



**Universidade Federal de Itajubá**

**Instituto de Matemática e Computação**

**Aprendizado de máquina fuzzy:  
potenciais contribuições da teoria fuzzy  
na concepção de técnicas de aprendizado de máquina  
para a tarefa de agrupamento**

**RELATÓRIO FINAL  
PROGRAMA PIVIC**

**CICLO 2019/2020**

**Aluno: Thiago Silva Pereira  
Matrícula: 2018008209  
Curso: Ciência da Computação  
Orientador: Isabela Neves Drummond**

**Fase/Período: 3º ano, 6º período**

**Vigência: setembro de 2019 a setembro de 2020**

## RESUMO

Este trabalho de iniciação científica tem como objetivo principal avaliar um conjunto de modelos de agrupamento baseados na lógica *fuzzy*, buscando demonstrar suas principais características. O projeto apresenta um estudo comparativo de modelos de agrupamento clássicos e *fuzzy*, através de um conjunto de experimentos em diferentes bases de dados. A aquisição de uma enorme quantidade de dados na última década demonstra a necessidade de estudos que envolvem o desenvolvimento de ferramentas automáticas que possam extrair dos dados conhecimento relevante para as mais variadas áreas e aplicações. Todas as implementações envolvidas neste projeto foram desenvolvidas em linguagem Python com o auxílio da ferramenta VScode, plataformas de código aberto como o Scikit-Learn e repositórios do Github. Foram selecionados seis (6) algoritmos para estudo, três clássicos: K-médias, Aglomerativo e Mean Shift; e três *fuzzy*: C-médias, C-médias possibilístico e Gustafson Kessel. A metodologia de trabalho é dividida em três experimentos. O experimento de validação emprega um conjunto de 6 bases de dados sintéticos criadas pelo autor, sendo conjuntos de dados em duas dimensões, facilitando a visualização gráfica do agrupamento. O experimento com bases de dados clássicas foi executado com 3 bases: “Íris”, “Wine” e “Boston”, disponíveis no repositório de dados do Scikit-Learn. E por fim, um experimento com outras duas bases de dados selecionadas aleatoriamente na plataforma do Kaggle: “Seguro Clínico” e “Consumo Online”. Na última etapa da pesquisa, é aplicada uma fase de ensemble, onde os algoritmos são combinados buscando melhores resultados. Durante a análise de resultados são demonstradas as principais características de cada um dos algoritmos, destacando: a velocidade do modelo Aglomerativo e Possibilístico, a automatização do Mean Shift e seus possíveis erros derivados, a convergência de centros possibilistas e a influência dos atributos sobre o algoritmo de Gustafson Kessel. O comparativo realizado permitiu verificar que os modelos que empregam lógica *fuzzy* tendem a apresentar agrupamentos com um maior número de clusters, e uma igualdade na divisão geométrica dos grupos. Foi possível perceber ainda que, a repetição de várias execuções dos modelos *fuzzy*, para a aplicação do ensemble, aumentou muito o custo computacional, quando comparado com a versão clássica dos modelos.

Palavras-chave: Aprendizado de máquina, Agrupamento clássico, Agrupamento *Fuzzy*.

## LISTA DE ILUSTRAÇÕES

Figura 1	A Hierarquia do Aprendizado. Fonte: (MONTEIRO, 2018) . . . . .	15
Figura 2	Exemplo do funcionamento do k-médias, fonte: (PIECH, 2013) . . . . .	17
Figura 3	Exemplo de um dendrograma representativo de um hierarquia de agrupamento, fonte: (BUITINCK et al., 2013) . . . . .	18
Figura 4	Esquema da divisão do banco de dados . . . . .	27
Figura 5	Fluxo da primeira fase de testes . . . . .	29
Figura 6	Figuras ilustrativas de cada base de dados sintética. . . . .	30
Figura 7	Fluxo da segunda fase de testes . . . . .	30
Figura 8	Fluxo da terceira fase de testes . . . . .	32
Figura 9	Coeficientes k-médias - Tríade (de 2 a 10) . . . . .	36
Figura 10	Coeficientes k-médias - Tríade (de 2 a 31) . . . . .	36
Figura 11	Coeficientes Aglomerativo - Tríade (de 2 a 10) . . . . .	37
Figura 12	Coeficientes Aglomerativo - Tríade (de 2 a 31) . . . . .	37
Figura 13	Coeficientes c-médias - Tríade (de 2 a 10) . . . . .	38
Figura 14	Coeficientes c-médias - Tríade (de 2 a 31) . . . . .	38
Figura 15	Coeficientes Gustafson - Tríade (de 2 a 10) . . . . .	39
Figura 16	Coeficientes Gustafson - Tríade (de 2 a 31) . . . . .	39
Figura 17	Coeficientes possibilístico - Tríade (de 2 a 10) . . . . .	40
Figura 18	Coeficientes possibilístico - Tríade (de 2 a 31) . . . . .	40
Figura 19	Coeficientes Aglomerativo - Anel Interno (de 2 a 10) . . . . .	41
Figura 20	Coeficientes Aglomerativo - Anel Interno (de 2 a 31) . . . . .	41
Figura 21	Agrupamentos trabalhados na base de dados "Tríade" . . . . .	43
Figura 22	Agrupamentos trabalhados na base de dados "Grupo de Anéis" . . . . .	44
Figura 23	Agrupamentos trabalhados na base de dados "Interlaço" . . . . .	46
Figura 24	Agrupamentos trabalhados na base de dados "Anel Interno" . . . . .	48
Figura 25	Agrupamentos trabalhados na base de dados "Ruídos" . . . . .	49
Figura 26	Agrupamentos trabalhados na base de dados "Luas" . . . . .	51
Figura 27	Resultado gráfico representativo do método aglomerativo agrupando o ponto isolado da base de seguros clínicos . . . . .	58
Figura 28	Resultado gráfico representativo do método do k-médias agrupando os custos da base de seguros clínicos . . . . .	58
Figura 29	Agrupamento FCM para Gênero X Idade do seguro clínico . . . . .	59
Figura 30	Agrupamento FCM para Região X Gênero do seguro clínico . . . . .	59
Figura 31	Agrupamento FCM para Custo X Gênero do seguro clínico . . . . .	60
Figura 32	Agrupamento FCM para nº filhos X IMC do seguro clínico . . . . .	60
Figura 33	Agrupamento FCM para Fumante X IMC do seguro clínico . . . . .	61

Figura 34	Agrupamento FCM para Região X IMC do seguro clínico . . . . .	61
Figura 35	Agrupamento FCM para Custo X IMC do seguro clínico . . . . .	62
Figura 36	Agrupamento FCM para Fumante X nº filhos do seguro clínico . . .	62
Figura 37	Agrupamento FCM para Região X nº filhos do seguro clínico . . . .	63
Figura 38	Agrupamento FCM para Custo X nº filhos do seguro clínico . . . . .	63
Figura 39	Agrupamento FCM para Região X Fumante do seguro clínico . . . .	64
Figura 40	Agrupamento FCM para IMC X Idade do seguro clínico . . . . .	64
Figura 41	Agrupamento FCM para Custo X Fumante do seguro clínico . . . .	65
Figura 42	Agrupamento FCM para Custo X Região do seguro clínico . . . . .	65
Figura 43	Agrupamento FCM para nº filhos X Idade do seguro clínico . . . .	66
Figura 44	Agrupamento FCM para Fumante X Idade do seguro clínico . . . .	66
Figura 45	Agrupamento FCM para Região X Idade do seguro clínico . . . . .	67
Figura 46	Agrupamento FCM para Custo X Idade do seguro clínico . . . . .	67
Figura 47	Agrupamento FCM para IMC X Gênero do seguro clínico . . . . .	68
Figura 48	Agrupamento FCM para nº filhos X gênero do seguro clínico . . . .	68
Figura 49	Agrupamento FCM para Fumante X Gênero do seguro clínico . . . .	69
Figura 50	Resultado gráfico representativo do método de Gustafson Kessel agrupando a base de seguros clínicos segundo idade x custos . . . . .	69
Figura 51	Agrupamento K-médias para Preço X Quantidade da base de consumos online . . . . .	70
Figura 52	Agrupamento K-médias para Cliente X Quantidade da base de consumos online . . . . .	71
Figura 53	Agrupamento K-médias para País X Quantidade da base de consumos online . . . . .	71
Figura 54	Agrupamento K-médias para Cliente X Preço da base de consumos online . . . . .	72
Figura 55	Agrupamento K-médias para País X Preço da base de consumos online	72
Figura 56	Agrupamento K-médias para País X Cliente da base de consumos online	73
Figura 57	Agrupamento Gustafson Kessel para Cliente X Preço da base de consumos online . . . . .	73
Figura 58	Agrupamento das bases Tríade, Grupo de Anéis e Interlaço, realizado pelas técnicas de ensemble. . . . .	74
Figura 59	Agrupamento das bases Anel Interno, Ruídos e Luas, realizado pelas técnicas de ensemble. . . . .	75

## LISTA DE TABELAS

Tabela 1	Tipo de Ligações que podem ocorrer um agrupamento hierárquico . . . . .	18
Tabela 2	Métricas não-supervisionadas usadas para validação interna dos métodos crisp e posteriormente para comparação e validação relativa entre os algoritmos. . . . .	23
Tabela 3	Métricas não-supervisionadas usadas para validação interna dos métodos <i>fuzzy</i> , fazendo uso da Tabela de pertinência. Os algoritmos foram fornecidos pelo repositório (GERMER, 2017) . . . . .	23
Tabela 4	Métricas com base supervisionada, usam rótulos para validar. Os algoritmos foram fornecidos por Scikit-Learn(BUITINCK et al., 2013), valores mais altos significam melhor compatibilidade e 1 é o valor máximo.	24
Tabela 5	Resumo dos algoritmos utilizados e seus parâmetros . . . . .	26
Tabela 6	Tabela com o número de grupos para cada um dos algoritmos referentes a cada base sintética. . . . .	40
Tabela 7	Resultados das métricas para comparação relativa com a base de dados 'tríade' . . . . .	42
Tabela 8	Resultados das métricas para comparação relativa com a base de dados 'Grupo de Anéis'. . . . .	44
Tabela 9	Resultados das métricas para comparação relativa com a base de dados 'Interlaço'. . . . .	45
Tabela 10	Resultados das métricas para comparação relativa com a base de dados 'Anel Interno'. . . . .	47
Tabela 11	Resultados das métricas para comparação relativa com a base de dados 'Ruídos'. . . . .	47
Tabela 12	Resultados das métricas para comparação relativa com a base de dados 'Luas'. . . . .	50
Tabela 13	Número de grupos para cada um dos algoritmos referentes a cada base científica. . . . .	52
Tabela 14	Resultados dos coeficientes supervisionados com a base de dados 'íris'. . . . .	52
Tabela 15	Resultados dos coeficientes supervisionados com a base de dados 'Wine'. . . . .	53
Tabela 16	Resultados das métricas para comparação relativa com a base de dados 'Iris'. . . . .	54
Tabela 17	Resultados das métricas para comparação relativa com a base de dados 'Wine'. . . . .	55
Tabela 18	Resultados das métricas para comparação relativa com a base de dados 'Boston'. . . . .	55

Tabela 19	Tabela com o número de grupos para cada um dos algoritmos referentes a cada base prática. . . . .	56
Tabela 20	Resultados das métricas para comparação relativa com a base de dados 'Seguro Clínico' . . . . .	56
Tabela 21	Resultados das métricas para comparação relativa com a base de dados 'Consumo Online' . . . . .	70
Tabela 22	Resultados dos coeficientes do ensemble da base de Íris . . . . .	72
Tabela 23	Resultados dos coeficientes do ensemble da base de Wine . . . . .	73
Tabela 24	Resultados dos coeficientes do ensemble da base de Boston . . . . .	74

## **LISTA DE ABREVIATURAS E SIGLAS**

AM Aprendizagem de máquina

FCM Fuzzy c-médias

GK Gustafson Kessel

IA Inteligência Artificial

PCM Possibilitic c-means

## SUMÁRIO

<b>1</b>	<b>INTRODUÇÃO</b>	<b>10</b>
<b>2</b>	<b>OBJETIVOS PROPOSTOS</b>	<b>13</b>
<b>3</b>	<b>REFERENCIAL TEÓRICO</b>	<b>14</b>
3.1	Aprendizado de máquina	14
3.2	Agrupamento	14
3.3	Lógica Clássica e <i>Fuzzy</i>	15
3.4	Modelos De Agrupamento	16
3.4.1	K-médias	16
3.4.2	Aglomerativo	17
3.4.3	Mean Shift	18
3.4.4	C-médias	19
3.4.5	Gustafson Kessel	20
3.4.6	C-médias Possibilístico	21
3.5	Validando Agrupamentos	21
3.5.1	Validação Interna	22
3.5.2	Validação Externa	22
3.5.3	Validação Relativa	22
3.5.4	Métricas empregadas nesta pesquisa	22
3.6	<i>Ensemble</i> de agrupamentos	24
<b>4</b>	<b>DESCRIÇÃO DAS ATIVIDADES DESENVOLVIDAS</b>	<b>26</b>
4.1	Algoritmos selecionados para estudo	26
4.2	Metodologia de testes e avaliações	27
4.2.1	Parametrização dos modelos	28
4.2.2	Experimento de validação	29
4.2.3	Experimentos em bases clássicas	30
4.2.3.1	Iris	31
4.2.3.2	Wine	31
4.2.3.3	Boston	31
4.2.4	Experimentos em outras bases de dados	32
4.2.4.1	Seguro Médico	32
4.2.4.2	Consumo Online	33
4.2.5	<i>Ensemble</i>	33
<b>5</b>	<b>RESULTADOS OBTIDOS E ANÁLISE</b>	<b>35</b>

5.1	Experimento de validação . . . . .	35
5.1.1	Análise do número de partições . . . . .	35
5.1.2	Visualização dos agrupamentos . . . . .	41
5.1.2.1	Tríade . . . . .	42
5.1.2.2	Grupo de Anéis . . . . .	43
5.1.2.3	Interlaço . . . . .	44
5.1.2.4	Anel Interno . . . . .	45
5.1.2.5	Ruídos . . . . .	46
5.1.2.6	Luas . . . . .	49
5.2	Experimentos em bases clássicas . . . . .	50
5.2.1	Número de Clusters . . . . .	50
5.2.2	Análise do agrupamento . . . . .	52
5.2.2.1	Validação Supervisionada . . . . .	52
5.2.2.2	Íris . . . . .	53
5.2.2.3	Wine . . . . .	54
5.2.2.4	Boston . . . . .	55
5.3	Experimentos em outras bases de dados . . . . .	56
5.3.1	Seguro Clínico . . . . .	56
5.3.2	Consumo Online . . . . .	61
5.4	Considerações finais . . . . .	65
5.5	<i>Ensemble</i> por votação . . . . .	68
5.5.1	Bases de dados de Validação . . . . .	69
5.5.2	Bases de dados Clássicas . . . . .	71
5.5.3	Outras bases de dados . . . . .	76
<b>6</b>	<b>CONCLUSÃO . . . . .</b>	<b>77</b>
	<b>REFERÊNCIAS BIBLIOGRÁFICAS . . . . .</b>	<b>79</b>

## 1 INTRODUÇÃO

Este trabalho de iniciação científica apresenta um estudo comparativo sobre os modelos de agrupamento clássicos e *fuzzy*, visando avaliar a aplicação da lógica *fuzzy* no aprendizado de máquina. A forma clássica de agrupar um conjunto de dados consiste em separar elementos em grupos contendo elementos similares, ou seja, indivíduos com características parecidas tendem a ficar na mesma partição, enquanto os dissimilares pertencem a grupos distintos; consistindo assim numa tarefa capaz de prover novas informações para um problema. Já a versão *fuzzy* do agrupamento segue o mesmo princípio, porém, além de definir as partições emprega também uma matriz referente a pertinência de um indivíduo a cada grupo gerado.

Um sistema clássico, ou também denominado *crisp*, emprega as ideias comuns de conjuntos na matemática básica, onde um ponto é definido como pertencente(1) ou não pertencente(0) a um determinado grupo. Já um sistema *fuzzy* (no português difuso ou nebuloso) trabalha com níveis de pertinências, onde cada ponto tem associado um valor no intervalo [0, 1] com relação a cada partição. A lógica *fuzzy* é desenvolvida em um teor linguístico com o objetivo de representar a imprecisão ou incerteza contida na informação. Sua aplicação nas principais tarefas de aprendizado de máquina, agrupamento e classificação, é sem dúvida uma área de estudo promissora. Algumas referências na literatura são (DEMBELE; KASTNER, 2003) e (MALHOTRA; KAUR; ALAM, 2014).

Alguns estudos já realizados no passado em cima da base teórica do agrupamento *fuzzy*, levaram pesquisadores a especular que esse é um método ineficaz ou que trás pouca relevância (MANY, 2007). Todavia essa é uma área que cresceu bastante na comunidade ao longo dos anos, novas métricas e implementações surgem para auxiliar e otimizar ambas abordagens sobre conjuntos. Este trabalho de pesquisa tem como objetivo principal estudar e testar um conjunto de algoritmos de agrupamento baseados tanto na lógica clássica, quanto na lógica *fuzzy*, utilizando das diversas métricas de validação descritas na literatura para tentar quantificar vantagens e desvantagens de cada lógica; buscando ainda apontar características que demonstrem as vantagens e desvantagens dos modelos estudados.

Como ferramenta de aplicação optou-se por utilizar implementações na linguagem Python (ROSSUM; JR, 1995). Trata-se de uma linguagem com um grande número de usuários e com a disponibilidade de bibliotecas de aprendizado de máquina que se apresentam adequadas para este estudo. Foram empregados também diversos algoritmos disponíveis pela plataforma do scikit-learn (BUITINCK et al., 2013) e seu derivado scikit-fuzzy (99991, 2019).

A metodologia que é apresentada na Seção 4.2, define um plano de testes dividido em três fases com uma revisão final fazendo uso de um modelo *ensemble*. Esse projeto é traçado utilizando as técnicas de validação descritas em 3.5, propondo demonstrar o funcionamento de 6 modelos de agrupamento, posteriormente fazendo a junção desses em um agrupamento múltiplo

descritivo para mais informações.

