5
Estudos de Caso

A seguir serão apresentados três estudos de caso. Os dois primeiros estudos de caso têm por objetivo demonstrar a facilidade de uso do aplicativo, e o último estudo de caso é focado em avaliar as vantagens do uso de métodos de natureza fuzzy em uma base de dados real.

5.1.
Base de Dados Exemplo do Matlab®

A Figura 28 mostra a representação gráfica dos dados presentes na base de dados fcmdata do Matlab® de dimensão 140x2 dividida em 2 grupos.

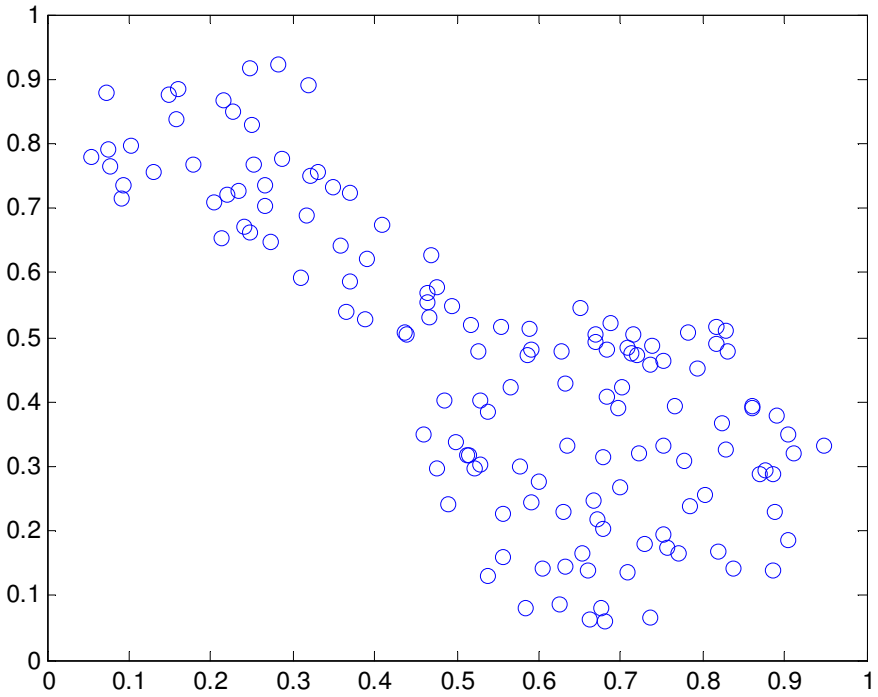

Figura 28: Base de Dados fcmdata do Matlab®.

Sobre essa base de dados foram aplicados os seguintes métodos: dois métodos hierárquicos aglomerativos, um método hierárquico divisivo, dois métodos particionais exclusivos e dois métodos particionais não-exclusivos.

A tabela de média se mostra mais útil quando se quer entender sobre os agrupamentos gerados. Para a base de dados fcmdata, os atributos não têm significado semântico algum, por isso a única informação útil que se pode ter dessa tabela é o total de dados em cada agrupamento. Pode-se sugerir que a média dos valores para as variáveis X e Y são os centros de cada agrupamento, porém isso não é verdade para muitos métodos.

No primeiro passo são selecionadas as duas variáveis presentes na base de dados, como mostra a Figura 29.

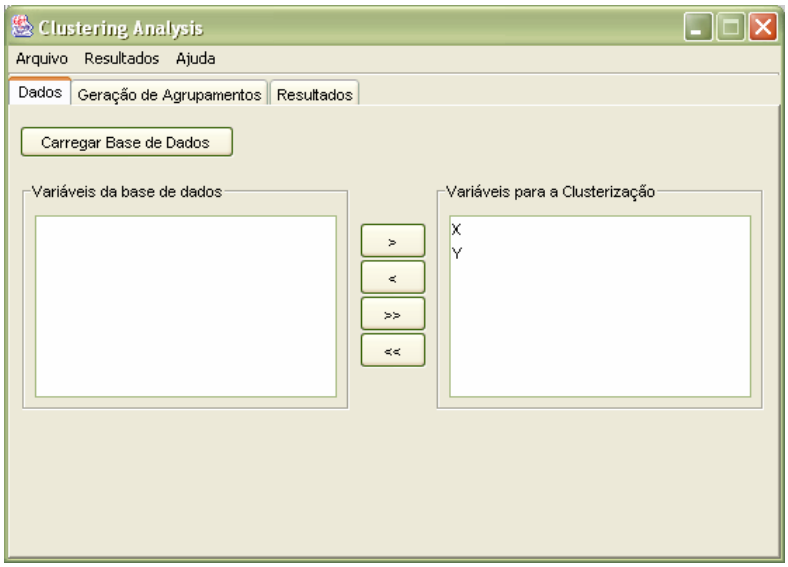

Figura 29: fcmdata – carregando base de dados

Em seguida são apresentados os resultados obtidos com a utilização de cada tipo de método de agrupamento, como se segue abaixo.

- *Método Hierárquico Aglomerativo – AGNES*
  - Método de Distância entre Grupos: Ligação Simples

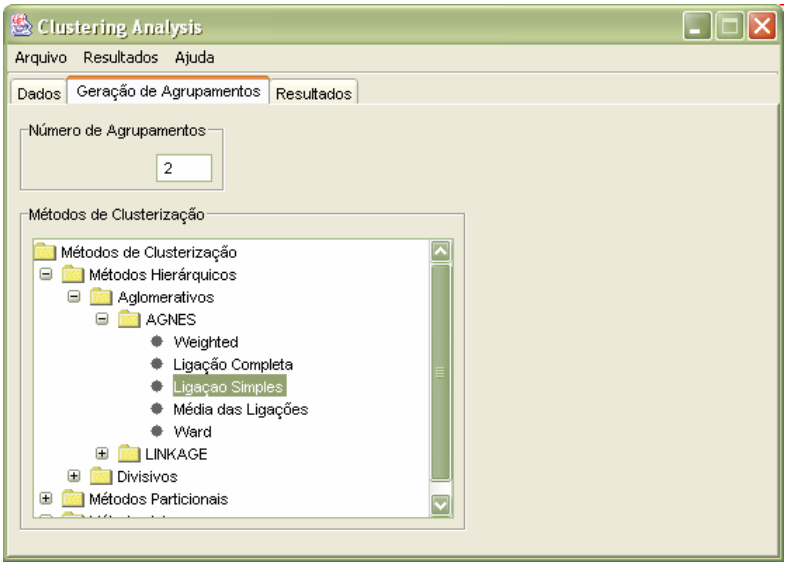

Figura 30: fcmdata – Geração de Agrupamentos (AGNES – Ligação Simples).

Resultados gerados:

Tabela 6: fcmdata – Tabela de Médias (AGNES – Ligação Simples).

	X	Y	Total
Grupo 1	0,5507552733812949	0,45679863309352525	139
Grupo 2	0,072686	0,87749	1

Pelos dados da tabela 6 já é possível verificar que os resultados obtidos com utilização desse método são muito ruins, onde um dos agrupamentos (grupo 2) tem apenas um dado. Isso pode ser facilmente observado no dendograma da Figura 31, onde não há uma divisão clara entre os agrupamentos. Com esses dados não é necessário verificar o gráfico da silhueta.

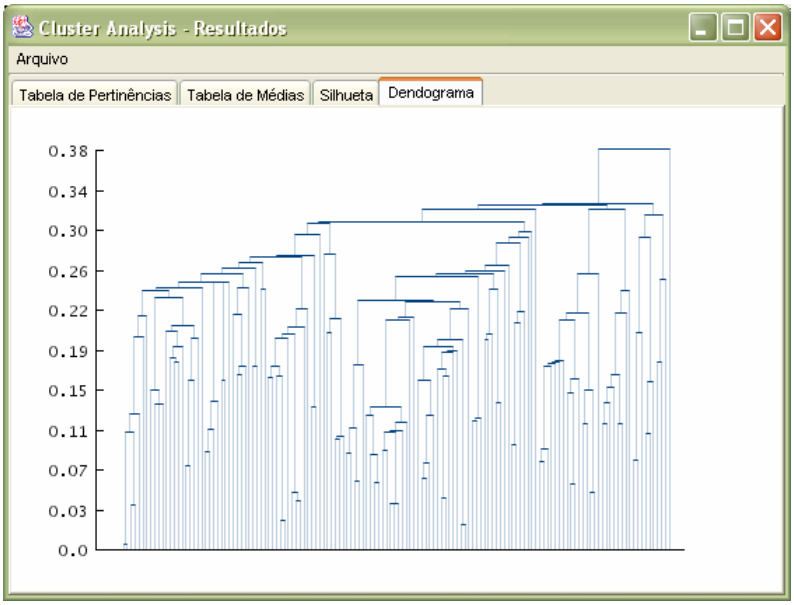

Figura 31: fcmdata – Dendograma (AGNES – Ligação Simples).

