seu respectivo centróide com coordenadas de dimensão p. Cada ponto p-dimensional, cada observação do conjunto de dados, terá sua distância calculada em relação a cada um dos centróides. Aquele centróide que for mais perto da observação, iremos dizer que ele pertence aquele grupo. Depois disto, calcularemos a média dos valores em cada dimensão, obtendo assim um vetor p-dimensional para grupo. Esta média irá virar o novo centróide do grupo. Baseado em algum critério de parada, iremos parar o algoritmo.

Em geral, a distância utilizada é a distância euclidiana ao quadrado.

1.2.2. *Partitioning Around Medoids (PAM).* Pode se chamar uma versão modificada do K-Means num aspecto de que embora iremos declarar centróides, a maneira de buscar o centróide ótimo é diferente. Não será a simples média do novo grupo, mas será o indivíduo que diminui o coeficiente médio de dissimilaridade, ou seja a distância entre as observações. Em geral, usa-se a medida de "silhouette"para medir essa dissimilaridade entre os grupos.

1.2.3. *Agrupamento Hierárquico Aglomerativo.* A idéia deste agrupamento começa de que cada observação é um grupo e depois por uma regra de associação baseado em uma medida ou matriz de distância, iremos decidir na "associação"ou criação de novos grupos baseado em quão mais perto eles estão um dos outros. A decisão por "k"grupos é arbitrário, onde decidimos por um corte de nível na Hierarquia proposta.

1.2.4. *Agrupamento por Misturas Gaussianas.* A idéia dele é de entender que as observações com p features que temos são amostras da mistura de K variáveis Gaussianas Multivariadas sendo que cada uma delas é de tamanho p. Em especial, essa mistura tem a estrutura de soma ponderada das Gaussianas Mult., onde também estimaremos esses pesos.

O método de estimação dos parâmetros das somas das K Gaussianas Multivariadas se consiste em definir uma função de log-verossimilhança entre duas variáveis as variáveis observadas e a probabilidade ou ponderação de pertencer a uma variável multivariada Gaussiana latente. Definido isso utiliza-se o algoritmo de otimização chamado *Expectation Maximization*, que consiste em procurar os parâmetros que maximizam a log-verossimilhança iterativamente.

1.3. **Classificação.** Por último, temos a família de técnicas de Classificação ou Análise de Discriminante. Concentraremos em duas técnicas específicas: Linear Discriminant Analysis (LDA) e K-Nearest Neighbors (K-NN).

1.3.1. *LDA.* LDA é um tipo de classificação criar pela a utilização da regra de Bayes para formular a a Probabilidade de má classificação em função da probabilidade de classificar em grupos diferentes aquilo que verdadeiramente pertence a um determinado grupo. Uma vez definido essa probabilidade de má classificação, minimiza-se esta utilizando um lema probabilístico, chegando se ao seguinte resultado:

Dentre dois grupos, alocasse x_0 ao grupo G_1, se:

$$\frac{f_1}{f_2} > \frac{\pi_2}{\pi_1}$$

Podendo conter muitas variações dessa fórmula inicial.

1.3.2. *K-NN.* A montagem deste modelo é extremamente intuitiva. Primeiramente, define-se uma medida de distância entre os dados. Depois, já sabendo quais são os nossos grupos nos dados, ao chegar uma nova observação calculamos a distância $D(x_0, X)$. Pegaremos os primeiros K vizinhos dos dados, o grupo que tiver mais quantidade de membros receberá essa nova observação.

2. Decomposição Estatística

Aplicaremos técnicas de Decomposição Estatística nos dados do LANDSAT 8, ou seja, dados de imagens advindas de satélites. Os dados de imagem de satélite podem ser pensados em um array de $(N1, N2, p)$, tais que p são os canais da imagem, e $N1$ e $N2$ são as coordenadas de cada pixel.

Na Figura 1 podemos ver as referências de cada canal. Vejamos que há 11 canais, porém um desses canais tem o dobro de resolução. Isso significa que esse canal tem $(2 * N1, 2 * N2)$ pixels. Isso complica a nossa análise e portanto iremos retirar esse canal.

A imagem que obtemos é a que se encontra na Figura 2.

Iremos cortar a imagem por questões de perfomance ficando com a imagem da Figura 3.

Comecemos por decompor a imagem de maneira estatística. As duas técnicas que iremos aplicar são *Principal Component Analysis* (PCA) e Análise Fatorial.

Antes de iniciarmos, vejamos o histograma de cada canal para termos idéia da natureza dos dados e de como eles se comportam.