As bases de dados que são usadas na realização dos testes são também separadas por cada fase da metodologia. Estão presentes na primeira fase, 6 bancos criados sinteticamente pelo autor, usando como referência o artigo (DAVE, 1996) e figurações comumente usadas para testes de agrupamento. Na segunda fase trabalha-se com três bases clássicas na comunidade científica, sendo elas fornecidas pelo próprio scikit-learn: o Íris<sup>1</sup>, Wine<sup>2</sup> e Boston<sup>3</sup>. Por último, são realizados testes práticos em bases de dados selecionadas na internet: um banco com informações sobre os gastos com seguros clínicos e um sobre compras realizadas online. Todas essas bases são melhor descritas em suas respectivas seções da metodologia.

Durante a fase de análises são destacadas características importantes presentes em cada algoritmo individualmente, com o avançar dos estudos, características que podem ser dadas como particularidades de cada lógica também surgem para dissertação. Esta pesquisa busca apontar trabalhos futuros a partir dos resultados alcançados.

Em um âmbito mais conclusivo, aponta-se duas características que parecem ser únicas à lógica *fuzzy* e que podem ser úteis quando utilizadas de maneira apropriada: uma procura por divisão equivalente ou "simétrica" dada pela divisão por pesos e uma facilidade de agrupar um maior número de partições. Além desses destaques principais, mostra-se também o potencial equivalente do uso dos modelos *fuzzy* perante os resultados clássicos, onde muito dos coeficientes tiveram respostas similares ou, em alguns casos, até mais eficiente. No lado negativo, também são descritos alguns fatores que tornam trabalhar com a lógica *fuzzy* um pouco mais complexo, como a dificuldade na escolha de um número de partições e visualização de alguns gráficos, mas não impedem sua execução. Outra desvantagem relacionada aos modelos *fuzzy* é o custo computacional, em execuções diretas ou que não exigem o retrabalho dos algoritmos, os resultados foram muito bons e eficientes, porém em casos como o *ensemble* que utiliza várias execuções dos modelos, as respostas exigiram mais que o dobro do esforço computacional, tempo de execução muito elevado.

Este relatório está dividido da seguinte maneira:

- No Capítulo 2 são apresentados os principais objetivos com a pesquisa e interesses específicos;
- o Capítulo 3 contém um levantamento da literatura para compreensão dos principais conceitos trabalhados na pesquisa, onde é apresentado a definição de agrupamento e os principais modelos;

---

<sup>1</sup> <[https://scikit-learn.org/stable/modules/generated/sklearn.datasets.load\\_iris.html?highlight=iris#sklearn.datasets.load\\_iris](https://scikit-learn.org/stable/modules/generated/sklearn.datasets.load_iris.html?highlight=iris#sklearn.datasets.load_iris)>

<sup>2</sup> <[https://scikit-learn.org/stable/modules/generated/sklearn.datasets.load\\_wine.html#sklearn.datasets.load\\_wine](https://scikit-learn.org/stable/modules/generated/sklearn.datasets.load_wine.html#sklearn.datasets.load_wine)>

<sup>3</sup> <[https://scikit-learn.org/stable/modules/generated/sklearn.datasets.load\\_boston.html#sklearn.datasets.load\\_boston](https://scikit-learn.org/stable/modules/generated/sklearn.datasets.load_boston.html#sklearn.datasets.load_boston)>

- o Capítulo 4 detalha a metodologia trabalhada para a execução dos testes propostos;
- no Capítulo 5 é feita a análise dos resultados com demonstrações gráficas e explicações contextualizadas;
- e por fim, o Capítulo 6 traz as conclusões sobre as principais características estudadas ao longo dos processos, e possíveis trabalhos futuros.

## 2 OBJETIVOS PROPOSTOS

A teoria *fuzzy* surgiu como uma forma de representar a informação vaga, o que permitiu melhor adequação da representação de conjuntos. Por exemplo, no idioma português o emprego das palavras “mais ou menos”, “um pouco frio”, “muito quente”, “alto” ou “médio”, são modificadores utilizados no dia a dia, que se representam na máquina empregando a lógica clássica, não reflete de fato o significado real da informação. A representação dos conjuntos *fuzzy* de Zadeh permite definir objetos que pertencem mais ou menos a mais de um conjunto (ZADEH, 1965).

Compreendendo esse propósito fica mais fácil visualizar o potencial da aplicação da técnica *fuzzy* no âmbito do agrupamento. O objetivo principal deste trabalho de pesquisa é avaliar um conjunto de modelos de agrupamento baseados na lógica *fuzzy*, buscando demonstrar suas principais características.

Dentre os objetivos específicos desta pesquisa destacam-se:

- Estudo e seleção de um conjunto de algoritmos.
- Definição das métricas de comparação entre os algoritmos.
- Definição dos conjuntos de dados para validação dos modelos.
- Seleção de bases de dados disponíveis para testes de modelos de agrupamento e classificação de dados.
- Definição da metodologia de testes.
- Execução dos experimentos e análise dos resultados.

Pretende-se ainda responder um conjunto de questões que possibilite compreender o papel da lógica *fuzzy* no modelo de agrupamento estudado.

Por fim, a última etapa compreende o estudo da combinação das técnicas implementadas, um *ensemble* por votação, uma técnica com o objetivo de unificar os algoritmos e tentar compreender melhor outras características mais específicas, buscando melhores resultados.

### 3 REFERENCIAL TEÓRICO

A revisão de alguns conceitos é essencial para a compreensão desta pesquisa. Neste Capítulo é apresentada uma breve explicação sobre o aprendizado de máquina e a tarefa do agrupamento, bem como os principais modelos de agrupamento clássicos e *fuzzy*. Também é realizada uma introdução ao *ensemble*, que consiste na combinação de resultados obtidos por diferentes algoritmos de agrupamento.

#### 3.1 APRENDIZADO DE MÁQUINA

A Inteligência Artificial (IA) é, em resumo, uma área da computação que estuda modelos matemáticos buscando desenvolver uma máquina inteligente. Apesar das diversas discussões acerca do termo inteligência, é definido como um dos seus principais atributos a capacidade de aprender. Surge-se então o campo de pesquisa denominado Aprendizado de Máquina (AM).

Segundo (COPPIN, 2010), a Inteligência Artificial envolve a utilização de métodos baseados no comportamento inteligente de humanos e outros animais para solucionar problemas complexos. O AM por sua vez, é a ciência que estuda e trabalha com tecnologias capazes de aprender com a experiência, uma área composta de diversos algoritmos capazes de induzir com suas entradas uma nova hipótese ou funções para solucionar o problema.

A indução de hipótese é um recurso muito utilizado até mesmo pelo próprio cérebro humano (MONARD; BARANAUSKAS, 2003), uma técnica que, para um conjunto de dados de treinamento informado, é capaz de criar um modelo capaz de deduzir regras generalizadas que sejam válidas para outras entradas fora de sua base de conhecimento. Esta é a característica principal de um bom modelo de AM, o quanto genérico ele se apresenta (FACELI, 2011).

O foco principal dessa prática é criar modelos de generalização capazes de conduzir novas informações em um banco de dados qualquer, ou desempenhar a tarefa de classificação, o resumo de informações e o agrupamento de dados similares por exemplo. É determinado, com base nestes fatores, uma relação hierárquica que compõe os ramos do AM, onde são apresentados grupos de tarefas, como pode ser observado na Figura 1.

#### 3.2 AGRUPAMENTO

Este trabalho de pesquisa se dedica ao estudo da tarefa de agrupamento, desempenhada pelos modelos não supervisionados. Na tarefa de agrupamento, apresentando uma base de dados, deseja-se partioná-la em 'N' grupos de elementos similares, objetivando a retirada de novas informações do ambiente. Sua aplicação está presente no *data mining* e na segmentação de imagens, produtos de software que se tornaram essenciais na vida moderna.

Diversos algoritmos são propostos na literatura, e geralmente se diferenciam pelas diferentes métricas de similaridade empregadas. O termo *cluster* é empregado como uma



Figura 1 – A Hierarquia do Aprendizado. Fonte: (MONTEIRO, 2018)

nomenclatura para os grupos formados pelos modelos de agrupamento, já que o problema abrange uma variedade de soluções diferentes que tornam o conceito de grupo, propriamente dito, muito abstrato (FACELI, 2011).

Diferente da classificação de dados, um ramo supervisionado da AM, o agrupamento é realizado sem nenhum referencial inicial, ou seja, o algoritmo deve ser capaz de observar e deduzir novas informações sem nenhum comparativo externo. Isto gera um dos principais problemas tratado nesta tarefa, a validação dos grupos obtidos.

Dentre os modelos de agrupamento, destacam-se neste trabalho também os modelos que empregam a lógica *fuzzy*. A teoria dos conjuntos *fuzzy* tem aplicação na área de reconhecimento de padrões desde sua criação, envolvendo a tarefa de classificação.

### 3.3 LÓGICA CLÁSSICA E FUZZY

A teoria clássica dos conjuntos descrita na matemática é aquela que define um ponto como pertencente ou não a um determinado grupo. Desta forma, numa tarefa de classificação, por exemplo, os elementos pertencentes a um conjunto são aqueles que representam exatamente o conceito expresso pelo conjunto. A informação que o ser humano trabalha nos problemas do dia a dia envolve imprecisão, e faz-se necessário uma teoria matemática que possilita a representação do dado impreciso, vago, incerto, incompleto ou até mesmo inconsistente.

A proposta da chamada lógica *fuzzy* envolve um elemento pertencer a um conjunto com um grau associado. Este valor, chamado de pertinência do elemento é um valor entre 0 e 1, podendo ser igual a 0 ou igual a 1, o que faz da Lógica *fuzzy* uma extensão da lógica clássica.

No aprendizado de máquina, para a tarefa de agrupamento envolver a lógica *fuzzy* nos modelos de agrupamento é extremamente interessante, já que muitos dos conjuntos de dados aos quais se aplica a tarefa de agrupamento envolve algum tipo de imprecisão.

Neste trabalho de iniciação científica o foco está no estudo do emprego da lógica *fuzzy*

nas técnicas de agrupamento, e os experimentos propostos tentar identificar características que demonstrem o potencial da lógica *fuzzy* nos modelos não supervisionados de aprendizado de máquina.

### 3.4 MODELOS DE AGRUPAMENTO

Nesta Seção são detalhados os modelos selecionados para estudo. São três modelos de agrupamento clássicos: K-médias, aglomerativo e Mean-shift; e três modelos que envolvem a lógica *fuzzy*: C-médias, Gustafson-Kessel e C-médias possibilísticos.

#### 3.4.1 K-MÉDIAS

O algoritmo K-médias é considerado uma das técnicas mais clássicas para solucionar a tarefa de agrupamento. Atualmente já possui mais de cinquenta anos nesse ramo de pesquisa e continua sendo amplamente utilizado pela sua ótima performance e resultados. Um dos principais artigos que descreve a história e modelagem desse algoritmo pode ser encontrado em (JAIN, 2010).

O k-médias inicializa com a escolha de  $K$  centroides aleatórias para seus clusters,  $K$  sendo um número de partições definido pelo usuário. Em cada etapa, o algoritmo desloca essas centroides objetivando encontrar a melhor combinação de partições, nesse caso sendo a divisão com menor erro quadrático possível (Equação 3.1). Esta equação basicamente é a representação do quão distante está uma partição das demais e o quão próxima estão as informações pertencentes a um mesmo grupo.

O erro quadrático também é considerado como um meio de validação de clusters, podendo até ser usado com a técnica do joelho de Elbow para encontrar um número adequado de clusters (Vide Seção 3.5).

$$J = \sum_{i=0}^n \min(||x_i - v_j||^2) \quad (3.1)$$

A cada iteração o K-médias procura atingir seu objetivo atribuindo a cada elemento do conjunto a partição com maior similaridade. Esta similaridade é calculada através de medidas como distância euclidiana, Manhattan ou Chebyschev, onde cada atributo de um dado é considerado uma dimensão para o cálculo. A distância é medida entre o elemento do conjunto e os centroides de cada partição, e o elemento é atribuído à partição cujo centroide é mais próximo. O deslocamento dos centroides, que objetiva diminuir o erro quadrático, é recalculado pela médias das informações atualmente presentes naquela partição a cada iteração. Quando um erro mínimo ou um número máximo de iterações é atingido, o algoritmo se encerra e a formação atual dos clusters é apresentada como resposta. O modelo também pode acabar se encontrando em um ciclo onde não há mais grande variação na posição dos centroides, também podendo ser este outro parâmetro de encerramento.

Uma das principais dificuldades na utilização da maioria dos modelos de agrupamento, incluindo o k-médias, é a descoberta de um número de partições apropriado para subdividir os dados, algo que geralmente deve ser feito pelo usuário da aplicação, o que torna esse processo menos automático e mais dependente de fatores humanos. Nesse trabalho as técnicas usadas para contornar esse problema são apresentadas na Seção 4.2.1, onde são apresentadas as métricas utilizadas nas validações dos algoritmos. Em alguns casos esse valor já é pré-definido no próprio “enunciado”, ou é mais simples de ser obtido, todavia mesmo usando uma heurística como base, impor esses valores pode influenciar na eficiência das respostas e deve ser utilizado com cautela.

A complexidade do K-médias é conhecida por ser  $O(n^2)$  (Pakhira, 2014), um resultado polinomial considerado bem eficiente e por isso ainda é altamente utilizado. Na Figura 2 é observado um exemplo de iterações do modelo.

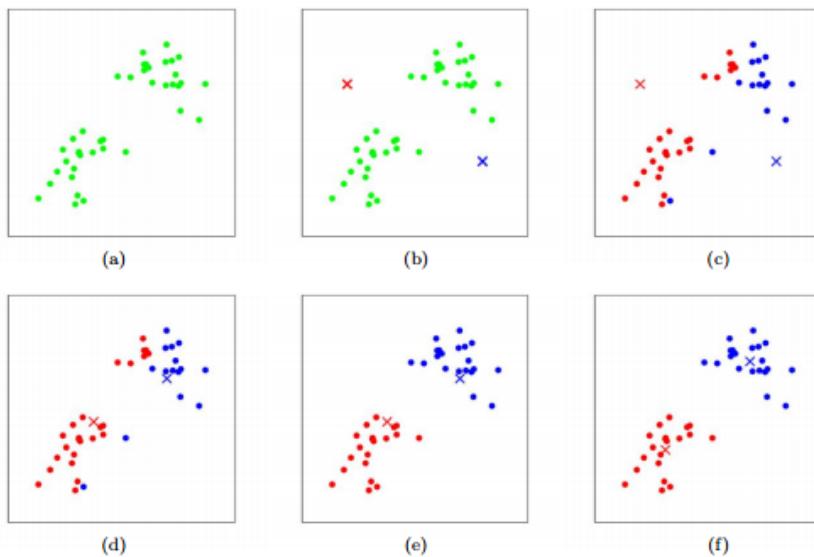


Figura 2 – Exemplo do funcionamento do k-médias, fonte: (PIECH, 2013)

### 3.4.2 AGLOMERATIVO

O algoritmo aglomerativo deriva de uma versão “*bottom-up*” dos modelos de agrupamento hierárquicos, muito conhecidos pela geração de seus dendrogramas de junção, exemplo Figura 3. A ideia desse sistema constitui-se em calcular a similaridade entre clusters, ao invés dos dados do banco em específico, unindo as partições mais próximas em um novo grupo pai. Na aplicação inversa “*top-down*”, tem-se um único cluster primário que é subdividido gerando outros grupos menores (K.SASIREKHA; P.BABY, 2013).

Mais uma vez o usuário administrando o modelo deve fornecer a quantidade de grupos encerrada, todavia o algoritmo permite que com apenas um corte na criação do dendrograma, ou um critério de parada simples, possa-se alterar esse valor sem grandes custos. Assim, o principal problema nessa metodologia se aplica em como calcular a similaridade de clusters, onde existem cinco métodos mais comumente aplicados (vide Tabela 1) (Takumi; Miyamoto, 2012).

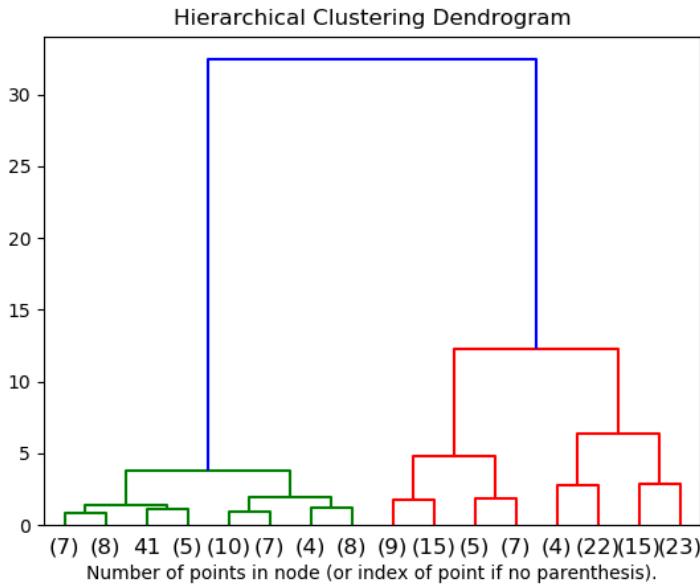


Figura 3 – Exemplo de um dendrograma representativo de um hierarquia de agrupamento, fonte: (BUITINCK et al., 2013)

Tabela 1 – Tipo de Ligações que podem ocorrer um agrupamento hierárquico

Ligaçāo	Tipo	Descriçāo
Ligaçāo mínima	(single)	Usa a distância dos pontos mais próximo entre os cluster.
Ligaçāo média	(average)	Usa a média das distâncias entre todos os pares de pontos que ligam os dois clusters.
Ligaçāo máxima	(complete)	Usa a distância dos pontos mais afastado entre os cluster.
Ligaçāo centroide	(centroid)	Usa a distância entre as centroides.
Ligaçāo extensa	(ward)	usa a diferença da soma das distâncias dos pontos a suas respectivas centroides, com a distância de todos os pontos ao que seria a nova centroide caso ocorra a união dos grupos.

As vantagens dos algoritmos hierárquicos proveem de sua eficiência em lidar com ruídos e trabalhar com rapidez, já sua maior dificuldade é trabalhar com a matriz de similaridade de informações, que pode ter diferentes resultados dado a métrica de distância usada, a técnica de ligação ou até mesmo permutações nos dados da Tabela (DONI; OLIVEIRA, 2004).