- *Método Hierárquico Aglomerativo – LINKAGE*
  - Método de Distância entre Classes: Média das Ligações

Resultados gerados:

Tabela 7: fcmdata – Tabela de Médias (LINKAGE – Média das Ligações).

	X	Y	Total
Grupo 1	0,6928909890109891	0,32406054945054946	91
Grupo 2	0,2770324285714285	0,7118977551020407	49

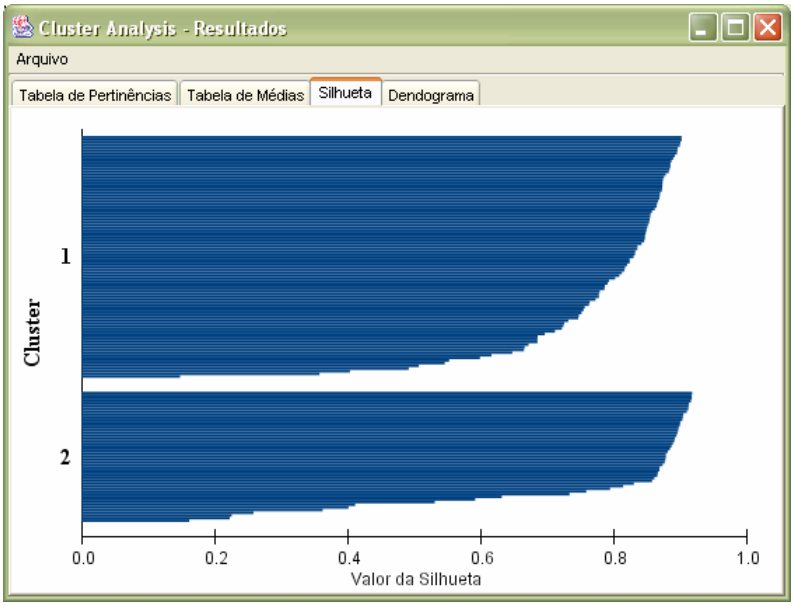

Figura 32: fcmdata – Gráfico da Silhueta (LINKAGE – Média das Ligações).

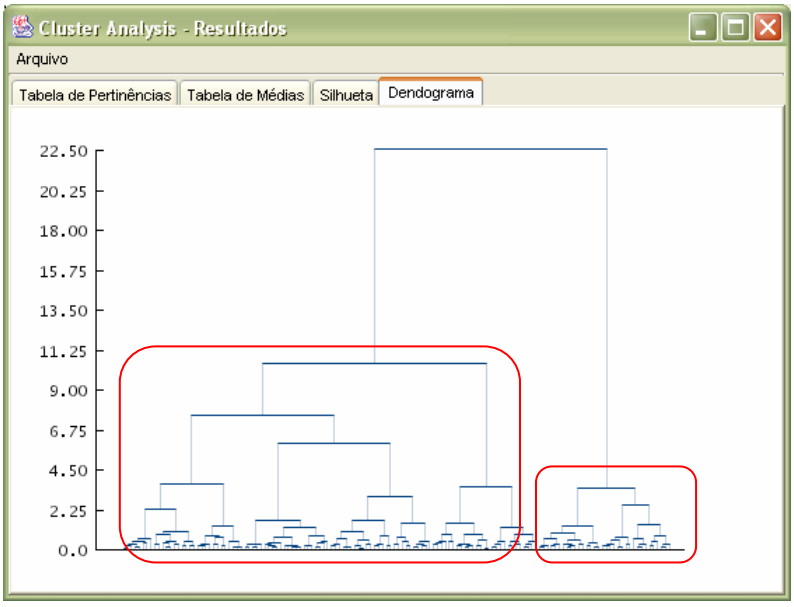

Figura 33: fcmdata – Dendograma (LINKAGE – Média das Ligações).

Ao se analisar o dendograma (Figura 33) pode-se observar que há uma divisão clara entre os agrupamentos. Pelo gráfico da silhueta (Figura 32) pode-se observar uma estrutura forte para os dois agrupamentos encontrados. Por essas razões conclui-se que esse método foi bastante eficiente para essa base de dados, definindo uma estrutura concisa sobre a mesma.

Pode-se ainda concluir que a eficiência dos métodos hierárquicos aglomerativos está muito relacionada com o método escolhido de distância entre classes. Isso pode ser observado quando foi utilizado sobre a mesma base de dados o mesmo algoritmo de agrupamento de dados, porém com métodos de distância entre classes diferentes (*Ligação Simples* e *Média das Ligações*), gerando resultados bem diferentes sobre a estrutura dos dados.

- *Método Hierárquico Divisivo – DIANA*

  Resultados gerados:

  Tabela 8: fcmdata – Tabela de Médias (DIANA).

	X	**Y**	**Total**
Grupo 1	0,2770324285714285	0,7118977551020407	49
Grupo 2	0,6928909890109891	0,32406054945054946	91

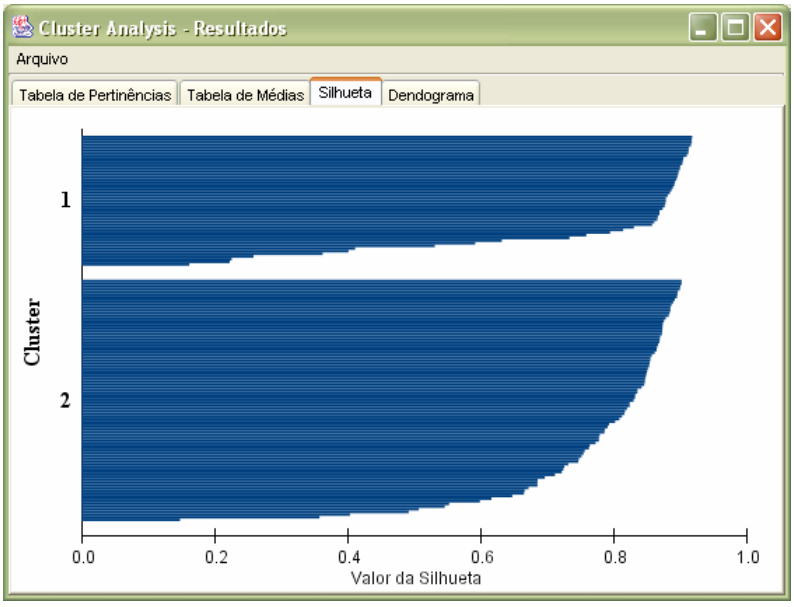

Figura 34: fcmdata – Gráfico da Silhueta (DIANA).

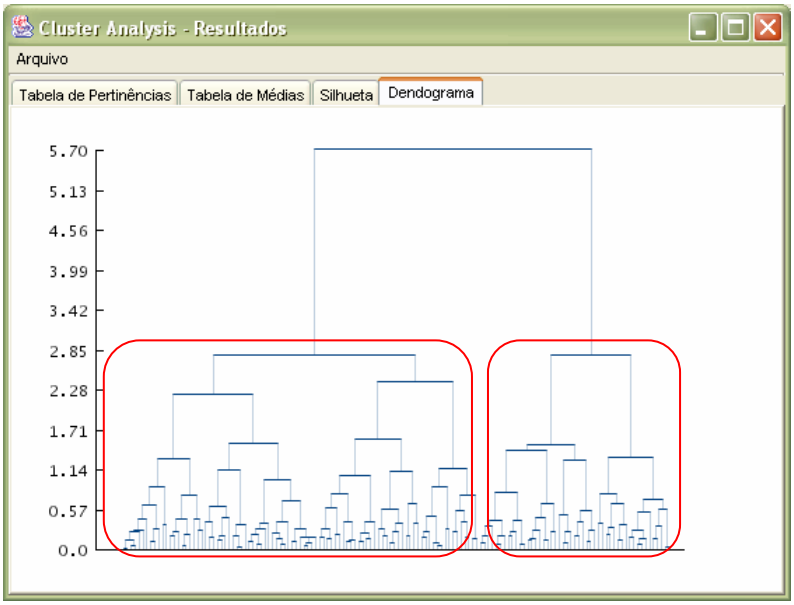

Figura 35: fcmdata – Dendograma (DIANA).

Pode-se observar que foi obtido o mesmo resultado do método anterior, com uma divisão clara entre os agrupamentos, bem como uma boa estrutura entre eles. Por essas razões, conclui-se que esse método foi bastante eficiente.

- *Método Particional Exclusivo – PAM*

Resultados gerados:

Tabela 9: fcmdata – Tabela de Médias (PAM).