Landsat 8		
Band Name	Bandwidth (µm)	Resolution (m)
Band 1 Coastal	0.43 – 0.45	30
Band 2 Blue	0.45 – 0.51	30
Band 3 Green	0.53 – 0.59	30
Band 4 Red	0.64 – 0.67	30
Band 5 NIR	0.85 – 0.88	30
Band 6 SWIR 1	1.57 – 1.65	30
Band 7 SWIR 2	2.11 – 2.29	30
Band 8 Pan	0.50 – 0.68	15
Band 9 Cirrus	1.36 – 1.38	30
Band 10 TIRS 1	10.6 – 11.19	100
Band 11 TIRS 2	11.5 – 12.51	100

Figura 1. Descrição dos canais da imagem do LANDSAT 8.

Dividimos o histograma em duas partes, considerando apenas valores abaixo de 6000 (Figura 5) e valores acima de 6000 (Figura 4) para termos uma melhor análise.

Podemos ver que não temos os valores comuns para imagens RGB onde em geral os pixels variam de 0 a 256. Vemos também comportamente separados para cada canal.

2.1. **PCA.** Primeiro, iremos normalizar os nossos dados para que tenham esperança 0 e variância 1. Desta forma, não ficaremos enviazados nos resultados pela diferença de escala. A maior variância será de fato dos dados e não pela escala.

Na Figura 6 podemos ver a explicação da variância de cada Componente Principal (PC). Vemos que com um PC apenas teremos 0.65 da variabilidade dos dados explicados. Juntos, os dois primeiros PCs tem 79% de explicação da variabilidade dos dados.

Podemos ver o comportamento dos primeiros dois PC no gráfico chamado *biplot*. Este gráfico se encontra na Figura 7.

Como um produto dessa redução de dimensionalidade, vê-se na Figura 8 a imagem em escala cinza da primeira PC. Na Figura 9 temos o seu histograma.

Para uma comparação, vejamos a imagem comprimida somente dos canais RGB na Figura 10 .

2.2. **Análise Fatorial.** Ao aplicarmos o processo da Análise Fatorial obtemos os seguintes resultados.

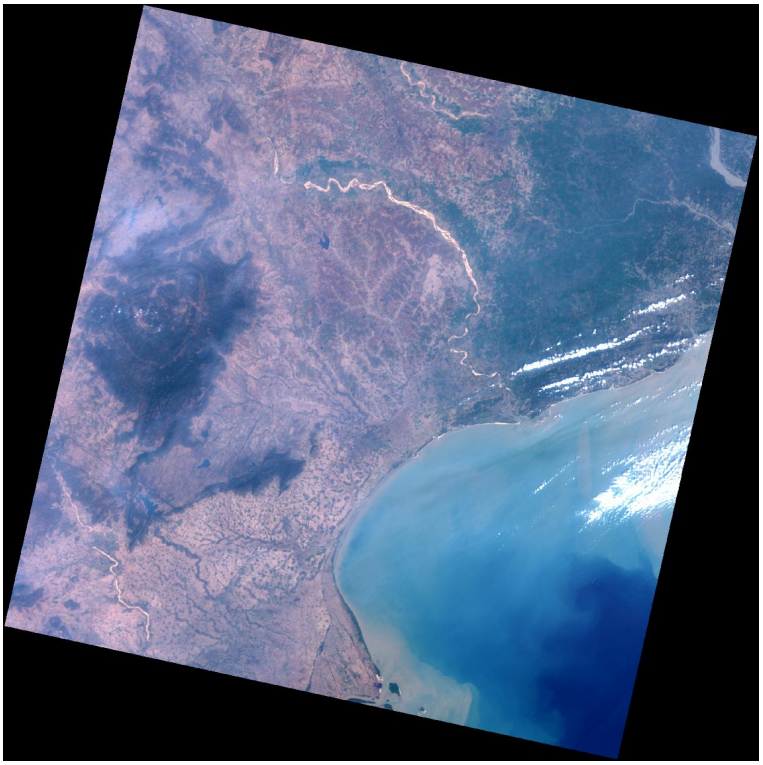

Figura 2. Imagem de Estudo.

Temos os *loadings* na Tabela 1. Para uma melhor interpretação coloquemos 0 nos loadings que forem menores que 0.5, vê se isso na Tabela 2. Existe interpretações associadas porque cada banda ou canal do satélite tem uma associação a um fenômeno físico.

F1	F2
0.946812	0.123880
0.193703	0.929151
0.195744	0.934924
0.963231	0.212297
0.924144	0.339988
0.818995	0.527240
0.415246	0.328909
0.543821	0.760991
0.609361	0.755275
0.024948	0.088758

Tabela 1. Loadings da FA

Podemos também ver a análise da variabilidade dos fatores na Tabela 3.