### 3.4.3 MEAN SHIFT

Como último modelo que utiliza apenas da lógica clássica dos conjuntos apresenta-se o *Mean Shift*, um algoritmo que otimiza a metodologia aplicada pelo k-médias. Esse é considerado um modelo de similaridade baseado em densidade gráfica, e possui uma aplicação que independe de uma pré-seleção do número de divisões, sendo essa uma das principais vantagens do método (Comaniciu; Meer, 2002).

A técnica é baseada na movimentações de centroides, usando os mesmos critérios do K-médias. Todavia, nessa versão todos os valores dentro da base serão seus próprios centroides de um cluster, aonde são deslocados baseando-se na vizinhança mais próxima, até que ao final exista a convergência das centroides presentes nos conjuntos mais densos, o que forma a resposta final. A convergência para mesma posição dos valores pode ser comparada analogamente ao fato do k-médias conseguir gerar as mesmas partições para diversas inicializações diferentes (JAIN, 2010) (DRINEAS et al., 2004).

O maior problema da utilização dessa metodologia é a dificuldade de trabalhar com muitos ruídos e a necessidade da definição de um raio de vizinhança, um valor que define a área de visão dos centroides para encontrar vizinhos. Existem técnicas para escolha automática desse valor que são bem úteis na prática, como a explicada na Seção 4.2.3. O que diferencia na preferência pela escolha automática ou manual é principalmente, o problema a qual está aplicado e os recursos utilizados, geralmente a escolha de um número pequeno já é suficiente para bons resultados, variando pela disposição dos valores da base proposta. No caso da escolha automática o programa está menos suscetível a erros, mesmo que ainda possam acontecer, o problema é que essa é uma função que geralmente demanda grande aumento nos custos computacionais, muitas vezes não sendo viável ao problema (BUITINCK et al., 2013).

A dificuldade para tratar de ruídos geralmente é solucionada por uma abordagem de pesos na aplicação do algoritmo. Muitos modelos que utilizam do *Mean Shift*, usam variações menores e maiores do raio de vizinhança escolhido, aplicando pesos aos resultados das convergências que utilizam as vizinhanças de cada tamanho (Comaniciu; Meer, 1999), raios menores têm pesos maiores e vice versa. Esse método pode gerar um melhor desempenho nesse ambiente, mas também traz um grande aumento na complexidade.

### 3.4.4 C-MÉDIAS

Assim como o K-médias é um dos principais algoritmos da versão clássica, o C-médias é fundamental na base *fuzzy*, sendo uma variação da técnica *crisp*, ou seja, é a versão *fuzzy* do K-médias. Esse modelo agora implementa a matriz de pertinência em seus cálculos (Pal; Bezdek, 1995). A lógica funciona com a adição de “pesos” a todos os valores do banco quando comparados a cada um dos clusters, fazendo com que o cálculo do deslocamento das centroides (Equação 3.4) sofra mais ou menos influência das variações entre os valores do banco.

Os pesos representam as pertinências de um ponto para um cluster, sendo calculados pela Equação 3.2. Esse fator é também o que define os algoritmos como *fuzzy*, já que agora nenhum ponto pertence completamente a um único grupo, pelo menos durante a fase de processamento das partições.

O erro mínimo buscado também sofre algumas alterações para incluir esses novos valores, sendo atualizado pela Equação 3.3.

$$u_{ij} = \left[ \sum_{k=1}^c \left( \frac{d_{ij}^2}{d_{ik}^2} \right)^{\frac{1}{m-1}} \right]^{-1} \quad (3.2)$$

$$J_m(U, v) = \sum_{k=1}^n \sum_{i=1}^c (\mu_{ik})^m (d_{ik})^2 \quad (3.3)$$

$$v_{ij} = \frac{\sum_{k=1}^n \mu_{ik} * x_{ik}}{\sum_{k=1}^n \mu_{ik}} \quad (3.4)$$

O parâmetro  $m$ , observado nas Equações 3.3 e 3.2 que incluem o *fuzzy*, é denominado como um fuzzificador, onde maiores valores acarretam em uma maior influência das pertinências nas equações, enquanto o contrário torna as funções mais parecidas com o modelo clássico. O valor de  $m$  igual a zero seria o mesmo que retirar os pesos das pertinências nas equações, já que todos os atributos resultariam em 1.

As iterações do algoritmo em sua execução funcionam da mesma maneira que o K-médias, inicializando randomicamente  $C$  centroides e movimentando-os em função de diminuir o erro mínimo proposto. A maior diferença é que dada as novas equações existe uma influência maior ou menor de alguns pontos na movimentação, além de ao final ser possível obter a matriz de pertinência que é útil para determinados fins mais específicos.

### 3.4.5 GUSTAFSON KESSEL

O uso de uma arquitetura *fuzzy* no ramo do agrupamento já vem sendo estudado e aplicado com diversos algoritmos. Os mais conhecidos são, o já citado C-médias, e junto a ele o modelo de Gustafson Kessel (GK), que nada mais é que uma otimização desse primeiro modificando o cálculo da similaridade (FIRMANSYAH; PRAMANA, 2018).

A técnica do C-médias têm uma dificuldade considerável para lidar com ruídos na base de dados (MALHOTRA; KAUR; ALAM, 2014), o que é algo comum na prática. O algoritmo GK propõe uma nova maneira de calcular a distância entre os pontos, que pode se ajustar de acordo com suas posições e com isso, se ajustar melhor a esses dados mais dispersos. A Equação utilizada é a técnica de Mahalanobis (representada em 3.5) que utiliza da matriz norma das distâncias euclidianas de um cluster  $i$ , para ser capaz de adaptar-se ao movimento desses grupos em cada iteração.

$$d_{ij} = (x_i - v_k)^T p_k set(F_k)^{1/p} F_k^{-1} (x_i - v_k). \quad (3.5)$$

Com a Equação 3.5 de mahalanobis temos  $F_K$  como a matriz de covariância que é calculada pela função (Equação 3.6). GK também apresenta uma nova função de erro a ser minimizada pelo algoritmo (Equação 3.7).

$$F_k = \frac{\sum_{i=1}^n \mu_{ik}^m (x_i - v_k) (x_i - v_k)^T}{\sum_{i=1}^n \mu_{ik}^m} \quad (3.6)$$

$$J_m(U, V) = \sum_{i=1}^n \sum_{k=1}^K \mu_{ik}^m (x_i - v_k)^T p_k \det(F_k)^{1/p} F_k^{-1} (x_i - v_k). \quad (3.7)$$

Este modelo tem um problema que pode complicar algumas aplicações práticas quando usado. Ele ocorre quando uma base de dados é muito pequena ou com informações perto de serem correlacionadas linearmente, onde por esse motivo a matriz de covariância usada na Equação de Mahalanobis não pode ser invertida, causando problemas matemáticos aos resultados que se tornam equivalentes (Babuka; van der Veen; Kaymak, 2002).

### 3.4.6 C-MÉDIAS POSSIBILÍSTICO

Outro algoritmo muito usado para solucionar o problema de ruídos do C-médias é o C-médias Possibilístico (PCM - Possibilistic C-Means). Esse é um método que trabalha com a possibilidade de um ponto pertencer a um grupo ao invés do seu grau de pertinência propriamente descrito. Em termos matemáticos a fórmula continua a mesma, entretanto a restrição de que a soma de seus valores para os clusters seja 1, é retirada (Equação 3.8 e 3.9).

$$\mu_{ki} = \frac{1}{1 + \left( \frac{\|x_i - v_k\|^2}{\eta_k} \right)^{\frac{1}{m-1}}} \quad (3.8)$$

$$\eta_i = \frac{\sum_{j=1}^N (\mu_{ij})^m (d_{ij})^2}{\sum_{j=1}^N (\mu_{ij})^m} \quad (3.9)$$

Por consequência a diferença entre os valores pode ser muito mais alta que o comum, diminuindo a influência de ruídos no cálculo do centroide. O lado negativo deste modelo, é que em alguns casos essa discrepância pode dificultar a análise de um valor apropriado de partições, ou até gerar uma convergência dos centroide.

Com a alteração do cálculo da pertinência, novamente tem-se uma adaptação no método que é usado para calcular o erro mínimo, agora dado pela Equação 3.10 e 3.9.

$$J_m(U, v) = \sum_{i=1}^C \sum_{j=1}^N (\mu_{ij})^m (d_{ij})^2 + \sum_{i=1}^C \eta_i \sum_{j=1}^N (1 - \mu_{ij})^m \quad (3.10)$$

## 3.5 VALIDANDO AGRUPAMENTOS

Como a tarefa de agrupamento é não-supervisionada, a avaliação dos clusters obtidos por cada modelo é uma tarefa essencial. Os diversos modelos tem diferentes características e os conjuntos de dados tem disposição distinta no espaço, o que torna o processo de avaliação necessário para que o modelo adequado seja selecionado.

são definidas na literatura diferentes métodos de avaliação. Estas métricas podem ser divididas em três categorias que representam a generalização dos objetivos de uma avaliação, são elas: externa, interna e relativa (CHARRAD et al., 2014) (KOUTROUMBAS; THEODORIDIS, 2008). É válido ressaltar que muitas das técnicas utilizadas em cada classe, podem também ser aplicadas a outra categoria quando empregadas de maneira correta.

### **3.5.1 VALIDAÇÃO INTERNA**

Este tipo de avaliação contabiliza as características internas do modelo de agrupamento trabalhado, determinando um conjunto de medidas quantitativas referentes a execução desse programa e sua configuração. As avaliações internas contabilizam dados como a corretude do algoritmo, seus grupos resultados e até tempo de execução.

Exemplos de métricas nessa categoria são os coeficiente de Dunn, Davis-Bouldin (Davies; Bouldin, 1979), Silhueta (ROUSSEEUW, 1987) e até mesmo a soma dos erros quadrático das partições geradas.

### **3.5.2 VALIDAÇÃO EXTERNA**

O método externo é usado apenas em bancos de dados que possam ser comparados a rótulos previamente definidos, assim como no ramo supervisionado do AM. Estas estatísticas são moldadas pela comparação probabilística dos pares de pontos e suas alocações em cada rótulo gerado, designando pares como falsos positivos ou falsos negativos, por exemplo.

Exemplos são o coeficiente de homogeneidade (ROSENBERG; HIRSCHBERG, 2007) e o índice de Folkes e Mallows (FOWLKES; MALLOWS, 1983).

Apesar de ainda ser uma técnica muito usada, algumas pesquisas mostram que essa não é uma forma correta para avaliar algoritmos de agrupamento, assim seu uso se justifica apenas como suporte a análises específicas. (FARBER et al., 2010).

### **3.5.3 VALIDAÇÃO RELATIVA**

A validação relativa é baseada na comparação entre diferentes métodos de agrupamento, determinando qual modelo é capaz de convergir um bom resultado com mais eficiência, geralmente sendo aplicados em situações previamente determinadas.

Geralmente não são determinadas métricas específicas dessa categoria, o que usualmente ocorre é a formação de uma metodologia de testes que usa as métricas das demais categorias, parametrizando-as para prover a comparação dos resultados e fornecer indicadores para uma comparação.

### **3.5.4 MÉTRICAS EMPREGADAS NESTA PESQUISA**

As métricas aplicadas na pesquisa são principalmente de validação interna e relativa, e são apresentadas nas Tabelas 2 e 3 a seguir.

**fuzzificador(m) = 2.0**

**nº máximo de interações = 1000**

**erro mínimo = 0.0001**

Tabela 2 – Métricas não-supervisionadas usadas para validação interna dos métodos crisp e posteriormente para comparação e validação relativa entre os algoritmos.

Métrica	descrição	objetivo
Silhouette_Score	Média da "silhueta" de cada dado. razões entre a variância intra-cluster e distância até a próxima partição vizinha. [scikit]	Max
Calinski_Harabasz_Score	razão entre a dispersão inter-cluster e a dispersão entre clusters. também é conhecido como Variance Ratio Criterion. [scikit]	Max
Davies_Bouldin_Score.	média do valor de similaridade de cada cluster comparado com o seu cluster mais próximo [scikit]	Min
Dunn_Score	razão entre o valor mínimo inter-cluster e o valor máximo intra-cluster. [jqmviegas]	Max

Tabela 3 – Métricas não-supervisionadas usadas para validação interna dos métodos *fuzzy*, fazendo uso da Tabela de pertinência. Os algoritmos foram fornecidos pelo repositório (GERMER, 2017)

Métrica	descrição	objetivo
Partition_Coefficient	média quadrática das pertinências de cada ponto em cada cluster	Max
Fuzzy_Hypervolume	somatório dos determinantes da matriz de covariância de cada cluster, define um grau de independência das partições e instabilidade do agrupamento	Min
Fukuyama_Sugeno_Score	descreve a compactação e a distância dos clusters, por um diferencial das pertinências com as distâncias dos pontos e as centroides.	Min
Xie_Beni_Score	coeficiente de otimização baseado na iteração compactação e a "força" entre os clusters.	Min

Observa-se a presença de métricas que usam o modelo de pertinências *fuzzy* e outros apenas dos rótulos respostas. Isto se deve aos modelos selecionados para estudo que apresentam as duas características. Todavia as técnicas com uso da pertinência são apenas usadas na validação interna, enquanto as demais métricas são também usadas de maneira relativa, já que suportam respostas de ambas teorias de conjuntos.

As métricas que usam rótulos verdade são todas apresentadas na Tabela 4.

Tabela 4 – Métricas com base supervisionada, usam rótulos para validar. Os algoritmos foram fornecidos por Scikit-Learn(BUITINCK et al., 2013), valores mais altos significam melhor compatibilidade e 1 é o valor máximo.

Métrica	descrição
Adjusted_Rand_Score	total de pares em que ambas as partições concordam em classificar sobre total de pares.
Completeness_Score	pontos que pertenciam a um mesmo grupo e continuam pertencendo a uma mesma partição.
Homogeneity_Score	baseado no número de rótulos diferentes deferidos a um mesmo cluster.
Folkes_Mallows_Score	média geométrica entre a precisão e memória de um cluster.
Normalized_Mutual_Info_Score	técnica estatística para avaliar a co-ocorrência de dados aplicado a dois clusters.
V_score	média harmônica entre a completude e a homogeneidade.

### 3.6 ENSEMBLE DE AGRUPAMENTOS

Os modelos de *ensemble* de agrupamentos buscam combinar conjuntos de agrupamentos obtidos por outras técnicas. A combinação pode ser de resultados gerados a partir de uma técnica, ou provenientes de técnicas distintas. Trata-se de modelos múltiplos descritivos que buscam superar limitações das técnicas de agrupamento, gerando resultados mais robustos e de melhor qualidade (FACELI, 2011).

Nos problemas de classificação e regressão os ensembles são comumente empregados. Os modelos de agrupamento são não supervisionados, o que faz com que os ensembles de agrupamentos sejam modelos adaptados, uma vez que um agrupamento não tem o rótulo ou a resposta verdadeira para a construção do modelo.

Existem diversas técnicas de *ensemble* que realizam o cálculo da função consenso de maneiras diferente, sendo esta a função que assimila os resultados dos algoritmos base separados em uma única resposta. o foco neste trabalho é com o método da votação, um dos mais bem empregados e com simples implementação. Deve ser ressaltado que para aplicar o *ensemble* de maneira correta é essencial que exista uma variação entre os resultados base, sendo esse o principal requisito. Duas categorias que dividem as metodologias de *ensemble* são os homogêneos, que utilizam apenas diferentes configurações de um mesmo agrupamento base, e heterogêneo com diferentes aplicações.

A votação parte da re-execução dos algoritmos base em diferentes configurações, para cada processamento são geradas diferentes rotulações que contaram como votos para o resultado final, assim quando diversos algoritmos e/ou configurações indicam um ponto ao rótulo do grupo x, esse será o grupo formado no resultado final. O maior problema dessa técnica é tratar da equivalência entre a nomenclatura das partições, pois em muitos casos dois agrupamentos

podem ser idênticos, mas terem considerado as mesmas informações como sendo do grupo 1 no primeiro caso e 0 para o segundo caso, as divisões não deixam de ser iguais, porem podem acarretar problemas para essa técnica.

## 4 DESCRIÇÃO DAS ATIVIDADES DESENVOLVIDAS

Neste Capítulo é detalhada toda a metodologia envolvida no processo de avaliação de modelos de agrupamento clássicos e *fuzzy*, buscando atingir os objetivos propostos. Para tanto são definidos 3 conjuntos de experimentos empregando 11 bases de dados distintas.

A metodologia empregada utiliza a linguagem Python 3.7.4, a partir de implementações retiradas e adaptadas de repositórios e plataformas *online* voltadas a pesquisa, utilizando como ferramenta o Visual Studio Code para a programação.

O plano traçado considera as diferentes características dos modelos de agrupamento e das bases de dados, visando observar as vantagens e desvantagens dos algoritmos selecionados para estudo.

### 4.1 ALGORITMOS SELECIONADOS PARA ESTUDO

Os modelos de agrupamento clássicos selecionados para estudo, disponíveis no Scikit-learn, foram: Aglomerativo, K-médias e *Mean Shift*.

Para as técnicas que aplicam a lógica *fuzzy*, encontrar repositórios otimizados online é mais difícil, já que são aplicações mais recentes quando comparadas à versão clássica, portanto, foram selecionados de plataformas distintas. O Scikit-learn disponibiliza uma plataforma, ainda em desenvolvimento, para trabalhar com algoritmos *fuzzy*, onde está disponível o algoritmo C-médias 3.4.4 empregado. Os outros dois algoritmos *fuzzy* selecionados foram a versão possibilística 3.4.6 do C-médias e o algoritmo de Gustafson Kessel (GK) 3.4.5, dois modelos considerados otimizações do C-médias. As implementações utilizadas são apresentadas por (EL-DIN; ALJABASINI, 2018) (Gustafson Kessel) e por (SKINNER, 2018) (a versão possibilística).

Na Tabela 5 são apresentados os nomes das funções que representam os modelos empregados e seus principais parâmetros.