	X	Y	Total
Grupo 1	0,2770324285714285	0,7118977551020407	49
Grupo 2	0,6928909890109891	0,32406054945054946	91

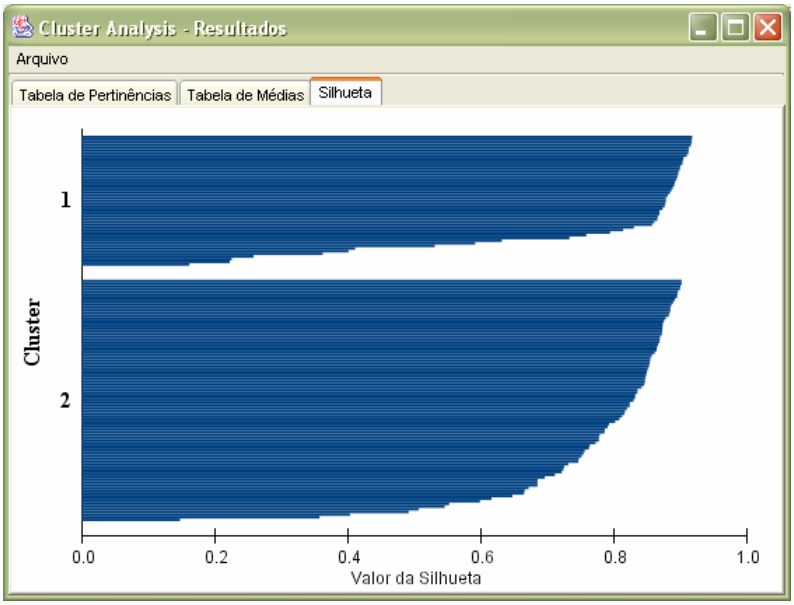

Figura 36: fcmdata – Gráfico da Silhueta (PAM).

Da mesma forma que o experimento anterior, um resultado similar foi obtido, com uma boa estrutura entre os agrupamentos, sendo, portanto, um método bastante eficiente.

- *Método Particional Exclusivo – K-Means*

Resultados gerados:

Tabela 10: fcmdata – Tabela de Médias (K-Means).

	X	Y	Total
Grupo 1	0,6928909890109891	0,32406054945054946	91
Grupo 2	0,2770324285714285	0,7118977551020407	49

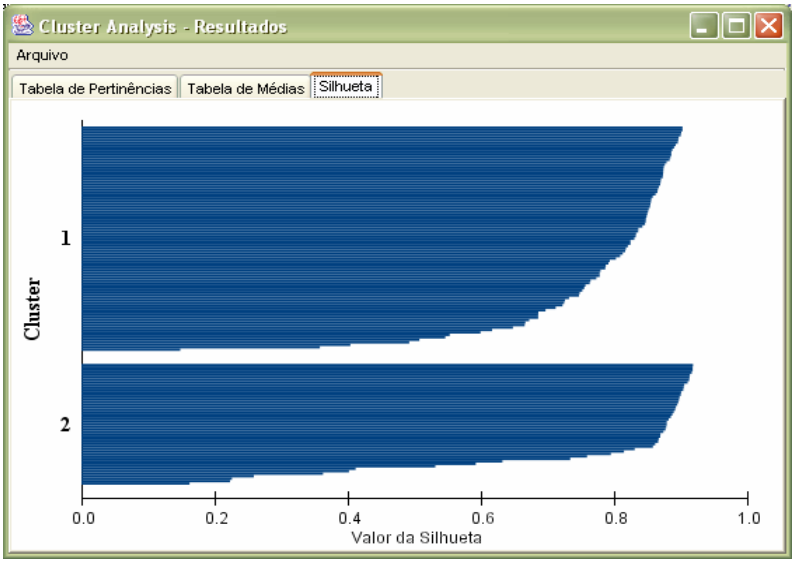

Figura 37: fcmdata – Gráfico da Silhueta (K-Means).

Também se percebe um resultado similar ao do método anterior. Através de comparações sobre as tabelas de pertinência do resultado desse método em relação ao resultado do método anterior, verificou-se que foram gerados os mesmos agrupamentos.

- *Método Particional Não-Exclusivo – FANNY*

  Resultados gerados:

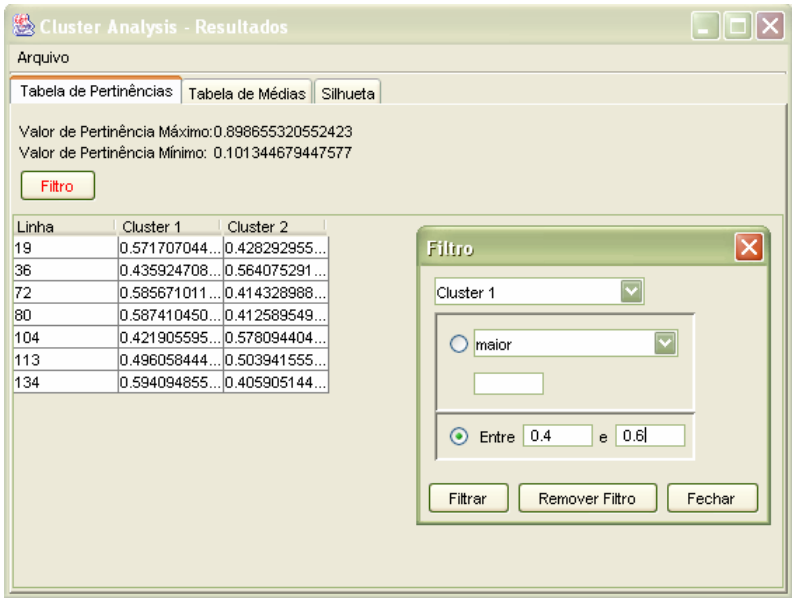

Figura 38: fcmdata – Tabela de Pertinências / Filtro (FANNY).

Tabela 11: fcmdata – Tabela de Médias (FANNY).

	X	Y	Total
Grupo 1	0,2770324285714285	0,7118977551020407	49
Grupo 2	0,6928909890109891	0,32406054945054946	91

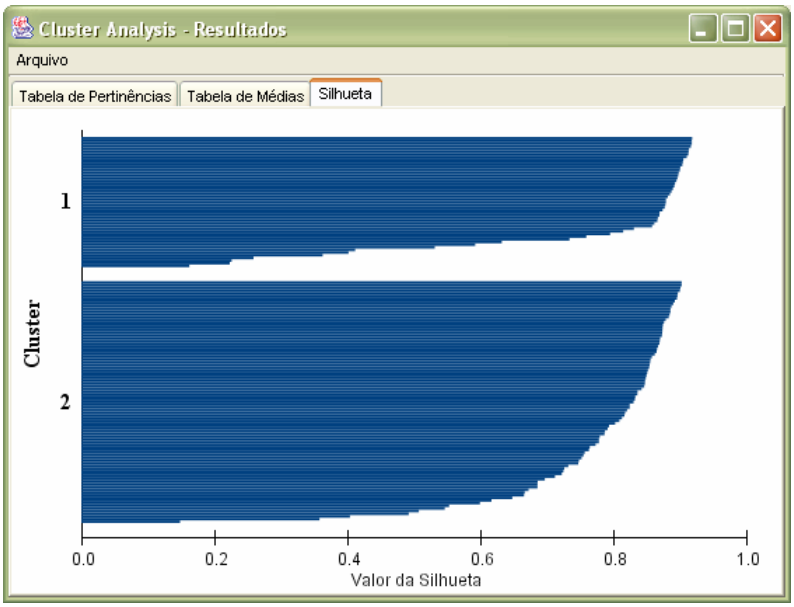

Figura 39: fcmdata – Gráfico da Silhueta (FANNY).

Pode-se observar que foi obtido um resultado bastante similar ao do método anterior. Comparações sobre as tabelas de pertinência geradas pelos dois métodos mostraram que foram gerados os mesmos agrupamentos.

Em uma análise fuzzy, pode-se observar a existência de pelo menos 7 dados na fronteira dos agrupamentos.

Pela simplicidade da base de dados em questão e por não se tratar de uma base de dados real, fica difícil analisar a estrutura fuzzy formada semanticamente, reduzindo a análise para agrupamentos de dados não-fuzzy.

- *Método Particional Não-Exclusivo – Fuzzy C-Means*

  Resultados gerados:

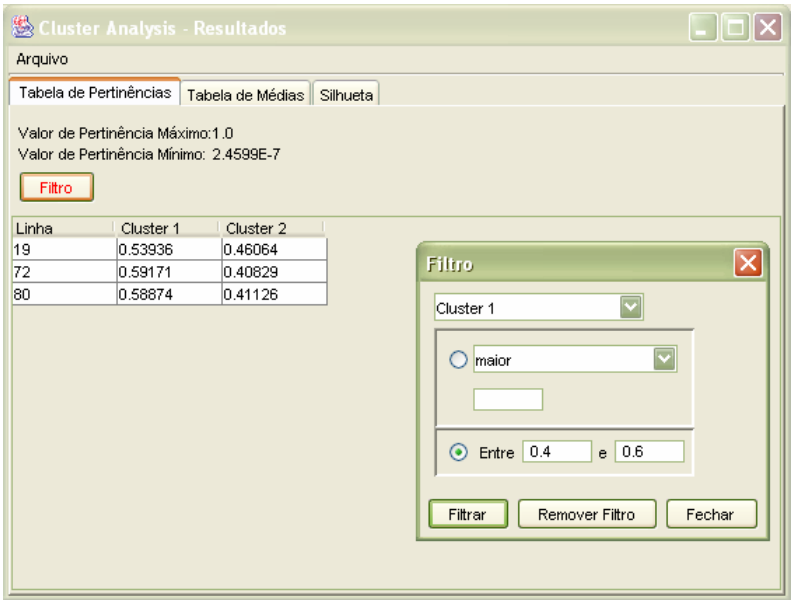

Figura 40: fcmdata – Tabela de Pertinências / Filtro (Fuzzy C-Means).