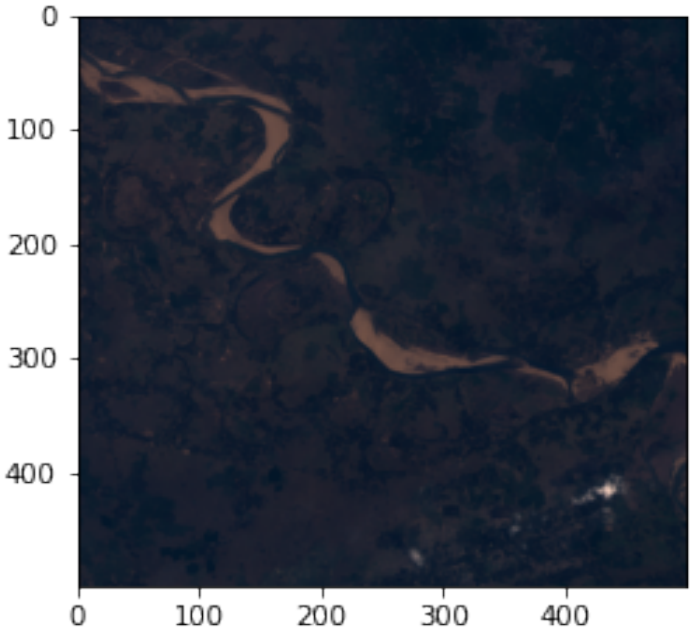

Figura 3. Imagem de Estudo cortada.

F1	F2
0.963231	0
0.946812	0
0.924144	0
0.818995	0.527240
0.609361	0.755275
0.543821	0.760991
0	0.934924
0	0.929151
0	0
0	0

Tabela 2. Loadings da FA ajustados

Finalmente, vejamos as figuras que os nossos fatores geraram para tentarmos inspecionar melhor o seus significados nas Figuras 11 e 13. Além disso, podemos ver os seus histogramas nas imagens 12 e 14.

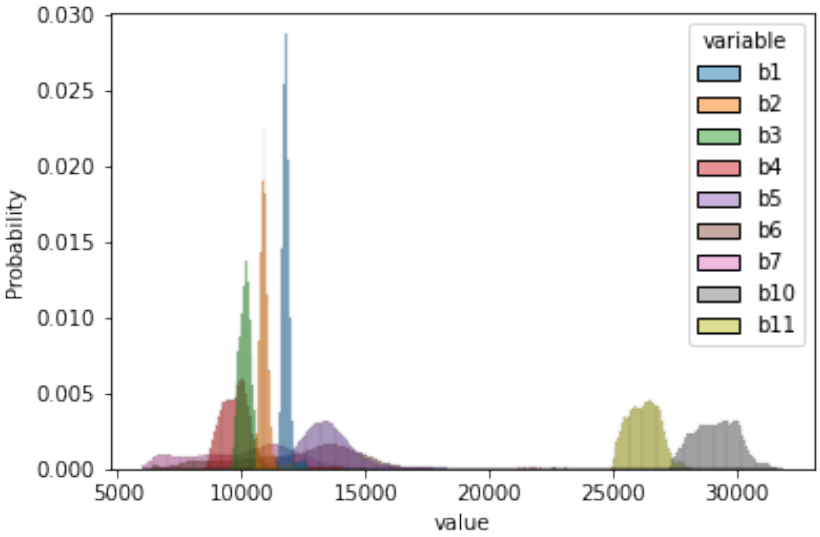

Figura 4. Histograma parte 1.

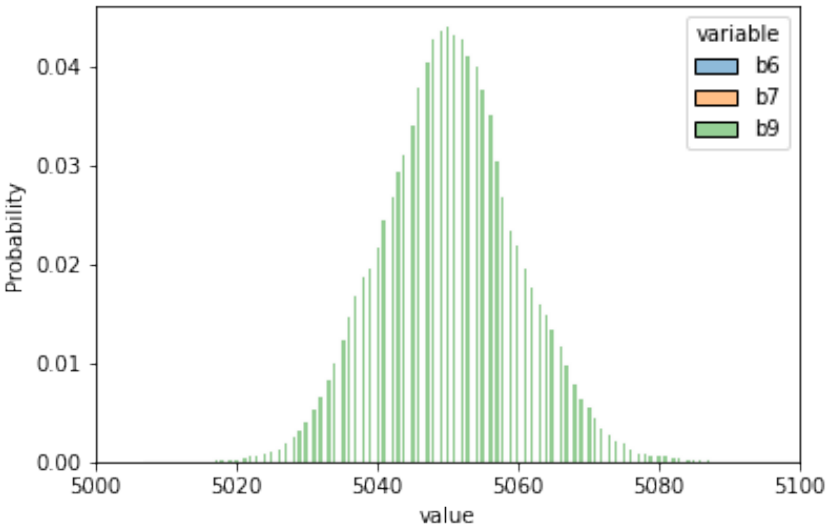

Figura 5. Histograma parte 2.

Podemos ver que um fator tentou comprimir a imagem como um todo, focando no gramado e estrada. Já o segundo fator, focou nos arbustos, nos detalhes da estrada e no enfoque de luz.