Tabela 5 – Resumo dos algoritmos utilizados e seus parâmetros

<i>Algoritmo</i>	<i>desenvolvedor</i>	<i>parâmetros</i>
AgglomerativeClustering	Scikit-learn	(Clusters, affinity, linkage)
KMeans	Scikit-learn	(n_clusters, init, max_iter, tol, random_state)
MeanShift	Scikit-learn	(bandwidth, cluster_all, max_iter)
cmeans	Scikit-Fuzzy	(data, c, m, error, maxiter)
Gustafson Kessel	(EL-DIN; ALJA-BASINI, 2018)	(n_clusters, max_iter, m, error)
Possibilístico	(SKINNER, 2018)	data, c, m, e, max_iterations)

Complementando a pesquisa, foi realizada uma última etapa onde os modelos gerados foram combinados. O *ensemble* é um algoritmo criado por (YANG, 2016), onde são feitos alguns ajustes para realizar a soma das pertinências para todos os agrupamentos *fuzzy*. Para obtenção dos rótulos das partições de maneira equivalente, é empregada uma matriz de custos, descrita na Seção 4.2.5.

Os demais algoritmos que geram os resultados para análise, normalizações, ajustes e a criação das bases sintéticas, são todos de autoria do autor desta pesquisa. Os gráficos foram gerados empregando as bibliotecas do Numpy, Matplot e Seaborn, principalmente.

## 4.2 METODOLOGIA DE TESTES E AVALIAÇÕES

Os testes foram divididos em três experimentos estruturados para explorar diferentes fatores em cada algoritmo, buscando obter informações para definir as vantagens e desvantagens em cada técnica de agrupamento.

O primeiro experimento consiste em um esquema de validação, onde foram gerados conjuntos de dados sintéticos de duas dimensões, buscando demonstrar as características dos algoritmos estudados, bem como possibilitar a visualização dos dados no espaço. Para o segundo experimento foram selecionadas bases de dados clássicas da literatura, possibilitando o estudo do desempenho dos modelos. E por fim, o terceiro experimento é realizado a partir de bases de dados maiores selecionadas de forma aleatória na internet, buscando avaliar como os algoritmos obtém resultados em um teste prático, ou mais próximo de um problema real.

Na Figura 4 é possível observar o esquema da metodologia de forma simplificada.

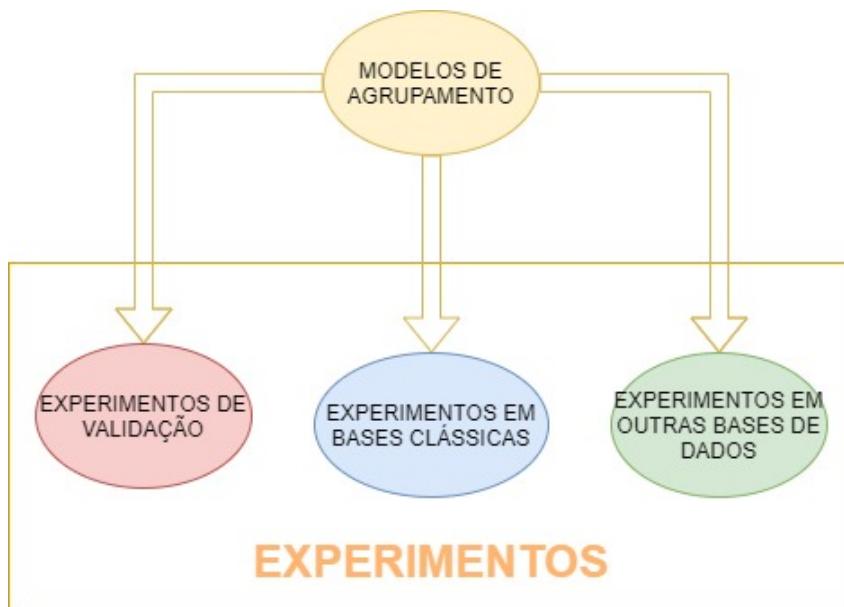


Figura 4 – Esquema da divisão do banco de dados

Para cada experimento, os modelos foram configurados buscando os parâmetros mais adequados. A parametrização dos modelos é detalhada na Seção 4.2.1.

#### 4.2.1 PARAMETRIZAÇÃO DOS MODELOS

Para definir os melhores parâmetros empregados nos modelos são empregadas algumas técnicas automáticas. A parametrização é o passo inicial para o agrupamento, sendo uma das três categorias na avaliação de um cluster (CHARRAD et al., 2014) (KOUTROUMBAS; THEODORIDIS, 2008). São definidos os principais argumentos a serem usados pelos modelos para gerar os grupos, alguns exemplos são: número de partições, número máximo de iterações ou, em modelos mais específicos como o Mean Shift, raio de vizinhança.

Essa parametrização normalmente pode ser definida por um padrão definido no próprio problema a ser solucionado, ou usando métricas de validação interna 3.5.1 para verificar valores mais otimizados, ou seja, que retornem a maior pontuação com as métricas. Na Tabela 2 são definidas as métricas utilizadas como esse propósito.

Para todos os experimentos é empregada a avaliação interna, com o propósito de verificar configurações que obtenham boas respostas para cada algoritmo. Foram avaliados os parâmetros: número de partições, tipo de ligação aglomerativa e raio de vizinhança para o *Mean Shift*. Para os demais parâmetros foram mantidos fixos os valores padrão.

A validação interna dos parâmetros iniciais é feita através da observação do comportamento gráfico dos coeficientes (Tabelas 2 e 3), através de variações, pontos de máximo ou pontos de mínimo, sempre de acordo com o objetivo de cada métrica. Como não existe um modelo que sempre apresente a resposta correta, é possível que os valores destacados durante os testes não apresentem o melhor desempenho possível, porém ao usar uma heurística de avaliação existe uma maior garantia de que os resultados representem boas respostas para os algoritmos.

Considerando todas as bases de dados e cada algoritmo selecionado são gerados os gráficos de respostas para as medidas avaliadas, estabelecendo configurações que vão de 2 a 10 clusters, e posteriormente de 2 a 31 clusters. Intervalos de busca menores são mantidos para os testes à medida que diminui o tamanho do banco de dados utilizado, e valores mais altos não apresentam boas respostas para retirada de informações.

A partir dos gráficos foram selecionados os melhores valores, buscando alternativas que se demonstrem mais viáveis para cada modelo de agrupamento testado. A análise feita sobre a viabilidade da respostas dos gráficos é demonstrada no experimento de validação na Seção 5.1.1.

A análise do raio de vizinhança e do método de ligação foram realizadas por técnicas diferentes. A área de busca do *Mean Shift* foi inicialmente estabelecida como 1,2, em caso de falhas no agrupamento, como a criação de apenas um grupo, o valor é reduzido em décimos até se adequar algum resultado. Prosseguindo a avaliação e objetivando apresentar toda a capacidade do *Mean Shift* de não necessitar da escolha de parâmetros iniciais, a partir dos experimentos em bases clássicas, o raio de vizinhança passa a ser calculado de forma automática por uma função de média disponibilizada pelo scikit-learn, explicada na Seção 5.2.2.

Para a ligação aglomerativa são utilizados dendrogramas por ligação extensa (*ward*), onde

são verificados valores iniciais de partições através de um corte horizontal no gráfico, usados apenas para comparação entre diferentes ligações. No final, a medida que demonstra melhores respostas é a escolhida. No decorrer dos próximos testes, se observado que uma ligação possui um melhor desempenho, essa alteração também é registrada.

Encontrados os melhores valores, os modelos são reconfigurados e uma nova execução é feita com o objetivo de gerar os gráficos e rótulos finais, gerando o agrupamento para cada algoritmo, indicando também o tempo de execução e outros objetivos definidos para cada experimento do projeto 4.2. Assim é permitido a comparação relativa das duas lógicas e seus resultados finais.

#### 4.2.2 EXPERIMENTO DE VALIDAÇÃO

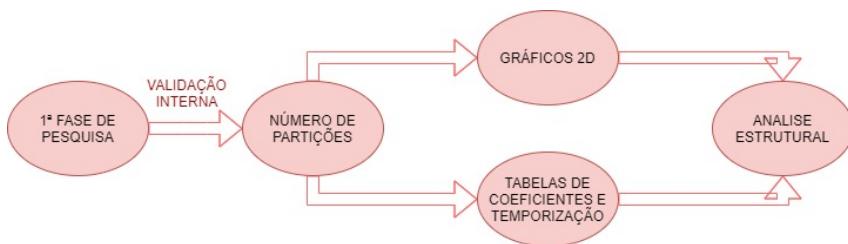


Figura 5 – Fluxo da primeira fase de testes

O primeiro conjunto de experimentos (Figura 5) emprega uma base de dados sintética (criada pelo próprio autor) que consiste nos seguintes conjuntos: “Tríade”, “Grupo de Anéis”, “Interlaço”, “Anel Interno”, “Ruídos” e “Luas”. As bases de dados que compõem esta etapa seguem padrões de figuras comuns para o agrupamento de dados, como intersecção de semi-círculos e ruídos espalhados, proximidade entre figuras internas e externas, regiões de maior e menor densidade, entre outros, como empregado nos estudos de (DAVE, 1996). Os gráficos que representam as bases de dados empregadas podem ser visualizados na Figura 6, apresentando o número de dados(pt) e dimensões(D) em cada uma..

Essa etapa consiste em construir conjuntos de dados de duas dimensões, onde é possível identificar agrupamentos a partir de uma análise visual. Para tanto, as configurações iniciais de cada algoritmo de agrupamento são definidas para cada base de dados, e o resultado é avaliado graficamente, onde as partições obtidas são apresentadas em cores distintas. Assim é possível apontar as diferenças entre as partições criadas por cada algoritmo.

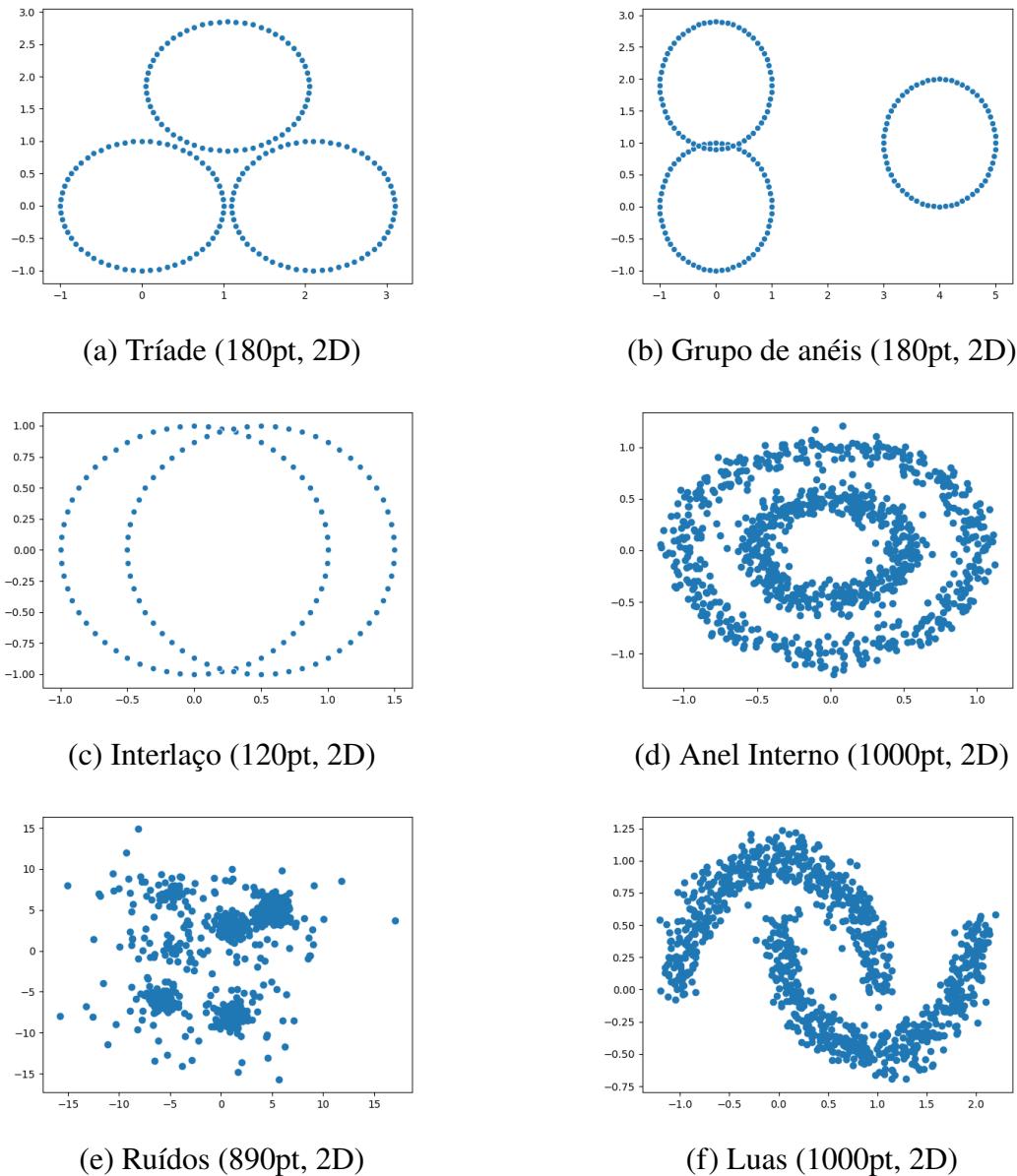


Figura 6 – Figuras ilustrativas de cada base de dados sintética.

#### 4.2.3 EXPERIMENTOS EM BASES CLÁSSICAS

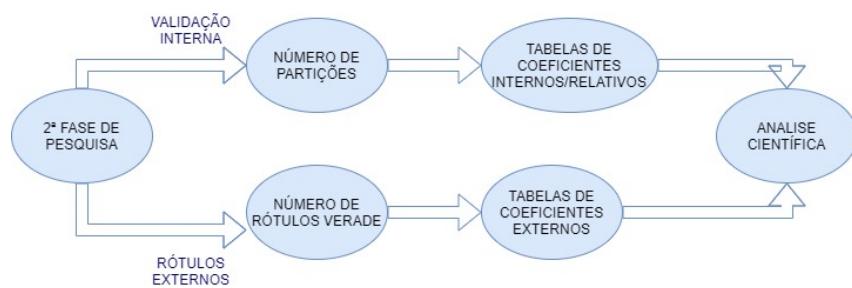


Figura 7 – Fluxo da segunda fase de testes

O segundo experimento (Figura 7), além de realizar o esquema comum de encontrar o número de clusters e avaliar os coeficientes relativos, utiliza bases de dados muito presentes na comunidade científica e que apresentam rótulos verdade para a realização da validação externa. Para as bases de dados selecionadas a representação gráfica não é interessante, por se tratar de dados que apresentam mais de 3 dimensões.

As base de dados utilizadas nesta segunda etapa do projeto são descritas nas Seções que se seguem.

#### 4.2.3.1 Iris

A base de dados Íris é datada de 1936 (FISHER, 1936), e foi primeiramente pensado para as tarefas supervisionadas, todavia ainda é comum o uso deste conjunto no campo de agrupamento. Ele é baseado em 50 plantas selecionadas de cada um dos três tipos de flores de íris: setosa, versicolor e virgínica, caracterizadas pela largura e comprimento de suas pétalas e sépalas (DEVASENA et al., 2011).

#### 4.2.3.2 Wine

O Wine é um conjunto de dados muito utilizado pelo seu número mais alto de dimensões, são 13 características para descrever 3 tipos de vinhos, dentre 178 amostras (DUA; GRAFF, 2017).

#### 4.2.3.3 Boston

Por fim, o conjunto Boston foi originalmente coletado por "U.S Census Service concerning housing", e é largamente utilizado, atualmente, para a tarefa de regressão. Pode ser resumido como um modelo de preço hedônicos das habitações na região de Boston, possui 506 amostras descritas em 13 dimensões (HARRISON; RUBINFELD, 1978). Esse é o único modelo dessa fase de testes que não possui rótulos verdade.

#### 4.2.4 EXPERIMENTOS EM OUTRAS BASES DE DADOS

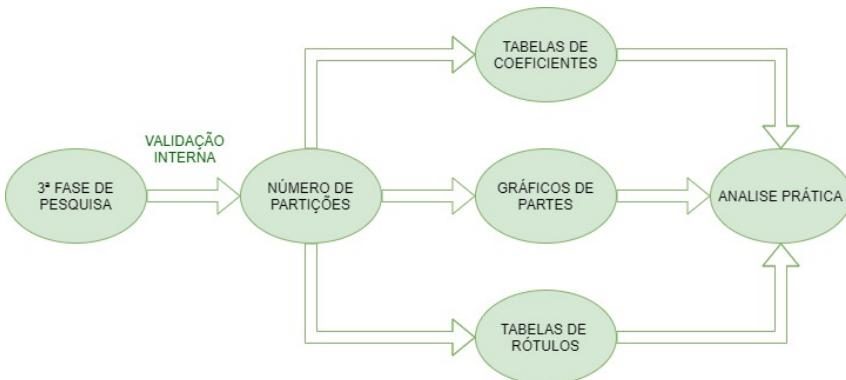


Figura 8 – Fluxo da terceira fase de testes

No terceiro experimento do projeto de pesquisa (Figura 8), todos os passos já descritos foram aplicados a bases de dados selecionadas de forma aleatória na internet. Os bancos trabalhados aqui foram selecionados a partir da plataforma Kaggle ([kaggle.com](https://kaggle.com)), um repositório de dados conhecido por sua inúmera reunião de informações, envolvendo a tarefas de aprendizado de máquina.

Com o foco mais prático é feito também uma avaliação voltada a sua aplicação real, objetivando após o agrupamento final, utilizar-se dos rótulos e de comparações gráficas bidimensionais para retirada de novas informações reais sobre aquele conjunto. Assim a comparação e análise final dessa etapa consiste em uma visualização de que conhecimentos cada algoritmo é capaz de trazer, proporcionalmente comparando também com os coeficientes relativos gerados.

##### 4.2.4.1 Seguro Médico

O primeiro banco selecionado, Seguro Médico <sup>1</sup>, apresenta as informações base perguntadas a qualquer paciente que necessita realizar um plano médico, além do valor do plano proposto. Essa base é bem intuitiva para a construção do agrupamentos, já que realmente existe uma classificação feita pelos centros médicos antes de catalogarem valores. Assim, uma análise nesse tipo de recurso pode retirar informações sobre como essa classificação é feita ou sobre um grupo de pacientes, por exemplo.

Este banco de dados apresenta 1338 registros e 7 características. Para melhor trabalhar com os dados fornecidos foram realizados alguns ajustes, como a transformação de valores textuais em índices. Um exemplo seria a característica que descreve o sexo, onde 0 representa os homens e 1 as mulheres.