Tabela 12: fcmdata – Tabela de Médias (FANNY).

	X	Y	Total
Grupo 1	0,2770324285714285	0,7118977551020407	49
Grupo 2	0,6928909890109891	0,32406054945054946	91

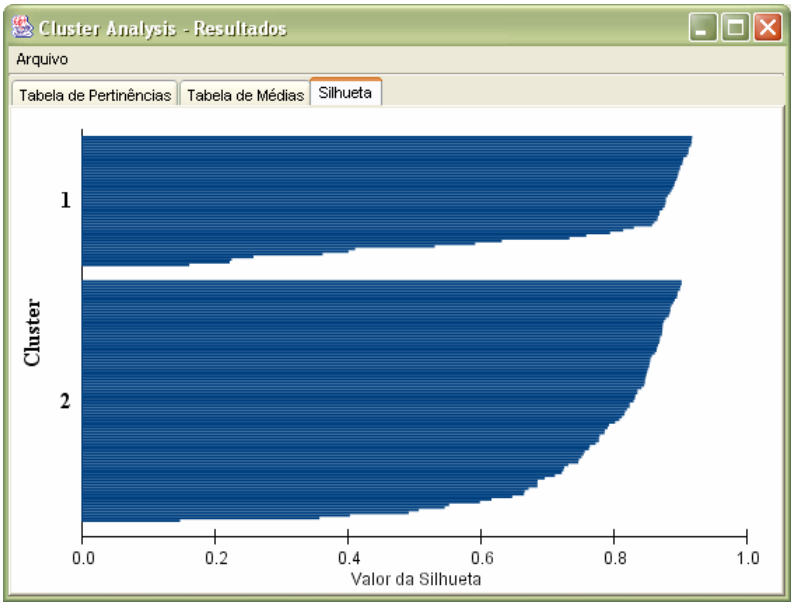

Figura 41: fcmdata – Gráfico da Silhueta (Fuzzy C-Means).

A execução desse método gerou os mesmo resultados do método anterior, o que pode ser comprovado através da análise sobre suas tabelas de pertinência.

Em uma análise fuzzy, pode-se observar a existência de pelo menos 3 dados na fronteira dos agrupamentos.

Assim como no método anterior, fica difícil fazer uma análise fuzzy mais conclusiva.

Os métodos listados foram executados utilizando como medida de proximidade a distância euclidiana. Alguns desses métodos foram executados utilizando também a distância de Manhattan, porém não foram notadas nenhumas mudanças significativas nos resultados.

Após a execução de diversos métodos de diferentes tipos sobre essa base de dados, pode-se observar que a maioria dos métodos gerou os mesmos resultados e que, pela simplicidade da base utilizada e por não se tratar de uma base de dados real, o uso de métodos de natureza fuzzy, apesar de mostrar uma estrutura mais rica em detalhes, não forneceu todo o benéfico que uma análise fuzzy permite.

5.2. Base de Dados Íris

Esta é uma base de dados bem conhecida que apresenta as medidas em centímetro da largura e comprimento das pétalas e sépalas de três espécies de flor íris (Setosa, Virginica e Versicolor) (Fisher, 1936). Esta base contém 150 amostras e 5 variáveis:

- Sépala.Largura
- Sépala.Comprimento
- Pétala.Largura
- Pétala.Comprimento
- Espécie

O objetivo desse estudo de caso é demonstrar a facilidade de uso do aplicativo para os seguintes objetivos:

- Determinar o número de agrupamentos.
- Identificar os agrupamentos.
- Avaliar o desempenho de cada método utilizado para realizar o agrupamento.

São conhecidos previamente o número de agrupamentos e a classificação de cada amostra, portanto essas informações serão usadas na avaliação dos objetivos traçados.

Para a determinação do número de agrupamentos serão usados um método hierárquico aglomerativo e um método hierárquico divisivo.

- *Método Hierárquico Aglomerativo – AGNES*
  - Método de Distância entre Classes: WARD

Resultados gerados:

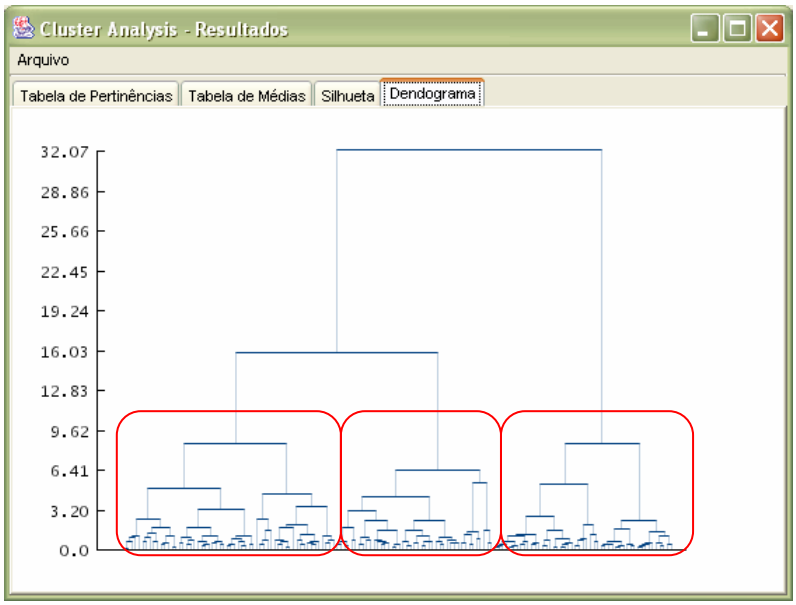

Figura 42: íris – Dendograma (AGNES – Ward).

Pelo dendograma da Figura 42 pode-se observar que há uma divisão clara entre pelo menos três agrupamentos.

- *Método Hierárquico Divisivo – DIANA*

  Resultados gerados:

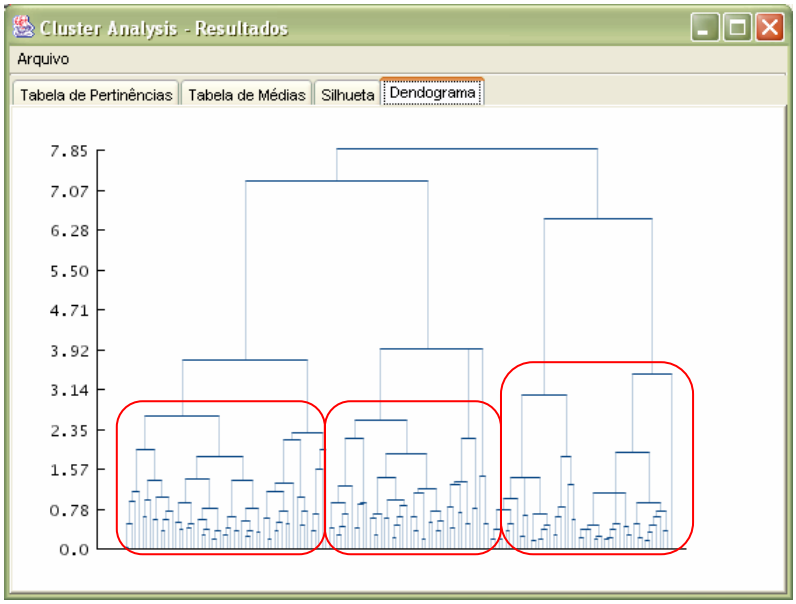

Figura 43: íris – Dendograma (DIANA).

Da mesma forma que o método anterior, pelo dendograma da Figura 43 pode-se observar que há uma divisão clara entre pelo menos três agrupamentos. Pode-se sugerir a presença de mais um agrupamento, porém o resultado obtido com o método anterior dá uma certeza maior sobre a presença de três agrupamentos.

Para a identificação dos agrupamentos serão usados um método hierárquico aglomerativo, um método hierárquico divisivo, dois métodos particionais exclusivos e dois métodos particionais não-exclusivos.

Nessa primeira etapa, para a avaliação dos agrupamentos gerados será utilizado o gráfico da silhueta.