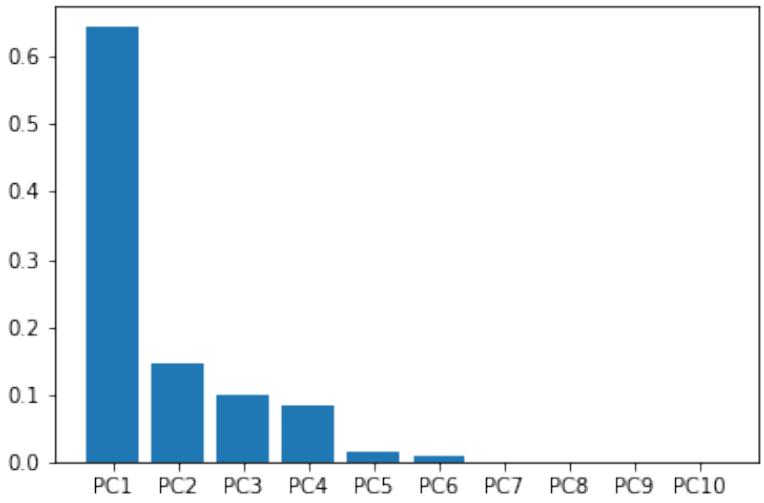

Figura 6. Explicação da Variância por cada PC.

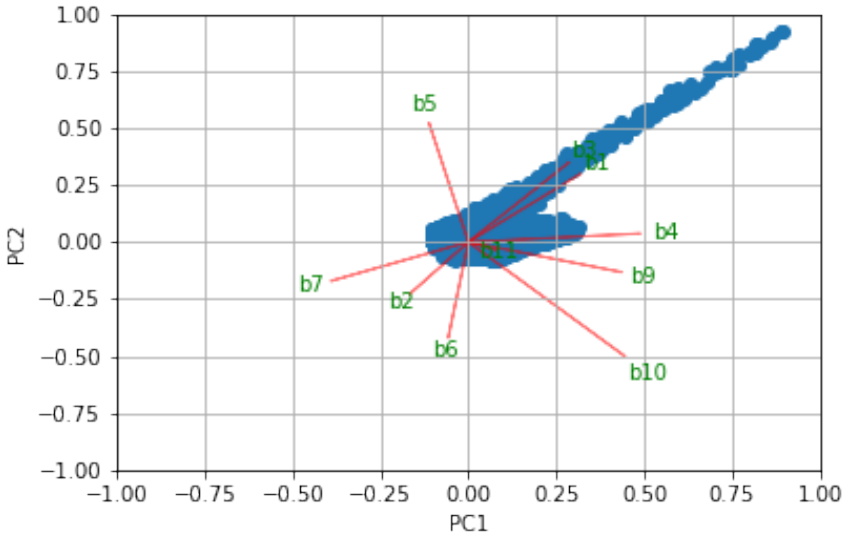

Figura 7. Biplot dos PC.

3. Agrupamento

3.1. **Dados.** Os dados que iremos usar para esta secção são dados macroeconômicos, retirados do site Kaggle, de quase todos os países do mundo. Os dados consistem de *child mort, exports, health, imports, income, inflation, life expec, total fer, gdpp.*

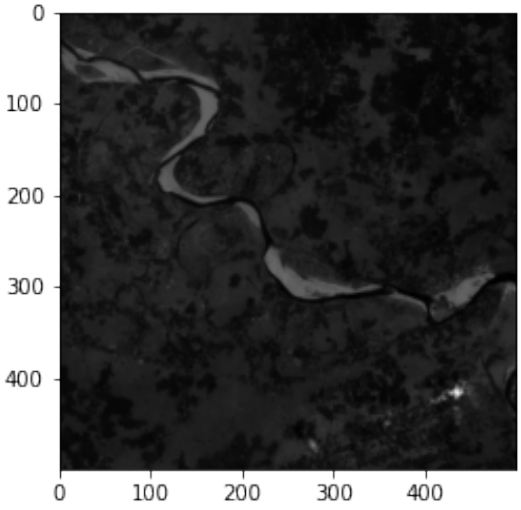

Figura 8. Imagem comprimida através de PCA.

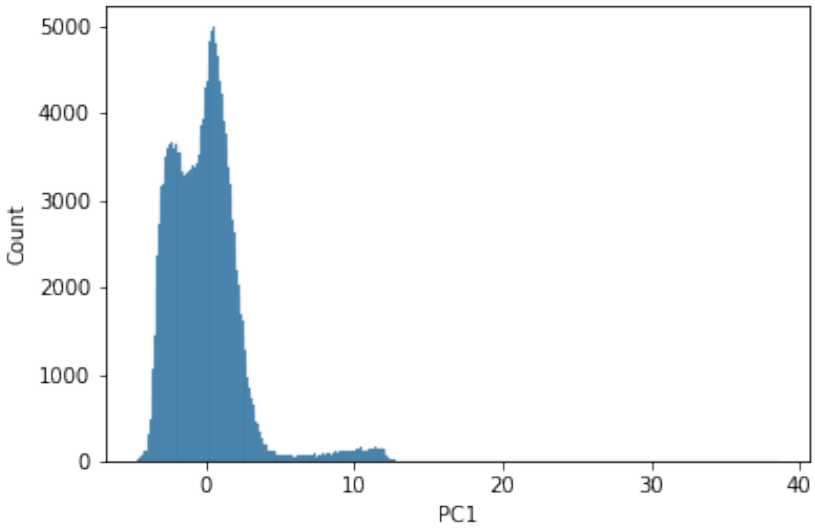

Figura 9. Histograma da imagem comprimida através de PCA.