<sup>1</sup> Acessado em 2020: <https://www.kaggle.com/mirichoi0218/insurance>

#### 4.2.4.2 Consumo Online

O banco de dados Consumo Online<sup>2</sup> reúne diversas informações sobre clientes e seus respectivos hábitos de compra, descrevendo características como preço dos produtos, quantidade comprada e qual região da compra. Apesar de mais simples com apenas essas informações é possível verificar aspectos como o poder de compra de determinados locais e sua industrialização, fatores onde o emprego do agrupamento é comum, ou adequado.

Para essa base de dados, a mesma conversão numérica da base anterior é feita para algumas das características, enquanto outras acabaram sendo excluídas, como identificadores que não possuem nenhum valor significativo, a não ser na organização da base de dados. Também foi necessário diminuir a quantidade de dados utilizados à medida que alguns algoritmos como o aglomerativo não suportavam. Para realizar esse processo de corte, foram selecionadas até 100 tuplas de cada país utilizado no banco.

Como resultado final as características usadas foram: Preço unitário, quantidade comprada, número cliente e país, gerando um total de 3154 tuplas.

#### 4.2.5 ENSEMBLE

Buscando investigar o desempenho da combinação dos modelos gerados, foi proposta a re-execução do conjunto de experimentos empregando *ensemble* por votação.

O modelo múltiplo descritivo gerado é baseado nas diferentes respostas de configurações aleatórias para o agrupamento de uma base de dados. Sempre que um algoritmo indicar que determinada informação é pertencente a um grupo, essa alocação conta como um voto para a alocação final que representa as partições mais votadas. No caso *fuzzy*, as respostas são inicialmente dadas por pertinências impedindo que o esquema de votação tire seu maior proveito quando tratada neste esquema, realizando as somas das pertinências. Assim, um valor assimilado com pertinência 0,5 a um grupo terá esse valor somado a matriz final. Para que isso ocorra sem problemas também é feita a normalização das respostas possibilistas para valores entre 0 e 1, fazendo a divisão da soma total de suas respostas por cada valor separado.

As diferentes partições dos algoritmos são geradas com a modificação randômica de seus parâmetros iniciais: tipo de ligação aglomerativa, cálculo de afinidade, valor inicial de centroides, raio de vizinhança (entre 0 a 100, até duas casas decimais) e parâmetro fuzzificador (entre 0 e 10, com uma casa decimal).

Como os modelos neste trabalho são de agrupamento, existe o problema da diversificação dos rótulos gerados a cada execução. A partir de duas execuções distintas ainda é possível que o resultado seja similar, porém os nomes dos grupos pode ser diferente. Isso porque o algoritmo não identifica rótulo, apenas retorna instâncias pertencentes a um conjunto de grupos. Para solucionar esse problema é utilizada uma técnica de equivalência com base no problema de atribuição de

---

<sup>2</sup> Acessado em 2020: <https://www.kaggle.com/carrie1/ecommerce-data>

trabalho a operários minimizando o custo, também conhecido como “Hungarian assingment problem” (KUHN, 1955). Os trabalhos e os custos são gerados por uma matriz de equivalência entre as partições a serem re-rotuladas, se o ponto x está na partição 1 do primeiro agrupamento, e na 3 do segundo, a posição [1][3] na matriz soma mais um ponto, e quanto maior o número de pontos menor o custo. Com a matriz gerada é possível utilizar uma biblioteca própria do Python que otimiza esse problema, o Munkres. A resposta para o problema é um par de equivalência de rótulos, onde por exemplo ’1 2’, é uma resposta que significa que o primeiro rótulo da execução x equivale ao segundo do agrupamento y.

## 5 RESULTADOS OBTIDOS E ANÁLISE

Neste Capítulo são apresentados os resultados obtidos a partir dos experimentos realizados seguindo a metodologia detalhada no Capítulo 4.

### 5.1 EXPERIMENTO DE VALIDAÇÃO

A primeira etapa é a descoberta dos parâmetros iniciais adequados, fazendo o uso da validação interna, feita separadamente para cada algoritmo, partindo do fato que cada técnica possui vantagens e tendências únicas ao agrupar as informações. Inicialmente, o raio de vizinhança do *Mean Shift* é definido de forma manual, partindo de um valor padrão igual a 1,2 e fazendo leves alterações quando necessário e possível.

#### 5.1.1 ANÁLISE DO NÚMERO DE PARTIÇÕES

Os gráficos das Figuras 9 a 18 representam as medidas obtidas a partir da análise interna: coeficiente de Silhueta, coeficiente de Calinski, coeficiente de Davies Bouldin e coeficiente de Dunn, para a base de dados sintética tríade. Nele observamos o melhor resultado marcado por uma linha vertical laranja, valores de máximo ou de mínimo de acordo com o exigido pelo coeficiente.

Para cada algoritmo, em cada uma das bases sintéticas, são considerados 2 intervalos distintos de número de partições, um variando de 2 a 10, foco principal dos valores buscados, e outro de 2 a 31 usado para observar melhor o comportamento dos coeficientes. Desta forma, considerando as 4 medidas de análise interna, são gerados 8 gráficos.

Os banco de dados utilizados nesses experimentos trabalham com um valor reduzido de informações, seja pela quantidade de dados ou o número de atributos em cada tupla. Assim, dado também o ambiente com foco experimental a qual são aplicadas, durante a escolha das configurações que melhor se adaptem aos coeficientes, a preferência é dada para números menores de partições. Logo, em caso de empate, a escolha é pelo menor.

Outro ponto observado, principalmente nos coeficientes *fuzzy*, é a tendência de alguns gráficos formarem um crescimento ou decrescimento constante, não permitindo apontar para nenhum valor confiável, consequentemente tornando a análise mais difícil. Em alguns casos com esse tipo de formação, pode ser aplicada a mesma técnica do joelho de Elbow, identificando fronteiras em que essa taxa de crescimento/decrescimento começa a se tornar muito pequena, podendo assim também identificar uma boa resposta.

Com os gráficos formados, a análise se resume na busca do número de grupos (ou clusters) que apresente os melhores resultados para a maioria dos coeficientes. A partir dos resultados dos modelos crisp (ou clássicos) é possível exemplificar como esse processo é feito, por exemplo, a partir dos gráficos das Figuras 9 e 10. Neste caso a análise foi mais simples, uma vez que ambos os

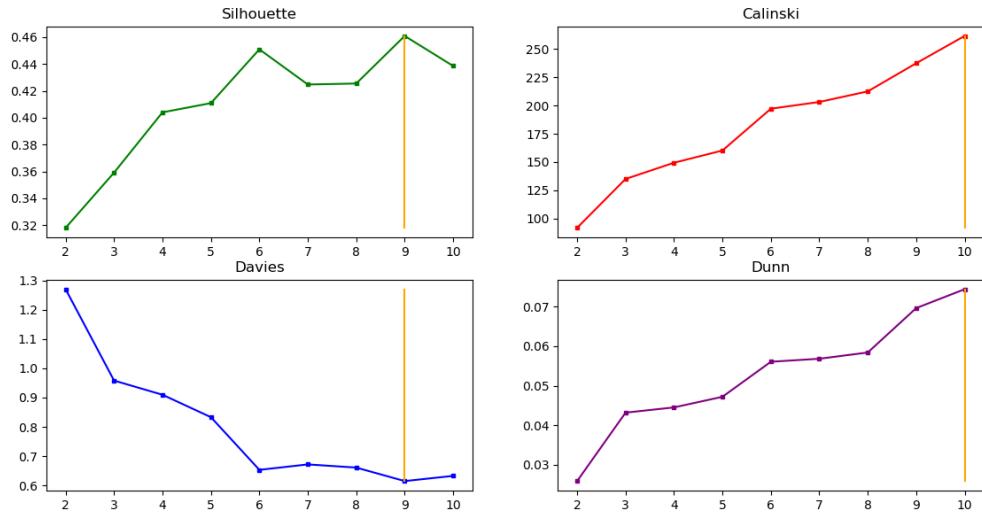


Figura 9 – Coeficientes k-médias - Tríade (de 2 a 10)

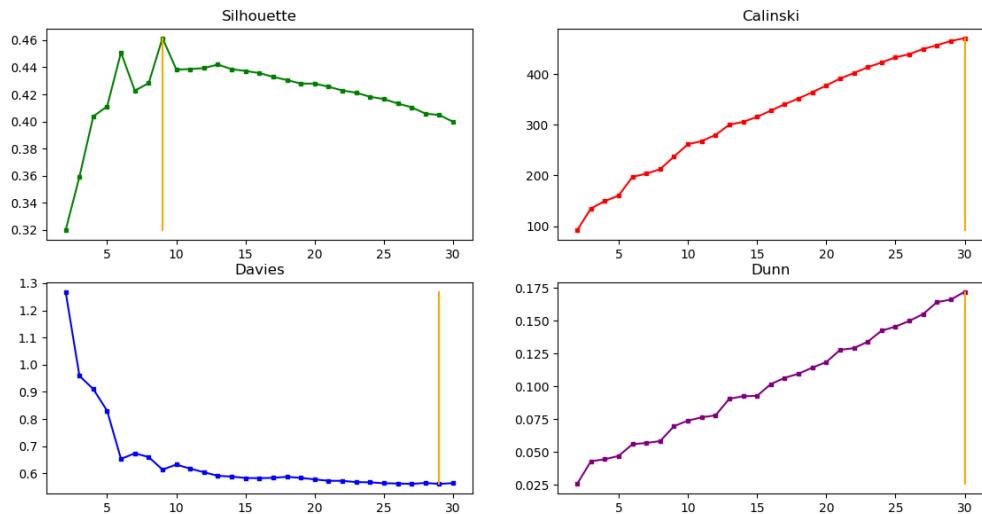


Figura 10 – Coeficientes k-médias - Tríade (de 2 a 31)

coeficientes de Silhueta e de Davies-Bouldin indicavam um mesmo valor, gerando uma maioria na preferência pelo número nove, enquanto os demais coeficientes resultaram apenas num padrão de crescimento constante. Já em respostas como a do método Aglomerativo, observadas nas Figuras 11 e 12, demonstram a tomada de decisão pelo menor valor, já que todos os coeficientes indicaram valores distintos.

Observando os resultados para os modelos *fuzzy* é preciso ter algumas outras considerações importantes:

- A primeira é uma tendência já comprovada e apontada em outros estudos, como (LI, 2011), do Coeficiente de Partição indicar o número dois como melhor número de partições e ter

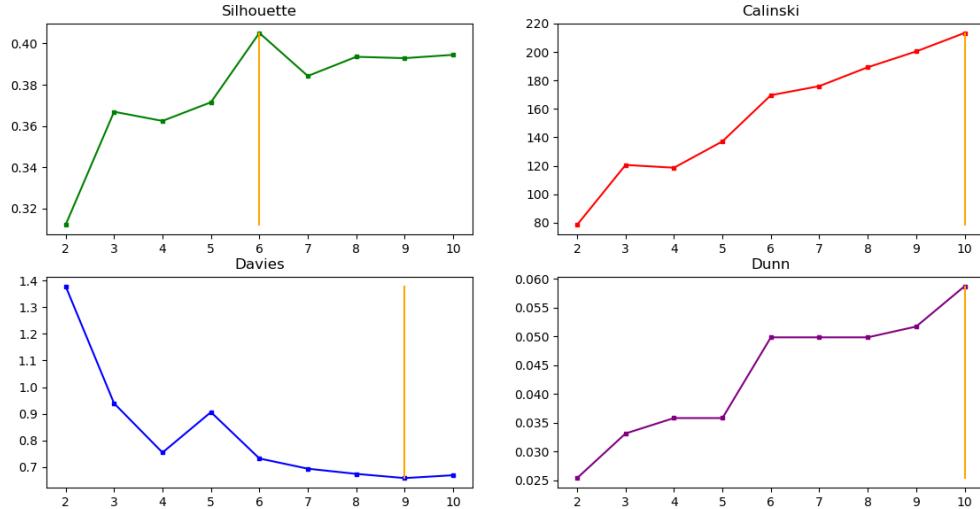


Figura 11 – Coeficientes Aglomerativo - Tríade (de 2 a 10)

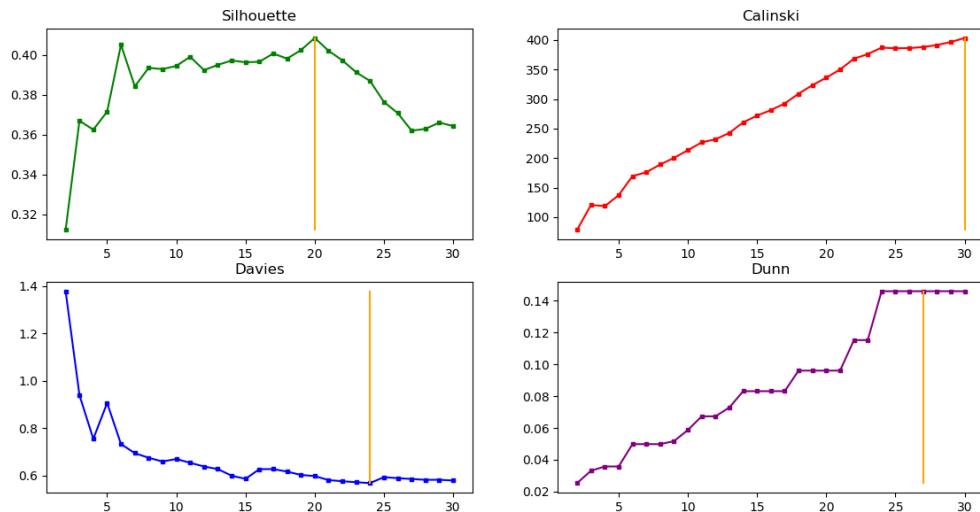


Figura 12 – Coeficientes Aglomerativo - Tríade (de 2 a 31)

um padrão constante de decrescimento. Seu uso ainda é relevante e notado ao usar técnicas como a de Elbow para identificar mudança de padrões no decrescimento ou verificando casos que fujam totalmente do padrão, sendo estes também indicadores de bons valores.

- Outra consideração importante é a discrepância entre as pertinências geradas por alguns agrupamentos, principalmente para modelos ilimitados quanto a pertinência, como o método Possibilístico, Onde podem surgir diferenças tão altas entre os valores que os coeficientes formam retas para uma visão mais ampla dos gráficos, o que prejudica a análise.

Diversas técnicas para avaliação de modelos *fuzzy* já foram propostas pela comunidade,

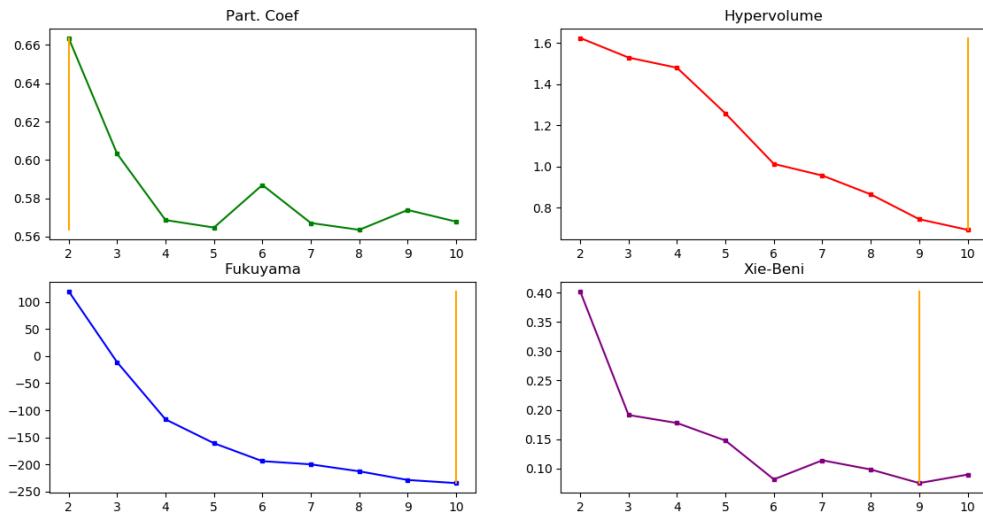


Figura 13 – Coeficientes c-médias - Tríade (de 2 a 10)

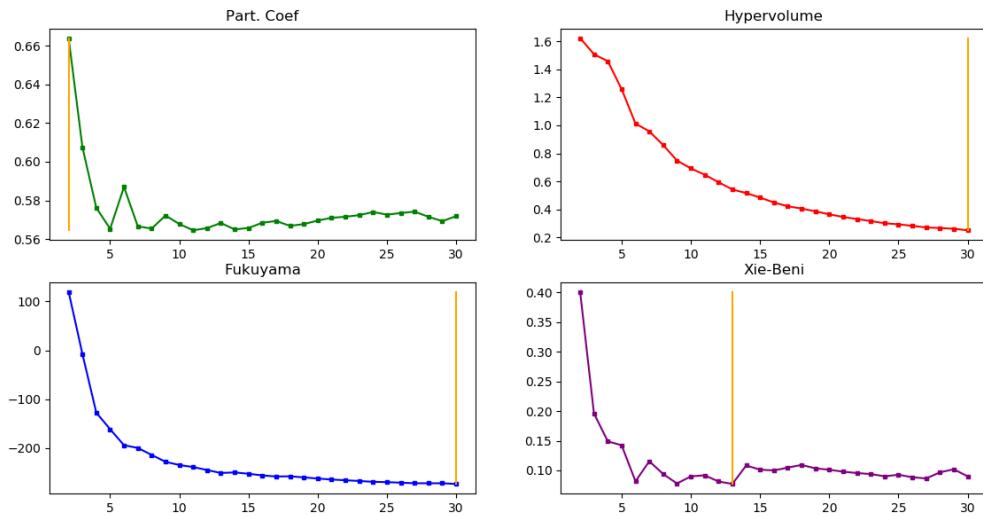


Figura 14 – Coeficientes c-médias - Tríade (de 2 a 31)

porém são técnicas mais novas, que veem ganhando espaço mais recentemente. Neste trabalho destacam-se os coeficientes de Xie Beni e Fukuyama, empregados para escolha do número de clusters durante os testes.

Com as respostas dos modelos *fuzzy*, também é possível observar casos onde a escolha do número de clusters foi mais direta. Para o modelo de Gustafson Kessel, observa-se nas Figuras 15 e 16, duas fortes indicações do valor dois (2) como melhor número de clusters, sendo indicado por Xie Beni e pelo Coeficiente de Partição, enquanto os outros dois coeficientes decrescem constantemente.