- *Método Hierárquico Aglomerativo – AGNES*
  - Método de Distância entre Classes: WARD

Resultados gerados:

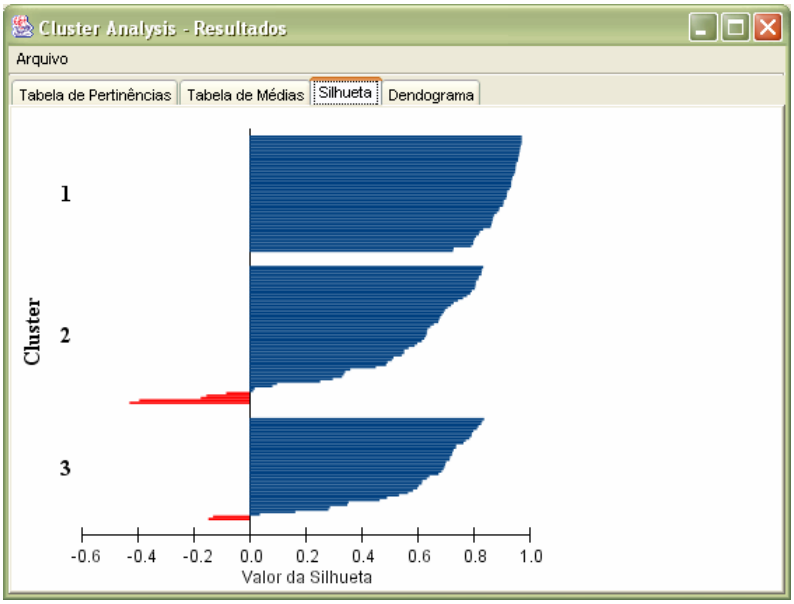

Figura 44: íris – Gráfico da Silhueta (AGNES - Ward).

Pode-se observar através do gráfico da silhueta da Figura 44 a presença de três estruturas bem definidas, apesar dos agrupamentos 2 e 3 conterem algumas amostras mal agrupadas.

- *Método Hierárquico Divisivo – DIANA*

Resultados gerados:

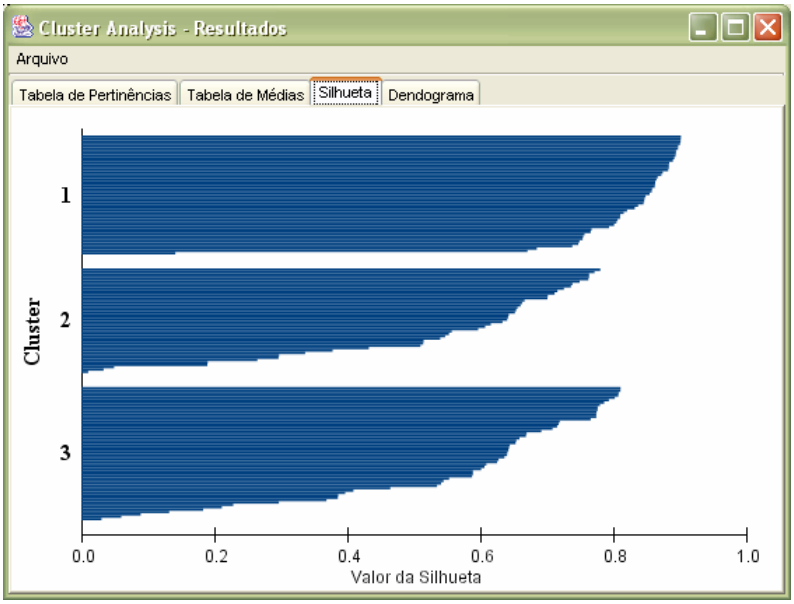

Figura 45: íris – Gráfico da Silhueta (DIANA).

Novamente observa-se a presença de três estruturas bem definidas (Figura 45). A utilização desse método apresentou um resultado melhor do que o apresentado pelo método anterior.

- *Método Particional Exclusivo – K-Means*

Resultados gerados:

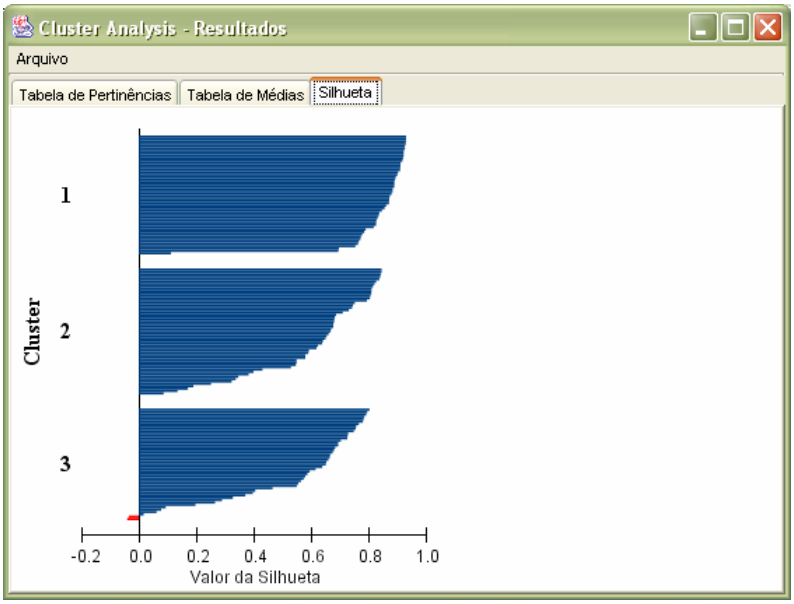

Figura 46: íris – Gráfico da Silhueta (K-Means).

Analisando o gráfico da silhueta da Figura 46 pode-se observar que, assim como os métodos hierárquicos, o *k*-means apresentou três estruturas bem definidas, com algumas amostras mal agrupadas no terceiro agrupamento.

- *Método Particional Exclusivo – PAM*

Resultados gerados:

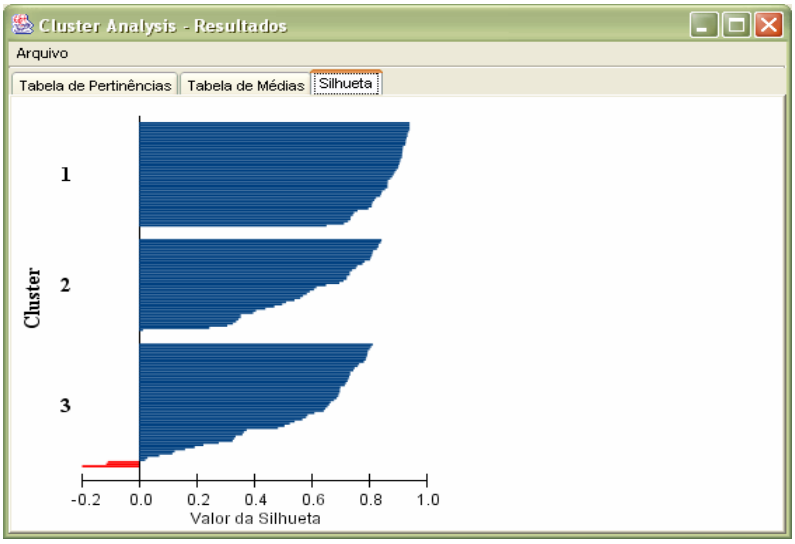

Figura 47: íris – Gráfico da Silhueta (PAM).

Analisando o gráfico da silhueta da Figura 47, o método *PAM* apresentou um resultado bastante semelhante ao *k-means*, com pouca diferença para o terceiro agrupamento, onde o método *PAM* se mostrou um pouco pior.

- *Método Particional Não-Exclusivo – Fuzzy C-Means*

Resultados gerados:

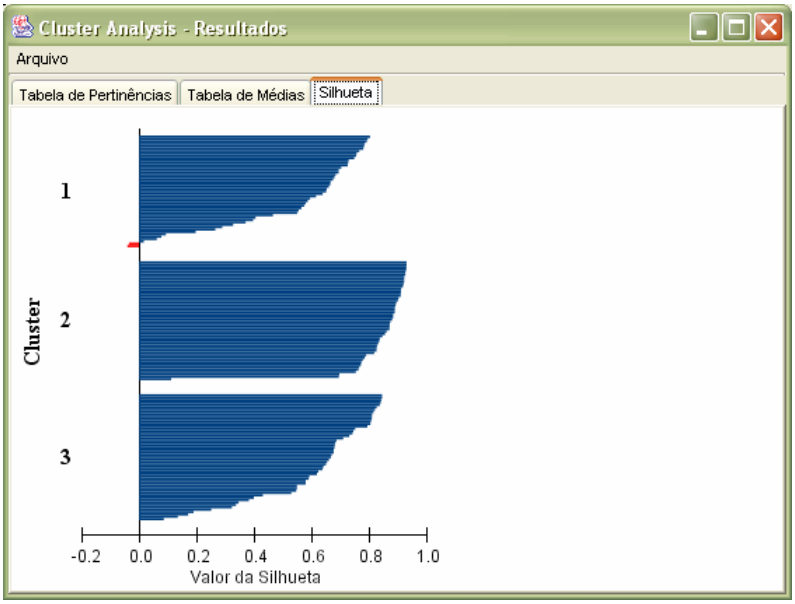

Figura 48: íris – Gráfico da Silhueta (Fuzzy C-Means).