3.2. **Encontrando o K ótimo.** Inicialmente, pelo próprio caráter não-supervisionado dos modelos, não sabemos nem quantos grupos queremos exatamente. Um método para amenizar essa incerteza é utilizar um critério de qualidade para as escolhas dos clusters. Em geral, usa-se o escore de silhueta ou "silhouette score"(SS). No qual, quanto mais próximo de 1 mais corretamente a observação foi colocada em relação a outros grupos, e -1 mais mal

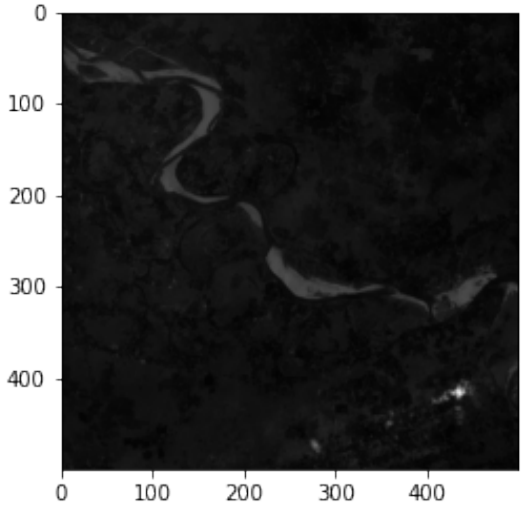

Figura 10. Imagem comprimida através de PCA somente dos canais RGB.

	F1	F2
SS Loadings	4.265013	3.457001
Variabilidade Prop	0.426501	0.345700
Variabilidade Cum	0.426501	0.772201

Tabela 3. Loadings da FA ajustados

colocada ela foi. Valores perto de 0 indicam que haveria outros grupos que a observação poderia fazer parte.

Sendo assim, iremos rodar cada modelo com K diferentes, tomar as medidas de qualidade dos clusters e comparar.

Vemos na Figura 15 o resultado desse processo. Vemos que todos os modelos acusaram que k=2 seria de melhor adequação. Isso já nos indica algo interessante. Os países são melhores divididos em 2 economicamente. Já começamos a supor que será entre países "bons"e países "ruins".

Vejamos os resultados de cada modelo na Figura 16.

Percebemos uma consistência muito grande de alguns países nos resultados de cada modelo. Países como EUA, Canáda, o continente europeu, AUS e o Japão, sempre andaram juntos. Isso nos indica que todos os modelos almejaram em dividir o mundo em "países de primeiro mundo e países de desenvolvimento".

Fatores socioeconômicos foram muito importantes como mortalidade infantil e taxa de fertilidade, etc. Provavelmente, que foi por esses motivos que mesmo sendo segunda maior potência econômica mundial, tenha sigo agrupado diferentemente.

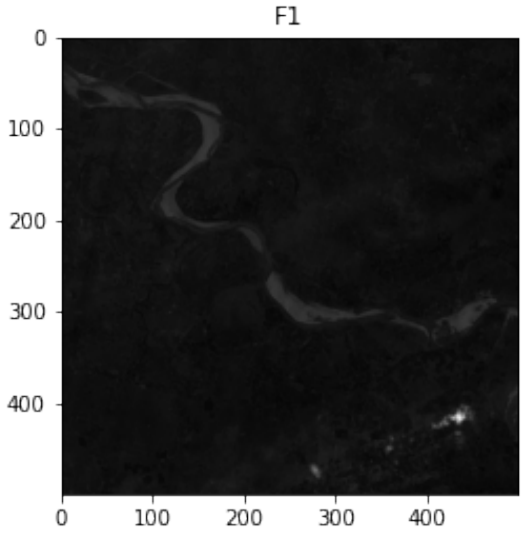

Figura 11. Imagem comprimida através FA: Fator 1.