Em alguns resultados para os modelos *fuzzy*, como os gráficos dos coeficientes para o

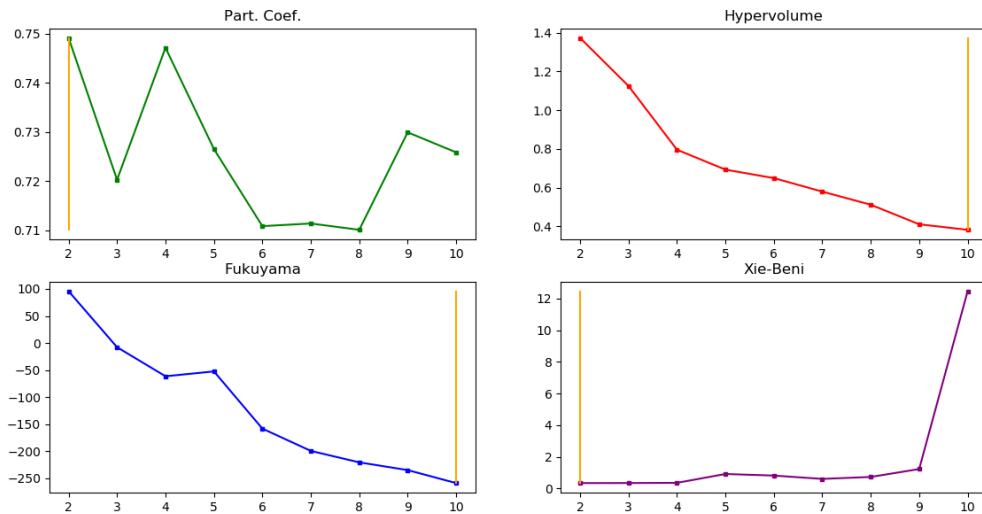


Figura 15 – Coeficientes Gustafson - Tríade (de 2 a 10)

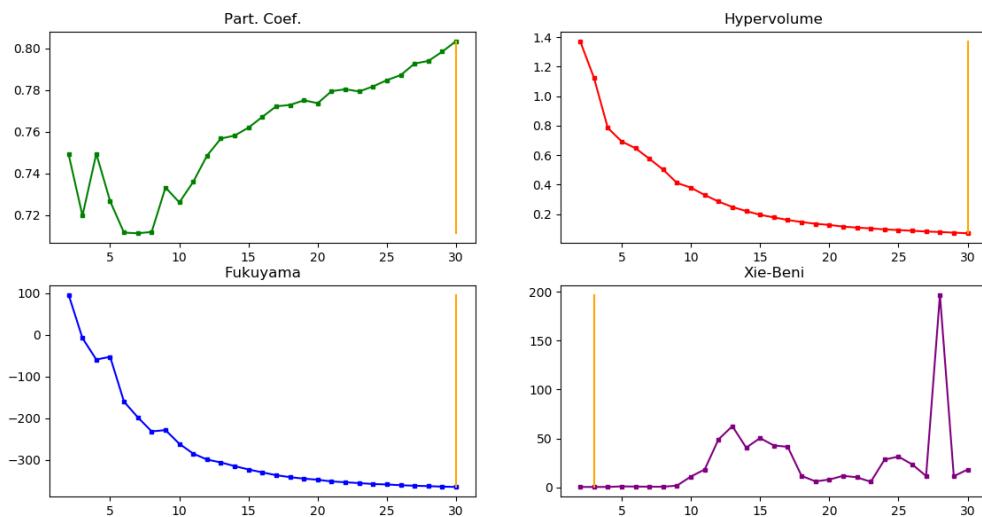


Figura 16 – Coeficientes Gustafson - Tríade (de 2 a 31)

modelo possibilístico (Figuras 17 e 18), apenas um dos coeficientes fez alguma indicação, nesse caso três (3) clusters indicado por Xie Beni. Nos resultados para o modelo do C-médias o mesmo pode ser observado, onde apenas Xie Beni fez uma indicação (Figura 13), porém nesse caso em específico, é possível perceber que o valor 6 possui respostas próximas ao 9, só que para 6 esse valor também é um ponto de pico no gráfico do Coeficiente de Partição, e se mostrou uma escolha mais balanceada entre os coeficientes.

A partir dos critérios de análise selecionados, o número de partições encontradas para as bases de dados sintéticas estão apresentados na Tabela 6. Nessa tabela também é apresentado qual método de ligação foi usado para o modelo Aglomerativo em cada caso.

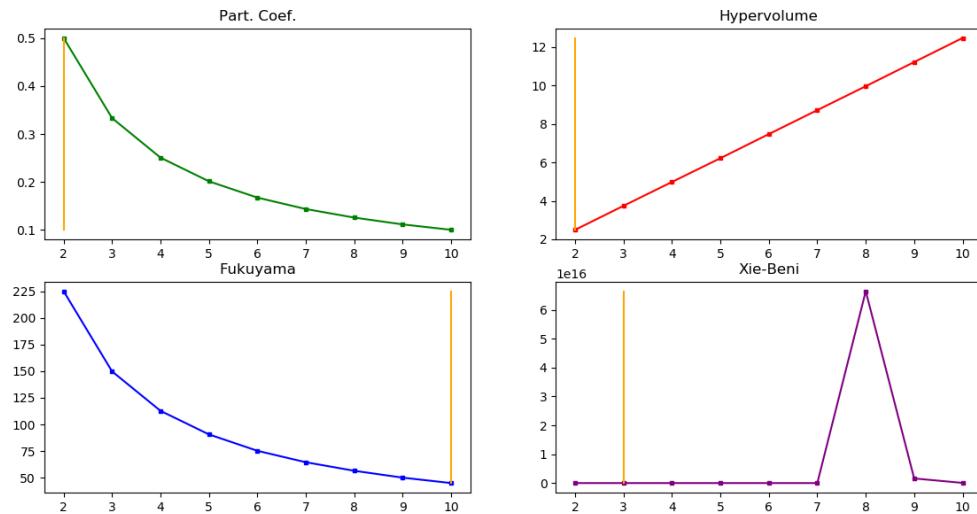


Figura 17 – Coeficientes possibilístico - Tríade (de 2 a 10)

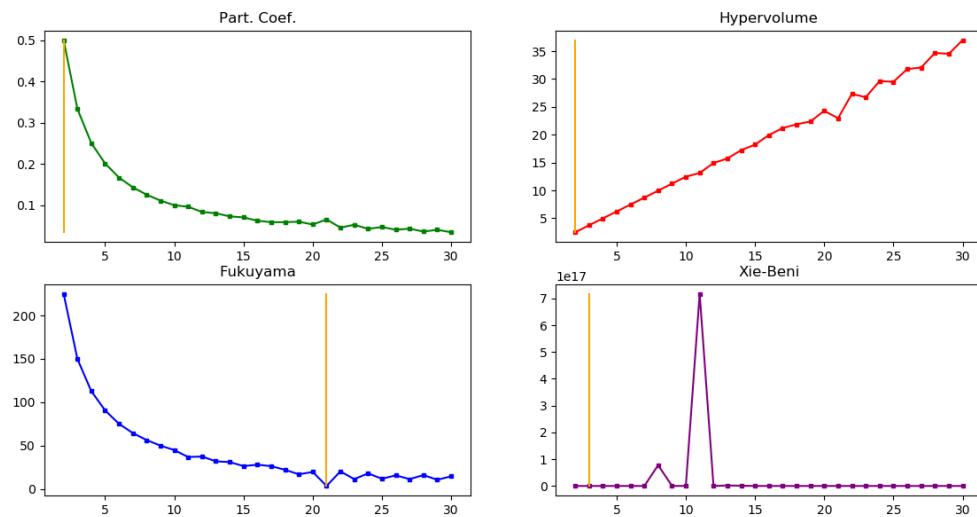


Figura 18 – Coeficientes possibilístico - Tríade (de 2 a 31)

Tabela 6 – Tabela com o número de grupos para cada um dos algoritmos referentes a cada base sintética.

Banco	<i>k</i> -médias	Aglomerativo	<i>c</i> -médias	Gustafson	possibilístico
Tríade	9	6(extensa)	6	2	3
Grupo de Anéis	2	2(extensa)	2	2	2
Interlaço	4	3(extensa)	6	4	3
Anel Interno	5	17(média)	3	3	6
Ruídos	4	4(extensa)	5	5	4
Luas	8	2(extensa)	6	4	6

Destaca-se o número de partições encontrado pelo modelo Aglomerativo para a base de

dados “Anel Interno”, onde nenhum valor foi encontrado dentro de intervalo de 2 a 10 clusters.

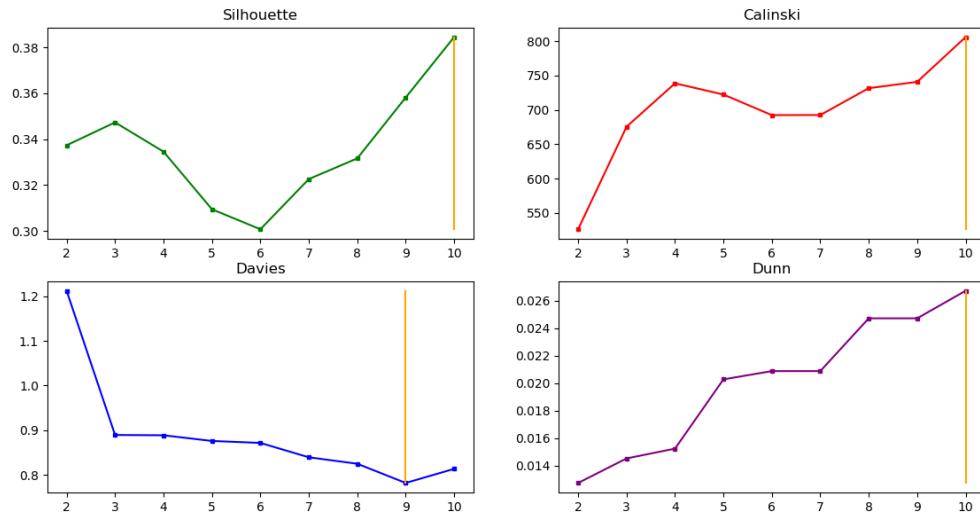


Figura 19 – Coeficientes Aglomerativo - Anel Interno (de 2 a 10)

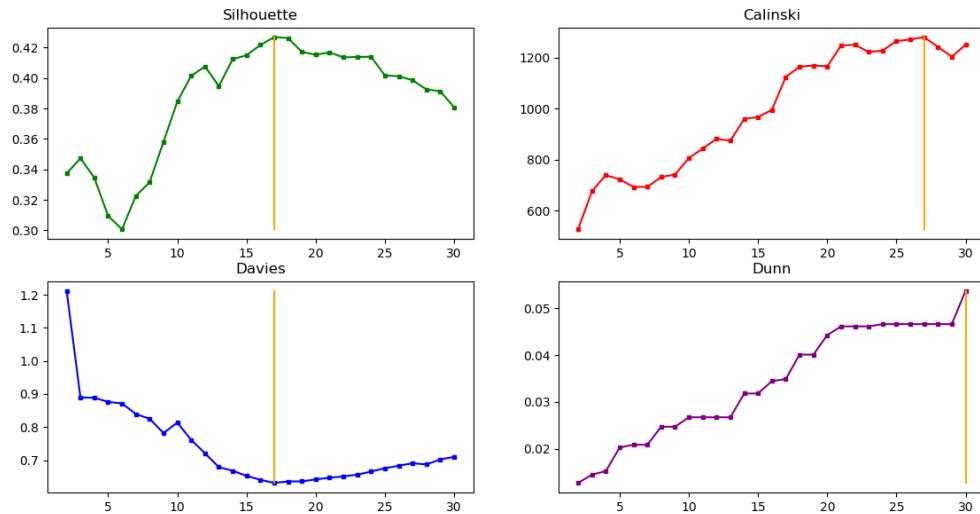


Figura 20 – Coeficientes Aglomerativo - Anel Interno (de 2 a 31)

Mostrando o porque da escolha de 17 como melhor número de clusters, observa-se na Figura 19 que a única indicação feita no intervalo de 2 a 10 clusters foi para um mínimo local em 9 pelo coeficiente de Davies, observado através do coeficiente de Davies-Bouldin. Porém, observando uma escala mais ampla, na Figura 20, os coeficientes de Davies e da Silhueta definem dezessete (17) como melhor valor.

### 5.1.2 VISUALIZAÇÃO DOS AGRUPAMENTOS

A base de dados sintética, por ser de duas dimensões, permite a representação e análise visual dos agrupamentos gerados. Nesta seção são apresentados os agrupamentos, o tempo de

execução dos modelos e uma comparação dos coeficientes empregados no estudo.

#### 5.1.2.1 Tríade

A base composta pela tríade de elipses compõem no espaço um conjunto com formas próximas que, aos nossos olhos exibe três grupos, porém para a maioria das técnicas de agrupamento a proximidade das formas faz com que sejam encontrados grupos distintos. Na Tabela 7 são demonstrados os resultados dos coeficientes para os modelos configurados de acordo com os resultados da avaliação interna, em negrito são destacados os melhores resultados e esse padrão segue para as demais tabelas desse capítulo. As imagens da Figura 21 apresentam também as respostas de cada modelo para essa base de dados.

Tabela 7 – Resultados das métricas para comparação relativa com a base de dados 'tríade'.

<i>Coeficientes \ Bases</i>	<i>k-médias</i>	<i>Aglomerativo</i>	<i>MeanShift</i>
Silhouette	<b>0,46334</b>	0,40507	0,39718
Calinski	<b>237,54603</b>	169,62177	144,00817
Davies	<b>0,61353</b>	0,73278	0,93789
Dunn	<b>0,07042</b>	0,04984	0,03681
<i>Coeficientes \ Bases</i>	<i>c-médias</i>	<i>Gustafson</i>	<i>possibilístico</i>
Silhouette	0,45028	0,33349	0,00713
Calinski	197,10208	90,46411	49,58269
Davies	0,65921	1,27451	1,26119
Dunn	0,05503	0,02553	0,02572
Part. Coef.	0,58695	<b>0,74906</b>	0,59806
Hypervolume	<b>1,01253</b>	1,37189	1,0401
Fukuyama	<b>-193,87329</b>	94,23809	74,64773
Xie Beni	<b>0,08187</b>	0,33303	1,82E+05

Comparando as duas classes de modelos de agrupamento em estudo, *fuzzy* e clássico, ambas mantém um mesmo balanceamento para o agrupamento dessa base, que por apresentar formas esféricas, fez com os modelos K-médias e o C-médias *fuzzy* obtivesse melhores valores, em comparação com os valores obtidos pelos outros sistemas. Especificamente para essa base, a única diferença mais notável foi o fato dos algoritmos *fuzzy* gerarem um número menor de divisões, algo que nesse caso é mais compatível com a percepção visual dos agrupamentos.

O agrupamento que mostrou um comportamento diferente foi obtido a partir do método Possibilístico, que mesmo indicando 3 como melhor divisão de grupos, constitui uma tendência de convergir centroides em alguns casos, ou seja, as pertinências levam os centros de dois ou mais grupos a terem uma mesma posição como resposta. Isso acontece por não existir um limite nos valores de sua pertinência, o que pode levar a formações como a observada na imagem da Figura 21 (f), sugerindo também um número menor de partições.

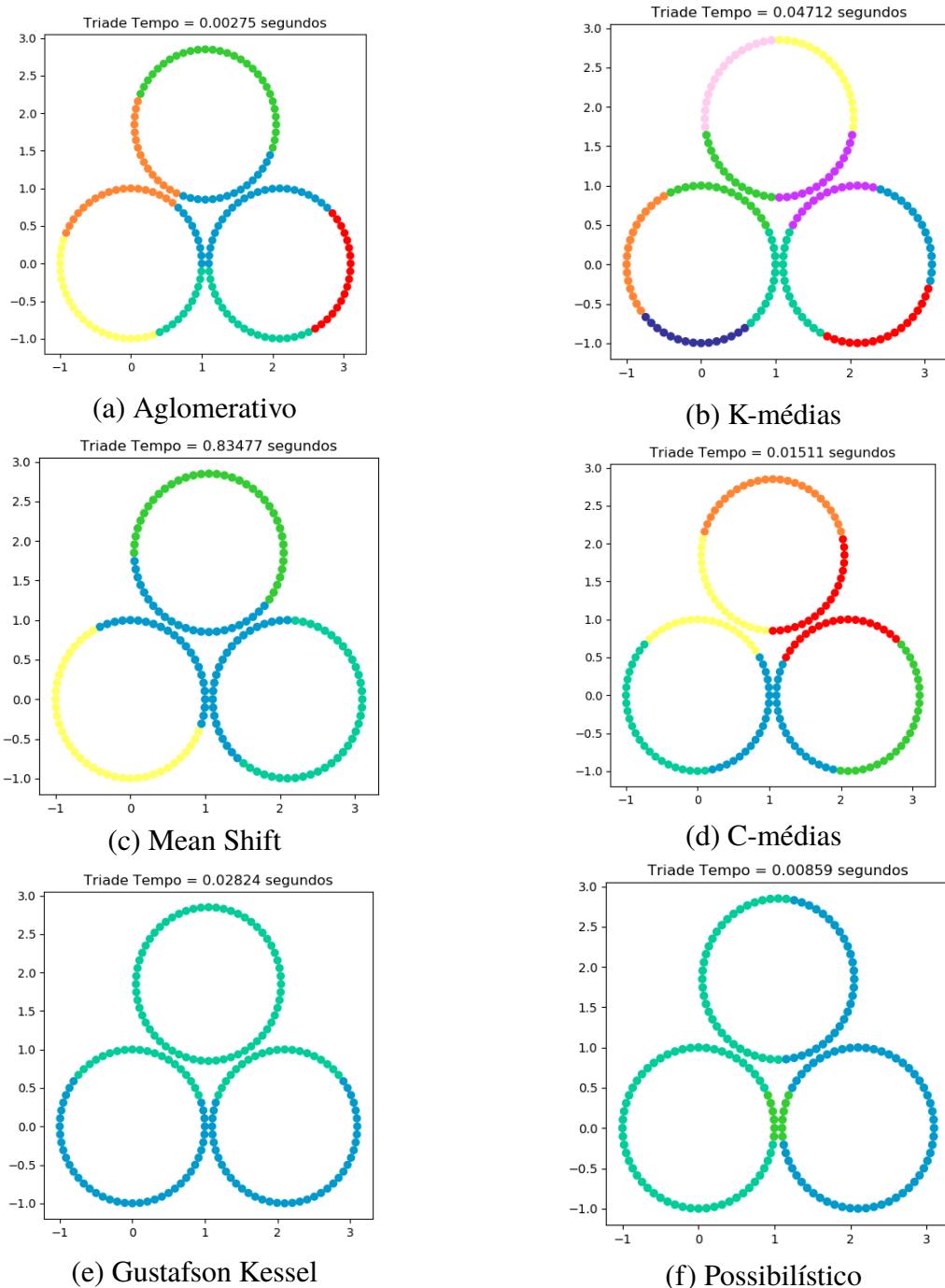


Figura 21 – Agrupamentos trabalhados na base de dados "Tríade"

#### 5.1.2.2 Grupo de Anéis

Para esta base de dados todos os algoritmos apresentaram agrupamentos semelhantes, com exceção do Mean Shift, como pode ser observado na Figura 22 (a). Na Tabela 8 é possível observar os resultados dos coeficientes e a resposta dos demais agrupamentos na Figura 22 (b).