O método *FCM* apresentou exatamente o mesmo resultado gerado pelo método *k-means*, o que pode ser observado através da Figura 48.

- *Método Particional Não-Exclusivo – FANNY*

Resultados gerados:

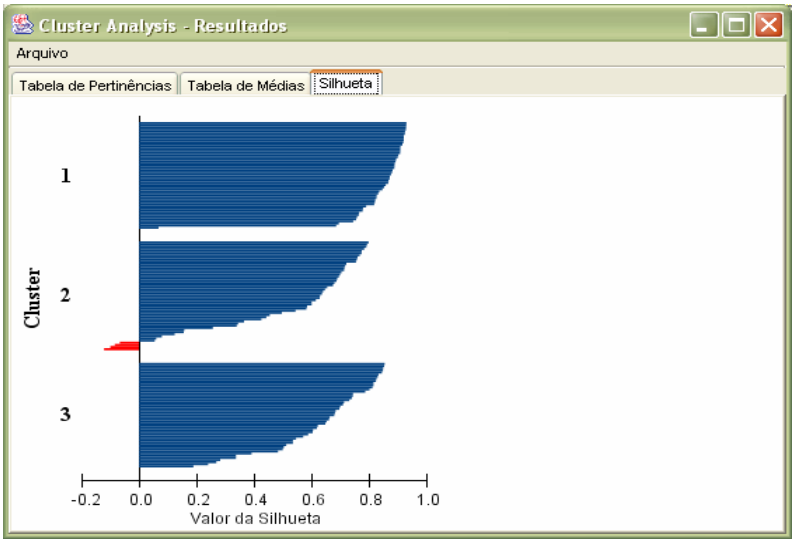

Figura 49: íris – Gráfico da Silhueta (FANNY).

O método *FANNY* apresentou um resultado semelhante aos demais métodos particionais, contendo poucas amostras mal agrupadas no segundo agrupamento, o que pode ser observado pelo gráfico da silhueta da Figura 49.

A seguir é apresentada uma tabela com o grau de acerto de cada método aplicado.

Tabela 13: íris - Avaliação dos métodos de agrupamento de dados.

	AGNES	DIANA	K-Means	PAM	FCM	FANNY
Acerto (%)	82,67	81,33	83,33	84,67	83,33	84,67

Pela tabela acima podemos verificar que os métodos utilizados apresentaram um desempenho muito semelhante, embora os métodos particionais tenham tido um desempenho um pouco melhor.

Pelos testes realizados pode-se verificar a facilidade de uso do aplicativo na determinação do número de agrupamentos – através da análise dos dendogramas na utilização dos métodos hierárquicos – e na comparação de desempenho dos métodos disponíveis.

5.3. Base de Dados de Hipertensão na Ilha do Governador

A seguir é apresentado um estudo de caso com dados reais colhidos através de um inquérito domiciliar da população da Ilha do Governador (Klein, 1995), com o objetivo de estimar a prevalência de hipertensão arterial (HA) na população adulta (acima de 20 anos de idade) e de analisar as possíveis associações da pressão arterial com algumas variáveis pré-definidas. Foram selecionados e entrevistados 1270 indivíduos, moradores de 750 domicílios da I.G., nos quais aplicou-se um questionário padronizado e realizaram-se medidas objetivas.

Utilizou-se neste estudo de caso um método hierárquico aglomerativo, um método hierárquico divisivo, um método particional exclusivo e um método particional não exclusivo. Esse estudo de caso é focado na utilização de métodos de natureza fuzzy, com o objetivo de apresentar sua potencialidade de identificação de alguns indivíduos cuja pertinência a um ou outro agrupamento pode não estar bem definida. O número de grupos foi estabelecido como sendo 3 e utilizaram-se as seguintes variáveis no processo de agrupamento de dados:

- Idade;
- Consumo de cigarros industriais;
- Peso;
- Altura;
- Pressão arterial sistólica;
- Pressão arterial diastólica.

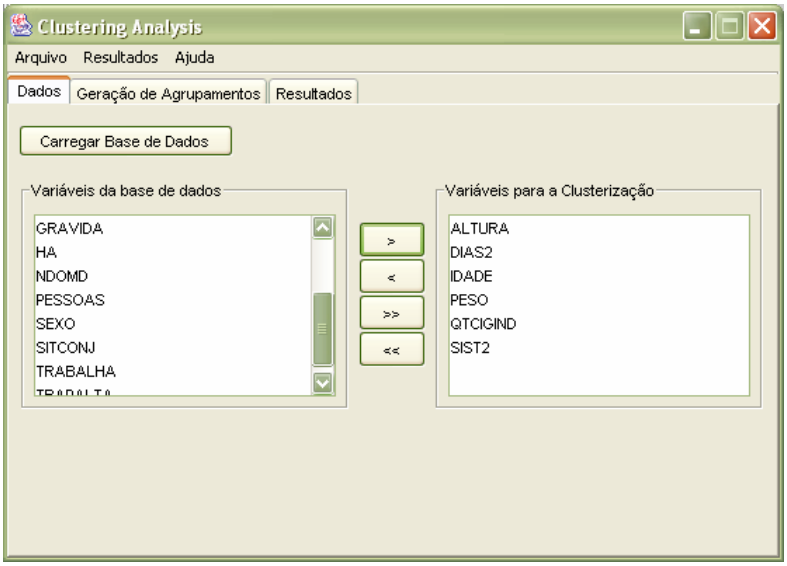

Figura 50: Base de Dados da Ilha do Governador – Seleção de Variáveis.

A seguir são apresentados alguns dos resultados obtidos com diferentes métodos.

- *Método Hierárquico Aglomerativo – LINKAGE*
  - Método de Distância entre Classes: Média das Ligações

Resultados gerados:

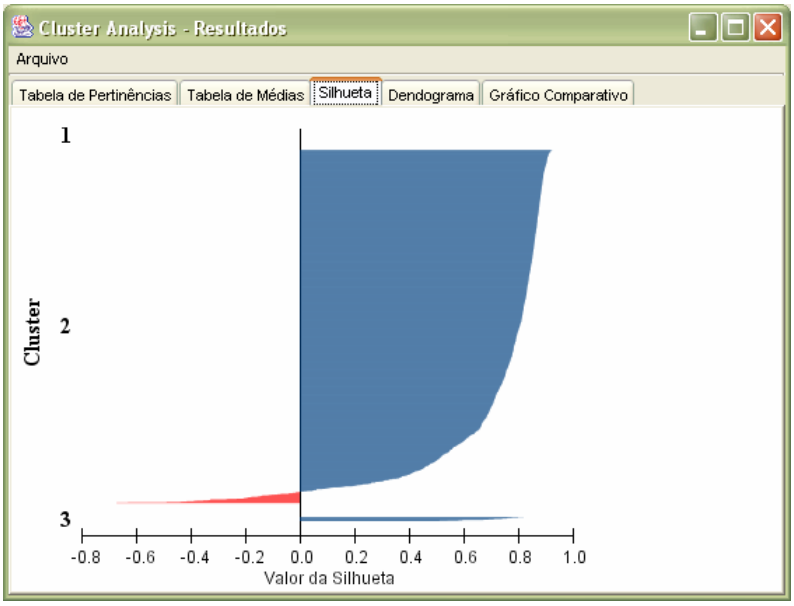

Figura 51: Base de Dados da Ilha do Governador – Gráfico da Silhueta (LINKAGE – Média das Ligações).

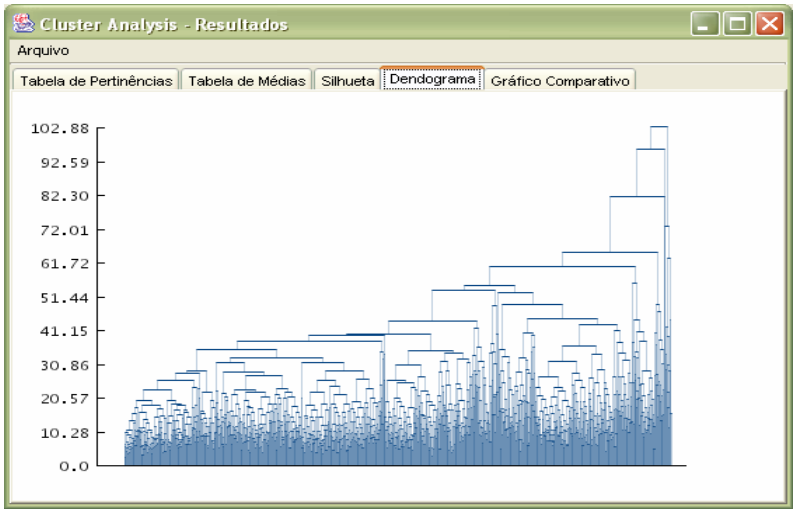

Figura 52: Base de Dados da Ilha do Governador – Dendograma (LINKAGE – Média das Ligações).