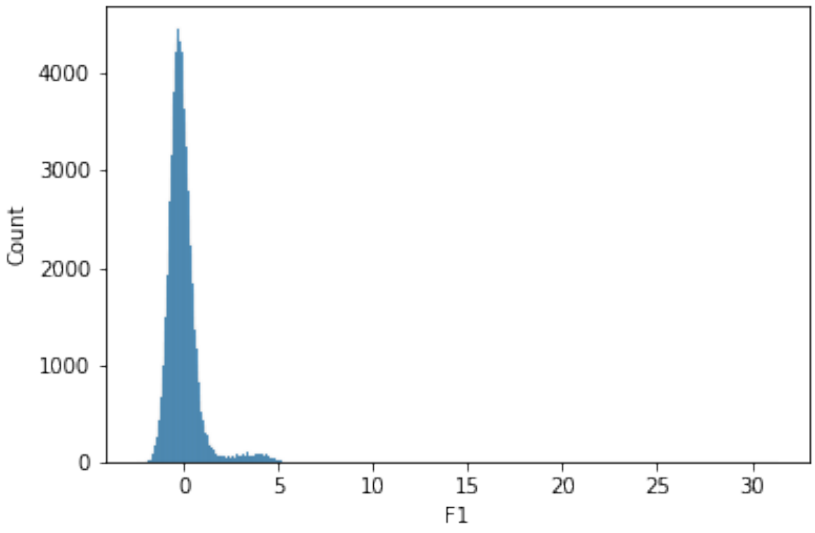

Figura 12. Histograma da imagem comprimida através da FA: Fator 1.

Da mesma forma, séries puramente econômicas como importação e exportação podem ter feito a Arábia Saúdita ter sido agrupada frequentmenete no grupo de países de primeiro mundo.

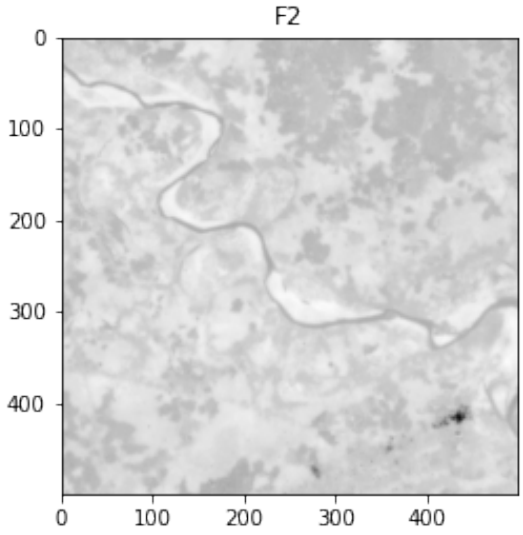

Figura 13. Imagem comprimida através de FA: Fator 2.

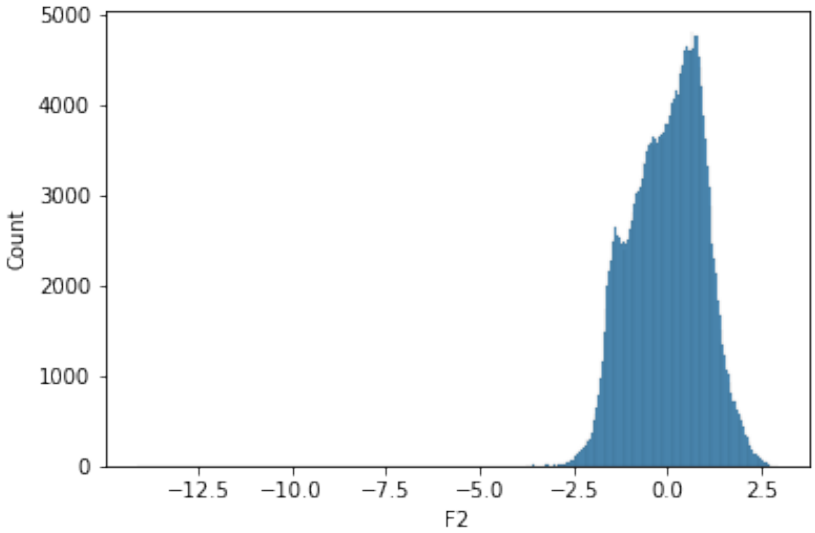

Figura 14. Histograma da imagem comprimida através de FA: Fator 2.

4. Análise Discriminante

4.1. **Dados.** Tomamos uma base de dados que consiste em dados médicos que descrevem medidas tomadas quando o paciente estava tendo um derrame ou AVC, e pacientes que aparentavam estarem tendo AVC mas não o estavam, de fato.

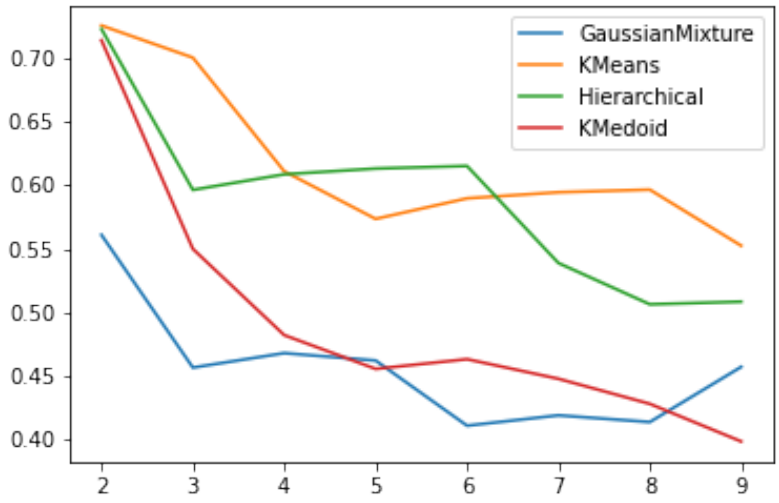

Figura 15. Número ótimo de Clusters por Modelo.