Tabela 8 – Resultados das métricas para comparação relativa com a base de dados 'Grupo de Anéis'.

<i>Coeficientes \ Bases</i>	<i>k-médias</i>	<i>Aglomerativo</i>	<i>MeanShift</i>
Silhouette	0,61505	0,61505	0,40246
Calinski	395,20569	395,20569	317,84723
Davies	0,56077	0,56077	1,02358
Dunn	0,53853	0,53853	0,04564
<i>Coeficientes \ Bases</i>	<i>c-médias</i>	<i>Gustafson</i>	<i>possibilístico</i>
Silhouette	0,61505	0,61505	0,61505
Calinski	395,20569	395,20569	395,20569
Davies	0,56077	0,56077	0,56077
Dunn	0,53853	0,53853	0,53853
Part. Coef.	0,84809	0,84665	0,30806
Hypervolume	1,51938	1,51029	1,01062
Fukuyama	-339,55505	-336,90994	-82,89306
Xie Beni	0,09045	0,09131	3,07E-02

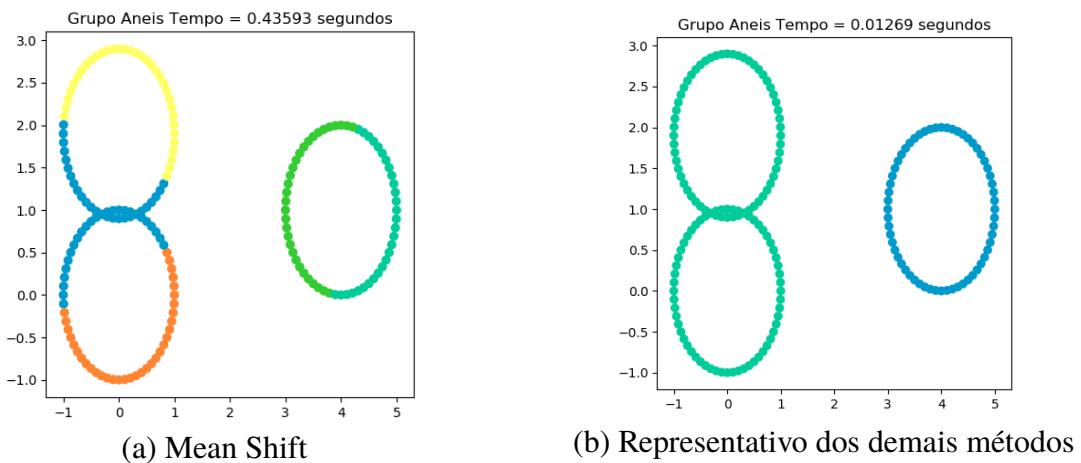


Figura 22 – Agrupamentos trabalhados na base de dados "Grupo de Anéis"

#### 5.1.2.3 Interlaço

Essa base foi construída com apenas o interlaço de duas elipses, mas que foram propositalmente bem unidas para identificar como isso afetaria os agrupamentos. Assim, na maioria dos casos, os modelos trataram a figura como um único círculo distorcido nos polos. Os resultados podem ser visualizados pela Tabela 9 de coeficientes e pelas imagens da Figura 23.

Tabela 9 – Resultados das métricas para comparação relativa com a base de dados 'Interlaço'.

<i>Coeficientes \ Bases</i>	<i>k-médias</i>	<i>Aglomerativo</i>	<i>MeanShift</i>
Silhouette	<b>0,43483</b>	0,40805	0,42658
Calinski	143,78279	102,55757	112,39145
Davies	<b>0,67538</b>	0,79454	0,76679
Dunn	0,06683	0,04393	0,04951
<i>Coeficientes \ Bases</i>	<i>c-médias</i>	<i>Gustafson</i>	<i>Possibilístico</i>
Silhouette	0,41159	0,41886	0,43149
Calinski	<b>177,8416</b>	134,77353	115,35661
Davies	0,69175	0,67972	0,75434
Dunn	<b>0,08231</b>	0,05694	0,04962
Part. Coef.	0,59696	<b>0,7309</b>	0,50802
Hypervolume	0,3482	<b>0,3082</b>	0,81548
Fukuyama	<b>-58,62839</b>	-54,08514	9,91626
Xie Beni	<b>0,08393</b>	0,09407	3,68E-01

Destaca-se que as avaliações dos agrupamentos para os métodos *fuzzy* foram melhores, mesmo com um maior número de partições. No geral, tanto os modelos crisp quanto os modelos *fuzzy* obtiveram agrupamentos semelhantes, porém o custo computacional dos modelos *fuzzy* ainda é maior.

#### 5.1.2.4 Anel Interno

Esta base, assim como as que se seguem nas próximas seções, são pequenas, mas trabalham com ruídos para influenciar a formação dos grupos.

A primeira observação é que nenhum dos algoritmos foi capaz de separar o anel interno do anel externo. O algoritmo Mean Shift, que tem a característica de se comportar de forma mais adequada em bases de dados deste tipo, também teve seu comportamento influenciado pelo ruído e não conseguiu realizar bem a tarefa.

Os resultados são observados nas imagens da Figura 24 e na Tabela 10.

Do lado crisp, como já visto na identificação do número de clusters, o algoritmo aglomerativo é o que teve maiores dificuldades para agrupar um número pequeno de partições, mas como demonstrado pelos coeficientes, isso não necessariamente representa um erro, já que ele foi o maior destaque dentre os resultados. Todavia, é válido observar que quanto mais grupos são gerados, mais difícil é extrair informação sobre a descrição do grupo gerado.

Do lado *fuzzy*, o método Possibilístico gerou novamente um dos seus erros já mencionado, que é proveniente da junção de um ou mais centros. Assim, mesmo que tenha sido requisitada uma divisão de 6 partições, os resultados geraram uma divisão confusa de apenas 2 grupos.

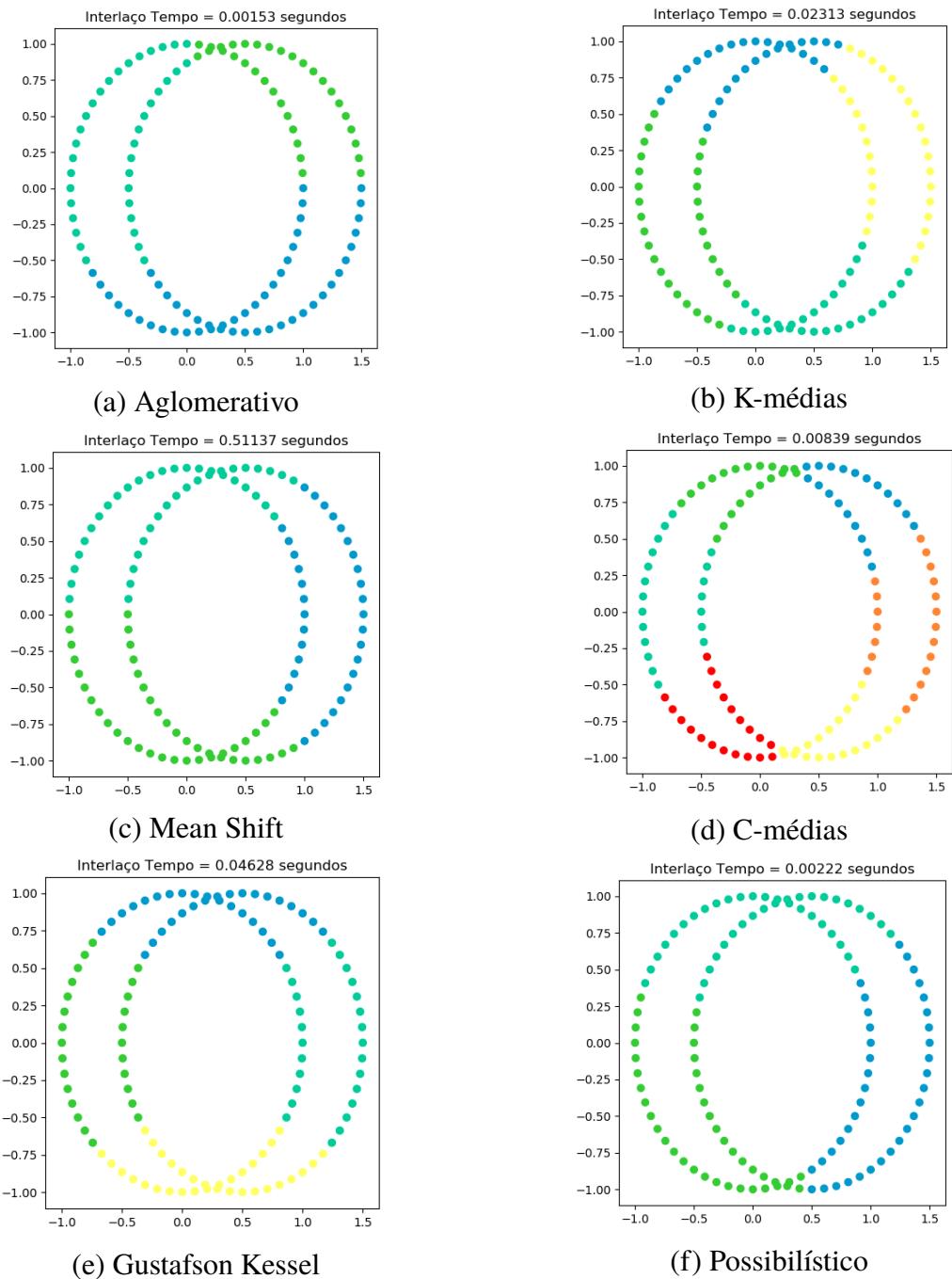


Figura 23 – Agrupamentos trabalhados na base de dados "Interlaço"

#### 5.1.2.5 Ruídos

Essa talvez é umas das bases sintéticas mais similares com o que pode ocorrer em casos reais e ideais para um agrupamento. Na Tabela 11 estão os resultados dos coeficientes e nas imagens da Figura 25 os resultados visuais dos agrupamentos.

Tabela 10 – Resultados das métricas para comparação relativa com a base de dados 'Anel Interno'.

<i>Coeicientes \ Bases</i>	<i>k-médias</i>	<i>Agglomerativo</i>	<i>MeanShift</i>
Silhouette	0,35603	<b>0,42693</b>	0,35339
Calinski	900,26488	<b>1123,60556</b>	690,4987
Davies	0,85398	<b>0,6313</b>	0,88081
Dunn	0,02118	<b>0,03488</b>	0,01033
<i>Coeicientes \ Bases</i>	<i>c-médias</i>	<i>Gustafson</i>	<i>possibilístico</i>
Silhouette	0,38423	0,37895	0,04827
Calinski	782,8232	774,73373	77,92336
Davies	0,84251	0,84878	1,80848
Dunn	0,01293	0,01268	0,00594
Part. Coef.	0,62248	0,64516	<b>1,14613</b>
Hypervolume	0,36593	<b>0,34978</b>	1,04984
Fukuyama	-58,48579	<b>-64,33655</b>	400,65446
Xie Beni	<b>0,15374</b>	0,15668	2,37E+05

Tabela 11 – Resultados das métricas para comparação relativa com a base de dados 'Ruídos'.

<i>Coeicientes \ Bases</i>	<i>k-médias</i>	<i>Agglomerativo</i>	<i>MeanShift</i>
Silhouette	<b>0,63039</b>	0,62639	0,22207
Calinski	1818,27991	1727,60341	764,63537
Davies	<b>0,55604</b>	0,58135	0,38812
Dunn	0,03119	<b>0,07799</b>	0,01426
<i>Coeicientes \ Bases</i>	<i>c-médias</i>	<i>Gustafson</i>	<i>possibilístico</i>
Silhouette	0,59628	0,59733	0,5893
Calinski	<b>1951,57402</b>	1860,29745	1702,34597
Davies	0,57913	0,59264	0,62386
Dunn	0,01634	0,023	0,00867
Part. Coef.	<b>0,78784</b>	0,77835	0,32715
Hypervolume	13,21628	<b>12,88593</b>	3,17649
Fukuyama	<b>-35936,03953</b>	-35446,79512	-17944,97403
Xie Beni	0,14602	<b>0,14082</b>	7,43E-03

Com exceção do Mean Shift, todos os algoritmos apresentaram uma divisão com grupos que podem ser percebidos numa análise visual, tendo também muitos destaques dentre os coeficientes. A maior comparação que pode ser feita aqui é que visualmente é notável na base cinco pontos de densidade, onde o método *fuzzy* conseguiu se aproximar melhor dessa visualização, mesmo que os coeficientes não apontem os melhores resultados.

Os algoritmos clássicos demonstram, em vários casos, formar grupos de formação visual menos aparente, mas com melhores resultados para os coeficientes. Já os modelos *fuzzy*, em suma, parecem tentar indicar uma divisão mais comum ao olhar humano, onde é possível observar melhor uma formação visual parecida entre os grupos.

Assim, é possível levantar questões como o que é mais interessante para uma aplicação, a observação de coeficientes adequados ou uma divisão visualmente interessante para o humano

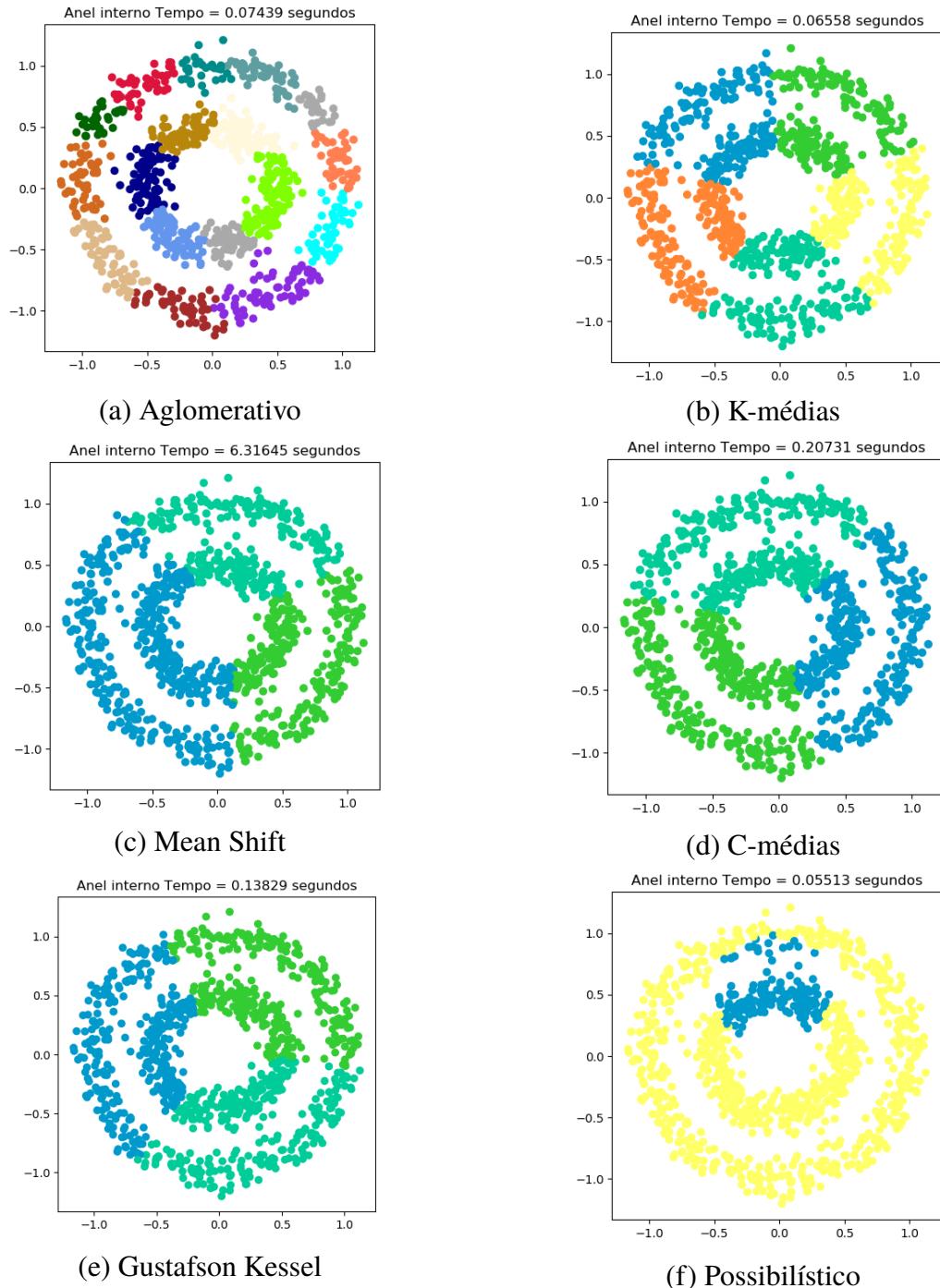


Figura 24 – Agrupamentos trabalhados na base de dados "Anel Interno"

que analisa o conjunto de dados.

Além dessas questões, deve-se também ter em mente que as análises foram realizadas com coeficientes inicialmente pensados para técnicas clássicas, voltando a mostrar como é difícil definir um parâmetro exato para dizer qual agrupamento é melhor.