Ao se analisar o dendograma da Figura 52, pode-se observar que não há uma divisão clara entre os agrupamentos. Pelo gráfico da silhueta da Figura 51 pode-se observar uma estrutura muito ruim entre os agrupamentos gerados. Embora esse método tenha funcionado muito bem no estudo de caso anterior, para essa base de dados ela se mostrou bastante ineficiente.

- *Método Hierárquico Aglomerativo – LINKAGE*
  - Método de Distância entre Classes: Ward

Resultados gerados:

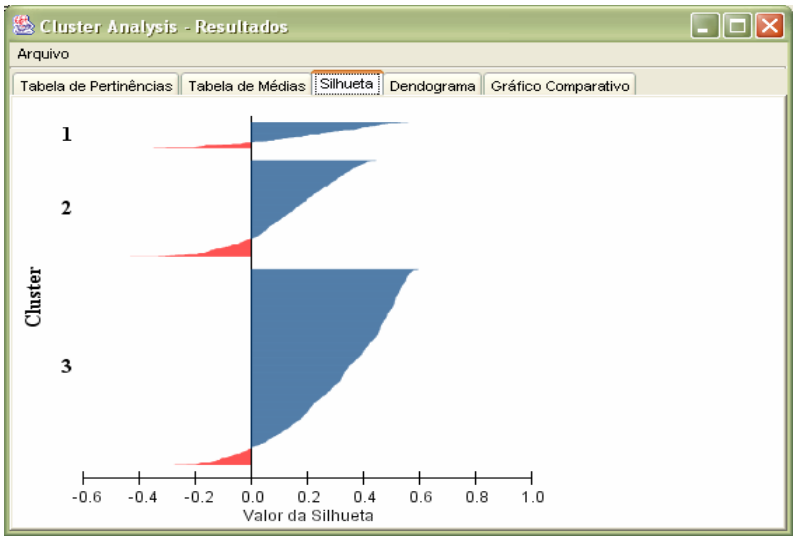

Figura 53: Base de Dados da Ilha do Governador – Gráfico da Silhueta (LINKAGE – Ward).

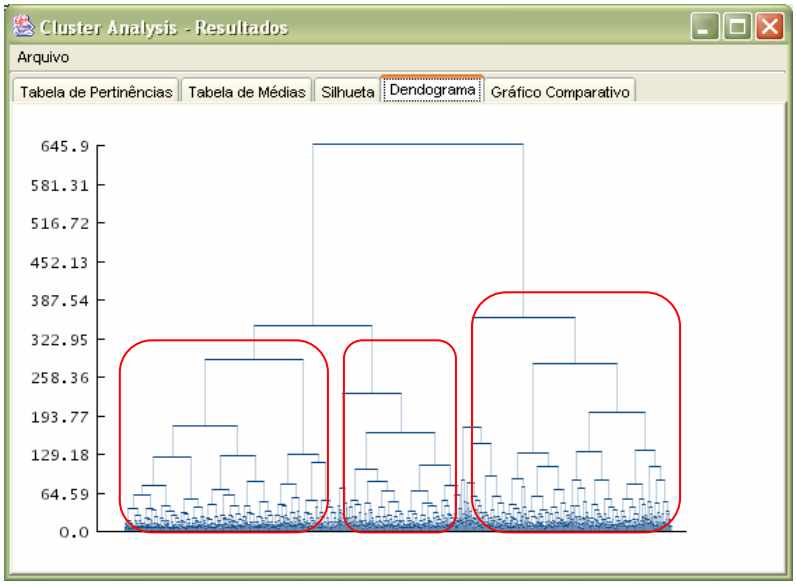

Figura 54: Base de Dados da Ilha do Governador – Dendograma (LINKAGE – Ward).

Pelo o dendograma da Figura 54, pode-se observar que há uma divisão clara entre os agrupamentos. O gráfico da silhueta da Figura 53 mostra uma estrutura entre os agrupamentos gerados muito melhor do que a apresentada pelo método anterior, porém a distribuição dos dados entre os agrupamentos está um pouco ruim.

- *Método Hierárquico Divisivo – DIANA*

  Resultados gerados:

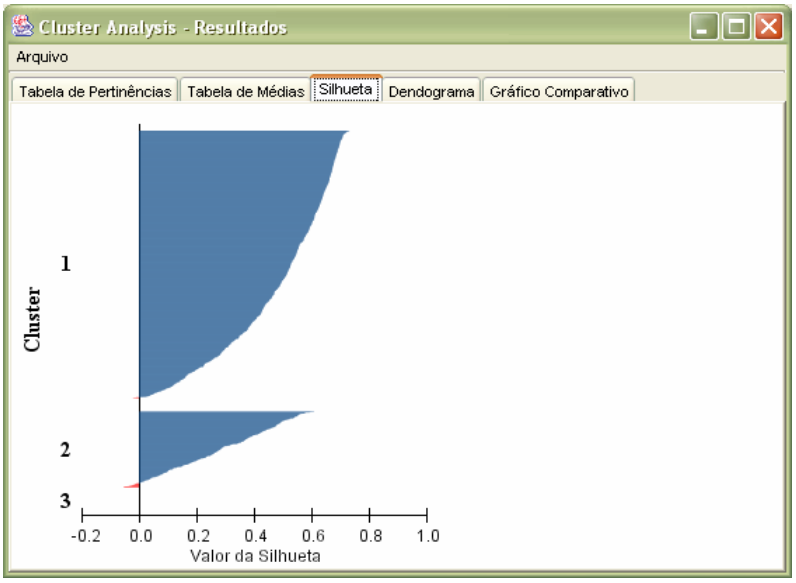

Figura 55: Base de Dados da Ilha do Governador – Gráfico da Silhueta (DIANA).

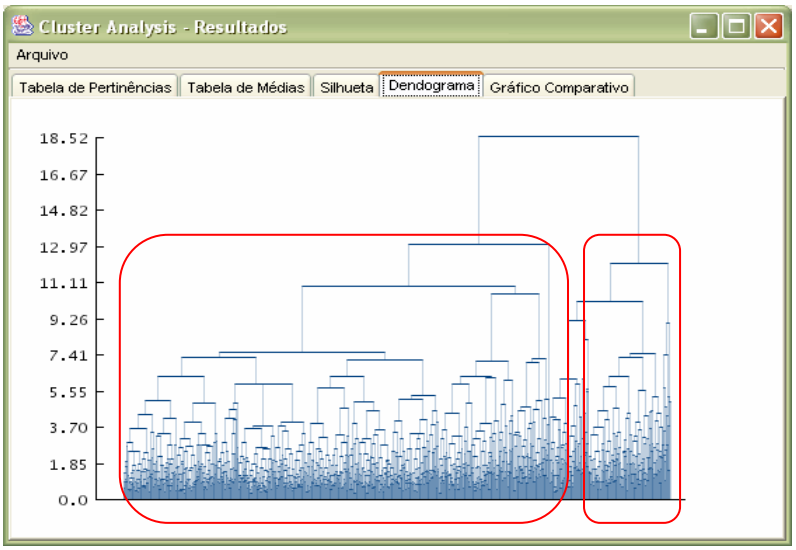

Figura 56: Base de Dados da Ilha do Governador – Dendograma (DIANA).

O dendograma da Figura 56 mostra uma divisão clara entre pelo menos dois agrupamentos, o que pode ser confirmado pelo gráfico da silhueta da Figura 55, onde se pode observar uma estrutura boa para dois dos três agrupamentos gerados.

A utilização desse método sugere a presença de dois agrupamentos, porém, como será visto mais adiante, o número ideal de agrupamentos para essa base são três, apesar da execução desse método indicar o contrário. Por isso é importante executar diferentes métodos a fim de se obter um melhor resultado.

A redução do número de agrupamentos reduz a interpretabilidade dos resultados, limitando o campo de visão sobre a estrutura dos dados. Ou seja, é importante encontrar um método que descreva a estrutura dos dados de forma razoável para o maior número de agrupamentos possível.

- *Método Particional Exclusivo – K-Means*

Resultados gerados:

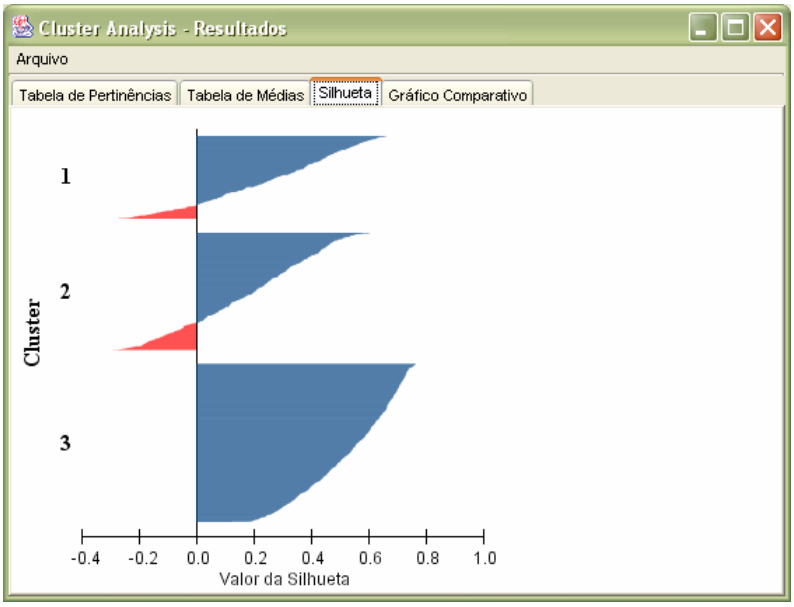

Figura 57: Base de Dados da Ilha do Governador – Gráfico da Silhueta (K-Means).