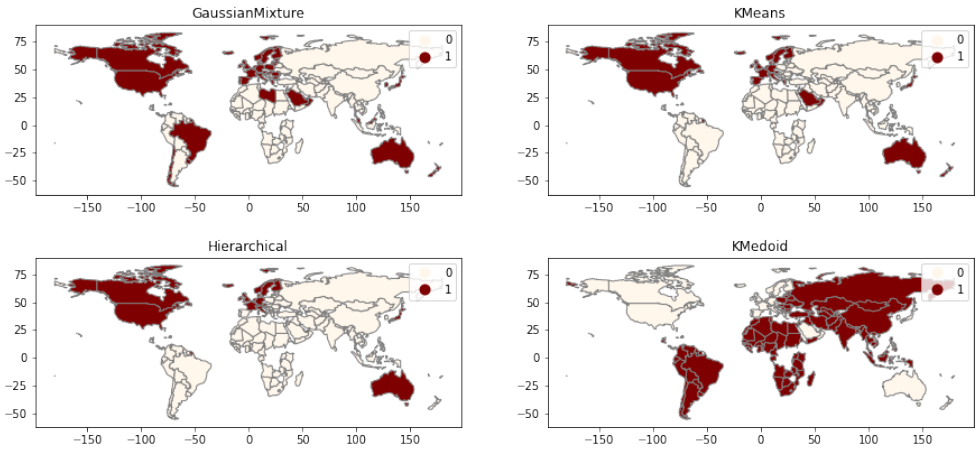

Figura 16. Resultado de cada modelo de clusterização.

As colunas são *'age', 'hypertension', 'heart-disease', 'avg-glucose-level', 'bmi', 'label', 'gender-Female', 'gender-Male', 'gender-Other', 'ever-married-No', 'ever-married-Yes', 'work-type-Govt-job', 'work-type-Never-worked', 'work-type-Private', 'work-type-Self-employed', 'work-type-children', 'Residence-type-Rural', 'Residence-type-Urban', 'smoking-status-Unknown', 'smoking-status-formerly smoked', 'smoking-status-never smoked', 'smoking-status-smokes'*.

Dividimos os dados entre teste e treino, 10% para o teste.

Encontramos a distribuição dos dados de AVC na Figura 17.

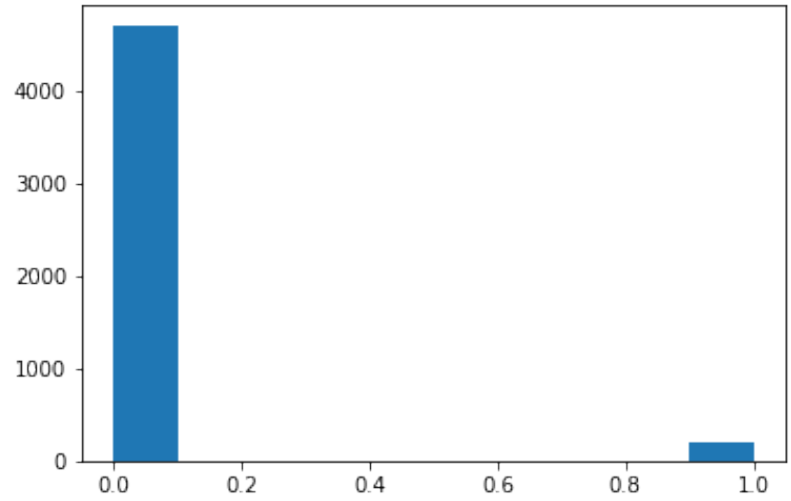

Figura 17. Distribuição dos dados de AVC.

4.2. **LDA.** Utilizando a técnica de LDA, conseguimos obter uma acurácia de 94% e a seguinte curva ROC na Figura 18. Mostrando bons resultados!