Analisando o gráfico da silhueta da Figura 57, pode-se observar a presença de três agrupamentos contendo um número significativo de dados bem agrupados.

- *Método Particional Exclusivo – Fuzzy C-Means*

Resultados gerados:

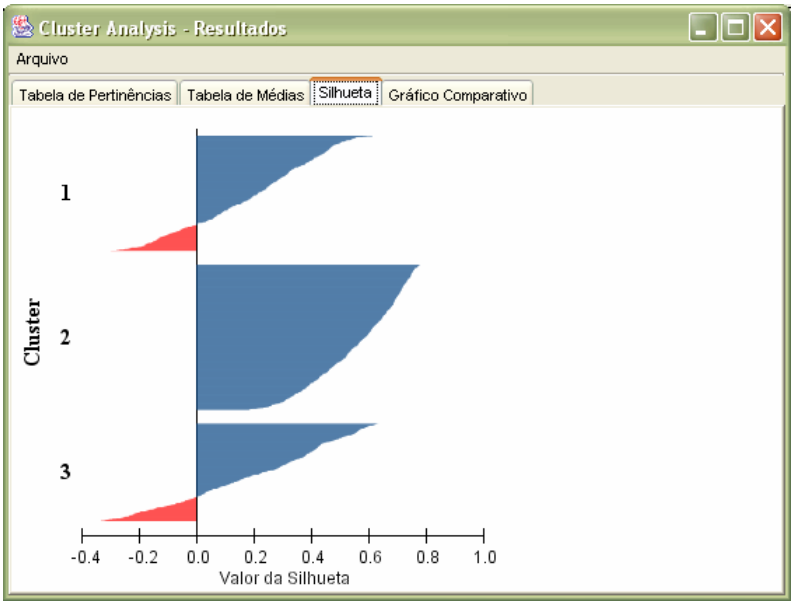

Figura 58: Base de Dados da Ilha do Governador – Gráfico da Silhueta (Fuzzy C-Means).

Pelo gráfico gerado (Figura 58), pode-se observar basicamente o mesmo resultado obtido através do método k-means, com a presença de três agrupamentos contendo um número significativo de dados bem agrupados. Mais adiante será discutido uma análise comparativa entre os métodos k-means e o fuzzy c-means.

Os métodos listados foram executados utilizando como medida de proximidade a distância euclidiana. Alguns desses métodos foram executados utilizando também a distância de Manhattam, porém não foram notadas nenhumas mudanças significativas nos resultados.

Após a execução de diversos métodos de diferentes tipos sobre essa base de dados é possível observar que os métodos que melhor descreveram de forma razoável sua estrutura foram os métodos k-means e o fuzzy c-means.

A partir da definição dos métodos que melhor descrevem os seus dados, é aconselhável fazer uma análise mais detalhada, como será feito a seguir.

As tabelas abaixo apresentam as estimativas de média de cada agrupamento utilizando os métodos k-means e FCM com coeficientes fuzzy m = 2 e m = 1, 5, respectivamente.

Tabela 14: K-Means: médias para cada agrupamento

Idade	Cigarros	Peso	Altura	Pressão Sistólica	Pressão Distólica	Total
40,49	16,481	77,55	171,8	131,52	84,099	405
36,48	6,5461	58,41	159,1	117,42	74,83	542
59,68	5,7123	65,76	156,1	157,01	91,6	285

Tabela 15: FCM com m=2: médias para cada agrupamento

Idade	Cigarros	Peso	Altura	Pressão Sistólica	Pressão Distólica	Total
40,393	16,122	77,45	171,9	131,43	84,142	394
33,882	6,2101	58,13	159,7	115,62	73,992	476
58,384	7,0249	65,24	156,2	151,48	89,37	362

Tabela 16: FCM com m=1,5: médias para cada agrupamento

Idade	Cigarros	Peso	Altura	Pressão Sistólica	Pressão Distólica	Total
40,456	15,917	77,71	172,1	131,78	84,196	397
34,76	6,5391	58,32	159,5	116,1	74,385	499
58,845	6,753	65,03	156	152,98	89,821	336

Embora se observe nas tabelas acima que os grupos formados pelos dois métodos são bem semelhantes, o FCM possibilita uma análise mais rica sobre a distribuição dos indivíduos nos agrupamentos, conforme se verá a seguir.

A Figura 59 apresenta uma divisão por sexo dos indivíduos, agrupados pelo FCM com m = 1,5. Observa-se que há, efetivamente, três grupos bem definidos; os agrupamentos de 1 a 3 são caracterizados, respectivamente, por conter indivíduos com médio, baixo e alto risco de hipertensão.

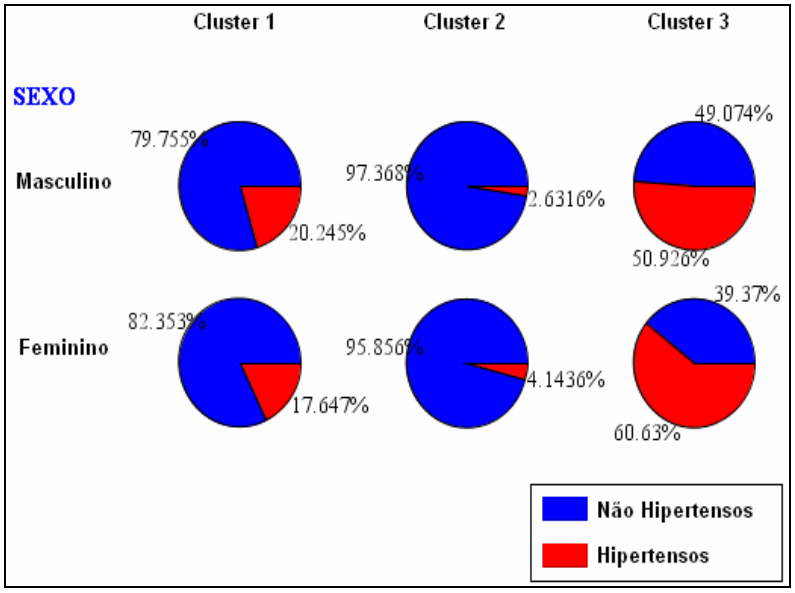

Figura 59: Distribuição de indivíduos nos agrupamentos

A tabela abaixo apresenta uma seleção de indivíduos que apresentam graus de pertinência altos e semelhantes a dois dos três agrupamentos, indicando que uma análise mais detalhada é aconselhável.

Tabela 17: Pertinências de indivíduos com graus de pertinência altos e semelhantes

Cluster 1	Cluster 2	Cluster 3
0,11391	0,45749	0,42859
0,048376	0,40946	0,54216
0,075556	0,51919	0,40525
0,097683	0,43017	0,47214
0,18912	0,41003	0,40085
0,062528	0,50821	0,42926
0,089248	0,48875	0,422
0,11252	0,43259	0,45489
0,084707	0,4415	0,47379
0,063739	0,44872	0,48754

Esses indivíduos apresentam as características explicitadas na tabela abaixo.

Tabela 18: Características dos indivíduos com graus de pertinência altos e semelhantes

Idade	Cigarros	Peso	Altura	Pressão Sistólica	Pressão Diastólica	Hipertensão
60	1	60	160	132	72	0
47	0	55	153	134	80	0
51	0	60	156	124	86	0
54	0	66	153	118	82	0
63	20	58	160	128	72	0
51	0	63	152	128	80	1
36	0	53	145	134	92	1
42	0	61	157	128	90	0
42	0	46	140	128	88	0
48	10	61	151	130	80	0

A média das observações pode ser observada na próxima tabela.

Tabela 19: Média dos indivíduos com graus de pertinência altos e semelhantes.

Idade	Cigarros	Peso	Altura	Pressão Sistólica	Pressão Diastólica	Hipertensão
53,571	2,7857	59,14	153,5	125,71	80	0,214

Pode-se observar que esses dados caracterizam indivíduos que em sua maioria não apresentam hipertensão, mas cujas idades e níveis de pressão arterial recomendam uma certa cautela em caracterizá-los como de baixo risco. Analisando os dados dessa forma, é possível extrair informações muito mais precisas e, nesse caso específico, permitir que um profissional da área de saúde tenha uma noção mais detalhada da situação do indivíduo e possa, com isso, adotar medidas preventivas.