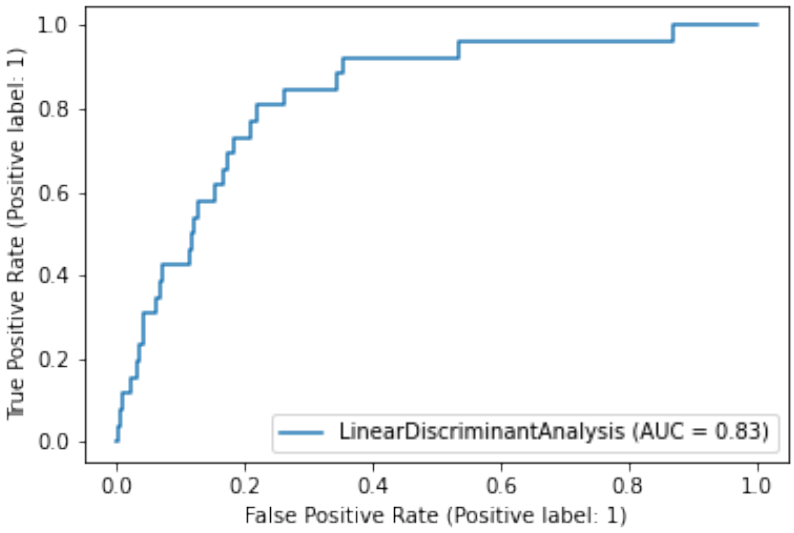

Figura 18. Curva ROC para os resultados do LDA.

4.3. **K-NN.** Inicialmente, o K-NN é muito sensível para a escolha do K. Sendo assim, iremos calculcar a acurácia média com a base de teste utilizando várias K diferentes. Vejamos o resultado na Figura 20.

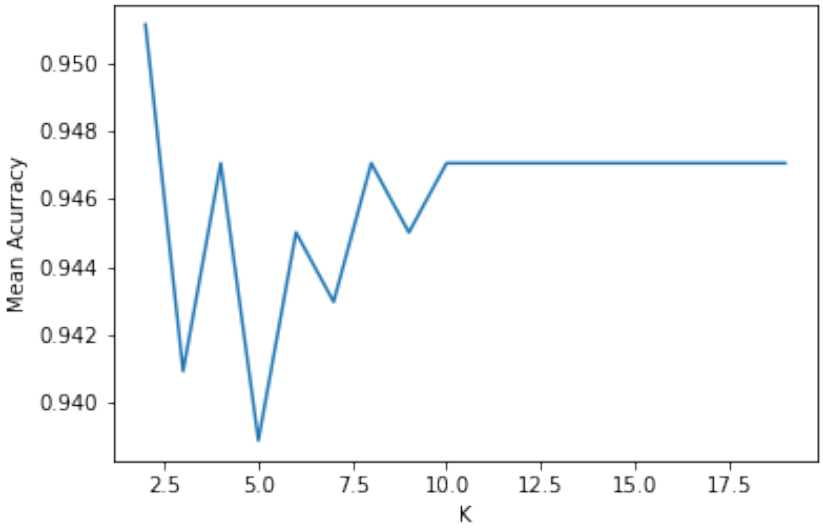

Figura 19. Escolha do K ótimo para K-NN.

Vejamos que K=2 obteve a melhor acurácia, porém não escolhemos K=2. Escolhemos K=50. Logo será explicado.

Obtivemos também uma acurácia de 94% porém, nossa curva ROC tem esse formato, o que tem um desempenho menor do que o LDA.

O motivo de escolhermos K=50 e não K=2 se dá pela curva ROC. Ao escolhermos K=2 aumentamos a acurácia, porém, aumento muito nosso erro tipo Falso Positivo. Nossa curva ROC ficou quase uma reta com resultado 0.5. Mostrando que estavámos muito tendenciosos, ao utilizarmos K=50 esse víeis desaparece ao custo muito baixo de acurácia.

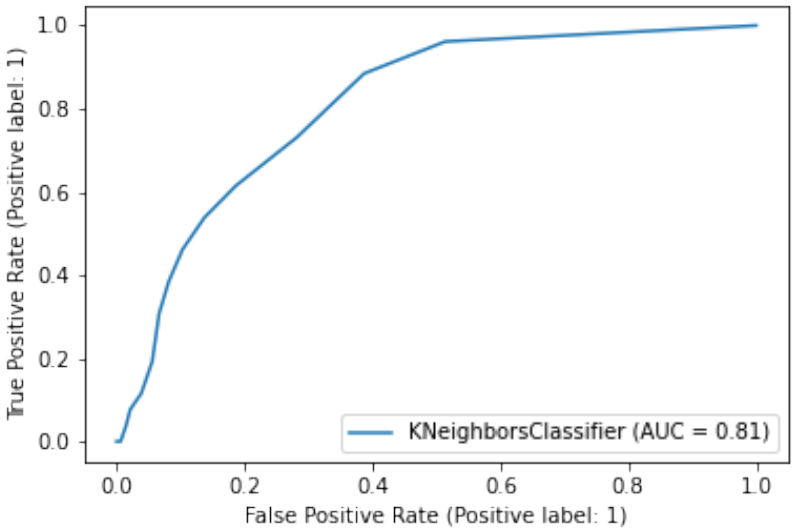

Figura 20. Curva ROC par ao modelo de K-NN.