Em relação ao Mean shift observa-se um dos problemas que podem ocorrer com a definição de um raio de vizinhança, que nesse caso foi pequeno demais para gerar uma resposta coerente, e apenas agrupou a maioria dos pontos em um cluster próprio. Esse erro pode ser

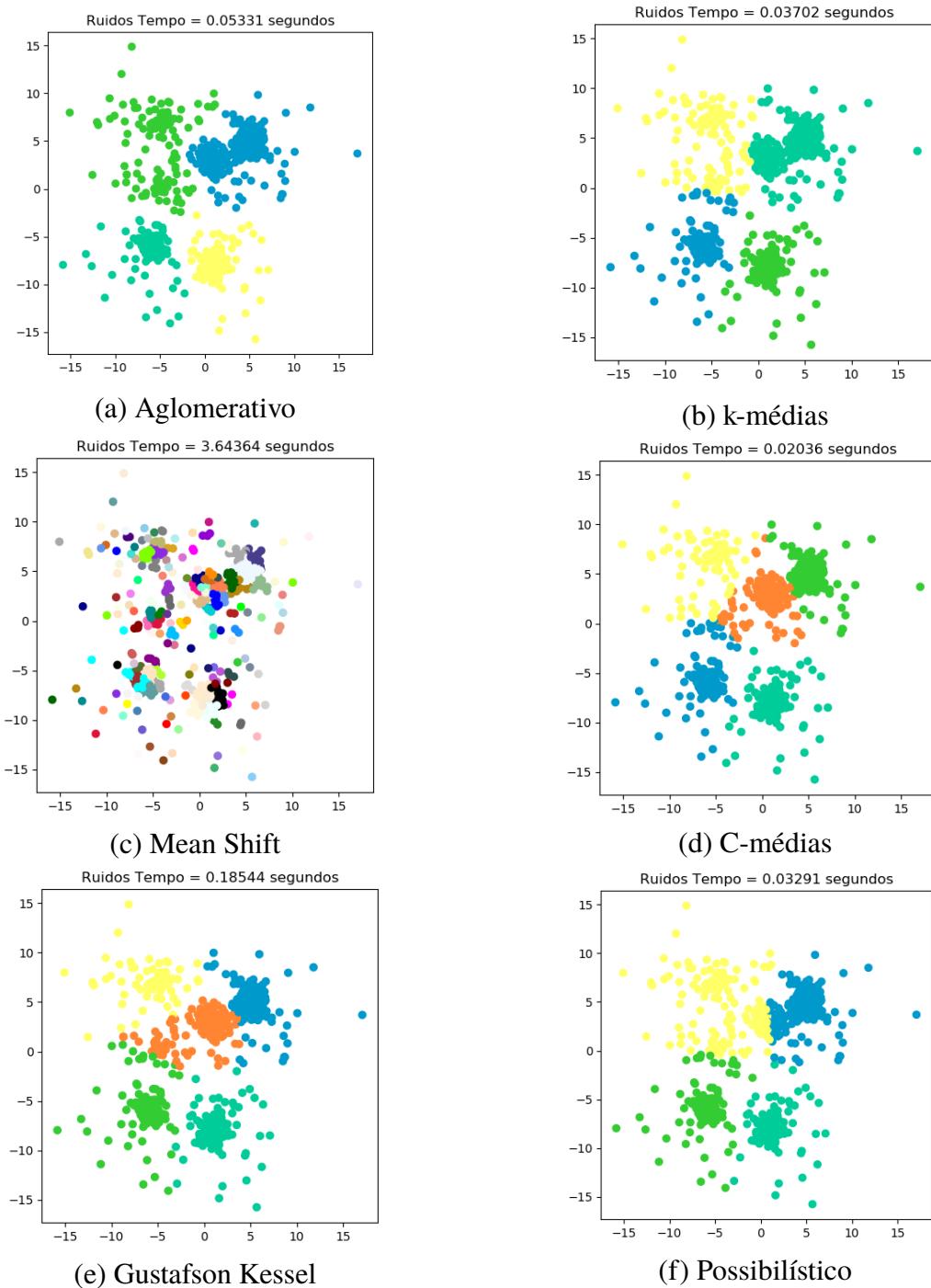


Figura 25 – Agrupamentos trabalhados na base de dados "Ruídos"

minimizado com o cálculo prévio das médias das distâncias, mas ainda assim, não impede que o valor de raio de vizinhança gerado não acarrete em agrupamentos inadequados.

#### 5.1.2.6 Luas

Para esta base de dados o algoritmo Agglomerativo e o Mean Shift foram os modelos capazes de chegar mais próximos ao que seria um agrupamento visual adequado, mesmo os valores de coeficientes para os modelos mais simples como o K-médias serem os melhores. Os

resultados são visualizados pela Tabela 12 e as imagens da Figura 26.

Tabela 12 – Resultados das métricas para comparação relativa com a base de dados 'Luas'.

<i>Coeficientes \ Bases</i>	<i>k-médias</i>	<i>Agglomerativo</i>	<i>MeanShift</i>
Silhouette	<b>0,49973</b>	0,48544	0,46303
Calinski	<b>2203,3117</b>	1453,89254	1258,25288
Davies	<b>0,59478</b>	0,78061	0,79049
Dunn	0,03431	0,01971	0,01081
<i>Coeficientes \ Bases</i>	<i>c-médias</i>	<i>Gustafson</i>	<i>possibilístico</i>
Silhouette	0,48019	0,46319	0,25992
Calinski	1836,48083	1296,73497	718,1968
Davies	0,65743	0,82853	1,16352
Dunn	<b>0,01555</b>	0,00885	0,00547
Part. Coef.	0,62263	<b>0,75969</b>	0,75782
Hypervolume	0,25656	<b>0,2244</b>	1,00522
Fukuyama	<b>-569,92825</b>	-493,66979	193,94456
Xie Beni	<b>0,07801</b>	0,2154	1,68E+02

O objetivo principal neste conjunto de experimentos foi justamente explorar os modelos com relação ao comportamento frente a diferentes formas espaciais na definição de agrupamentos.

## 5.2 EXPERIMENTOS EM BASES CLÁSSICAS

A segunda fase do projeto de pesquisa trabalha com bases de dados clássicas da literatura, descritas na Seção 4.2.3, largamente empregadas em estudos de algoritmos de classificação e agrupamento. Os procedimentos de avaliação seguem o mesmo padrão da seção anterior.

Neste experimento foram empregadas as métricas externas para comparação dos resultados com os rótulos reais de cada base, uma técnica geralmente aplicada ao aprendizado supervisionado, mas que é usufruída em algumas medidas quantitativas do agrupamento. Foi também empregada a técnica de média das distâncias, disponibilizada pelo próprio algoritmo da scikit-learn, com o propósito do cálculo automático do raio de vizinhança para o Mean Shift. Deve-se também levar em consideração o tempo para a interpretação da fórmula, sendo aconselhável pela própria plataforma o uso de valores pequenos de raio de vizinhança, já que a fórmula usada possui tempo de processamento no mínimo quadrático para  $N$  amostras (BUITINCK et al., 2013).

### 5.2.1 NÚMERO DE CLUSTERS

Na Tabela 13 podem ser observados os números de partições obtidos para as três bases consideradas neste experimento, também é apresentado o tipo de ligação usado para o método aglomerativo.

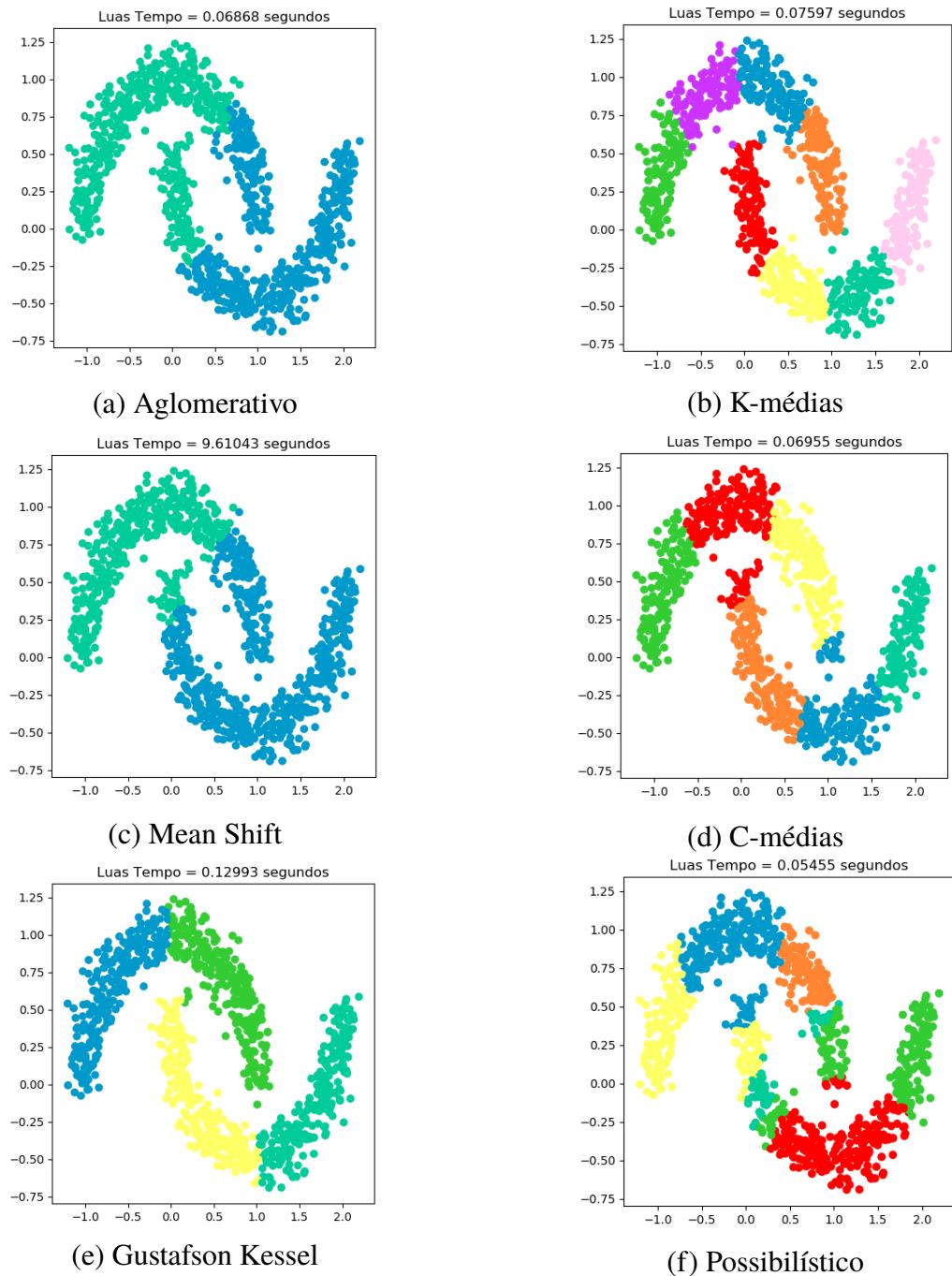


Figura 26 – Agrupamentos trabalhados na base de dados "Luas"

Tabela 13 – Número de grupos para cada um dos algoritmos referentes a cada base científica.

Banco	<i>k</i> -médias	Aglomerativo	<i>c</i> -médias	Gustafson	possibilístico
Íris	2	2(extensa)	4	5	2
Wine	2	2(extensa)	4	2	4
Boston	3	3(extensa)	4	3	5

É possível perceber que alguns algoritmos registraram os mesmos números de partição e que a variação tem uma média de 3 grupos (com exceção da base de dados Boston), valor que corresponde aos rótulos presente nos conjuntos de dados Iris e Wine.

### 5.2.2 ANÁLISE DO AGRUPAMENTO

Esta análise de resultados é também baseada nas métricas da Tabela 4 para as bases de dados Iris e Wine, no qual é realizada uma validação externa, pois estas bases possuem rótulo. O restante da análise se assemelha ao experimento 1, porém neste experimento são empregadas bases de dados muito exploradas na literatura considerando as tarefas de classificação e agrupamento de dados.

#### 5.2.2.1 Validação Supervisionada

Aqui são considerados como número de grupos o números de classes das bases Iris e Wine. Como o Mean Shift realiza a escolha automática do número de cluster, para essa validação em específico, isso pode ser uma desvantagem. Todavia, é válido ressaltar que, apesar de existirem meios de realizar a validação externa em clusters, ela não é eficiente, devendo ser usada apenas para ressaltar outros pontos observados na pesquisa.

Dos resultados é notado os valores representados pelas Tabelas 14 e 15

Tabela 14 – Resultados dos coeficientes supervisionados com a base de dados 'íris'.

<i>Coefficientes \ Bases</i>	<i>k</i> -médias	Aglomerativo	MeanShift
Rand	0,7302382723	0,7311985568	0,5583714438
Completeness	0,7649861514	0,7795958006	<b>0,9490204434</b>
Homogeneity	0,7514854022	0,760800847	0,5537492887
FolkesMallows	0,8208080729	0,8221697785	0,764206332
Mutual Info.	0,7582057278	0,7701409906	0,7249271657
V score	0,75817568	0,7700836616	0,699
<i>Coefficientes \ Bases</i>	<i>c</i> -médias	Gustafson	possibilístico
Rand	0,7294203486	<b>0,772631417</b>	<b>0,772631417</b>
Completeness	0,7542594808	0,80008891	0,80008891
Homogeneity	0,7450433681	<b>0,7882427871</b>	<b>0,7882427871</b>
FolkesMallows	0,8196711597	<b>0,8487259648</b>	<b>0,8487259648</b>
Mutual Info.	0,7496372616	<b>0,7941437605</b>	<b>0,7941437605</b>
V score	0,749623099	<b>0,7941216731</b>	<b>0,7941216731</b>

Para a base de dados Iris observa-se que apesar de nenhum algoritmo ter atingido 1 para os coeficiente, o mesmo que 100% de compatibilidade, teve-se uma aproximação bem grande com a realidade visto que a maioria dos valores estão acima de 0,7, com até 0,94 para o Mean Shift em uma das métricas.

Os algoritmos de Gustafson e os modelos possibilísticos geraram partições idênticas e com os melhores resultados, apesar do destaque do Mean Shift com o coeficiente de completude, contradizendo sua dificuldade de não ter um número de grupos previamente definido. A completude é um coeficiente que pode continuar gerando valores altos quanto menos grupos forem formados, já que quando todos os pontos estão em um mesmo grupo a divisão ainda é “completa” na visão do coeficiente. Além disso, observando os outros valores nota-se que o Mean Shift no final apresentou os piores resultados.

Tabela 15 – Resultados dos coeficientes supervisionados com a base de dados ‘Wine’.

<i>Coefficientes \ Bases</i>	<i>k-médias</i>	<i>Agglomerativo</i>	<i>MeanShift</i>
Rand	0,3711137182	0,2926269172	<b>0,397237</b>
Completeness	0,4287014139	<b>0,5236972239</b>	0,495682
Homogeneity	0,42881232	0,330083562	<b>0,430957</b>
FolkesMallows	0,5835370219	0,619215145	0,633028
Mutual Info.	0,4287568634	0,4157689804	<b>0,462188</b>
V score	0,4287568598	0,4049373046	<b>0,461</b>
<i>Coefficientes \ Bases</i>	<i>c-médias</i>	<i>Gustafson</i>	<i>possibilístico</i>
Rand	0,3539016593	0,3601252536	0,3559325357
Completeness	0,4173118545	0,5115855693	0,5058156919
Homogeneity	0,4162574248	0,3552449516	0,3527337437
FolkesMallows	<b>0,5727954036</b>	0,6366837526	0,6335332696
Mutual Info.	0,4167843062	0,426078335	0,4220870861
V score	0,4167839728	0,4188808557	0,415040051

Para a base de dados Wine, observa-se o contrário, onde agora o Mean Shift realmente gerou o melhor resultado geral, apesar de que observando todos os métodos, os algoritmos não obtiveram boa resposta final.

A utilização desse tipo de validação é um pequeno adendo para tentar verificar se a lógica *fuzzy* poderia ter resultados superiores para encontrar resultados parecidos com os rótulos humanos. No final, os resultados alcançados tanto para os modelos *fuzzy* quanto para os modelos clássicos foram similares, não sendo possível fazer muitos destaques.

### 5.2.2.2 Íris

Considerando as métricas de validação interna e relativa, são analisados nesta seção os valores alcançados na Tabela 13.

Nota-se que os maiores destaques foram aplicados pelos modelos aglomerativo e Possibilístico, que obtiveram agrupamentos idênticos. Outro ponto a se observar, foi que o método Possibilístico obteve os melhores resultados tanto na validação externa quanto relativa dessa base. Os resultados são destacados na Tabela 16.

Tabela 16 – Resultados das métricas para comparação relativa com a base de dados 'Iris'.

<i>Coeficientes \ Bases</i>	<i>k-médias</i>	<i>Agglomerativo</i>	<i>MeanShift</i>
Silhouette	0,68105	<b>0,68674</b>	0,68579
Calinski	513,92455	502,82156	509,70343
Davies	0,40429	<b>0,38275</b>	0,38855
Dunn	0,07651	<b>0,33891</b>	0,08111
<i>Coeficientes \ Bases</i>	<i>c-médias</i>	<i>Gustafson</i>	<i>Possibilístico</i>
Silhouette	0,49273	0,33621	0,68674
Calinski	<b>528,34728</b>	209,4967	502,82156
Davies	0,7814	2,33849	0,38275
Dunn	0,08234	0,04139	0,33891
Part. Coef.	0,70679	0,59381	<b>0,85055</b>
Hypervolume	0,02085	<b>0,01928</b>	0,02328
Fukuyama	<b>-476,00429</b>	-430,08252	-325,62179
Xie Beni	0,19533	3,35612	<b>0,06332</b>

Observando o parâmetro de tempo de execução, os modelos *fuzzy* não se apresentaram mais altos para as bases de dados testadas. O tempo para o modelo Possibilístico e suas ótimas respostas foram um ponto positivo nesse caso.

#### 5.2.2.3 Wine

Para a base de dados Wine, destaca-se novamente o modelo Aglomerativo, que vem se apresentando adequado, porém esse método não consegue lidar com o agrupamento bases de muito grandes.

Apresentasse os resultados obtidos para essa base na Tabela 17, com destaque aos melhores resultados.

Tabela 17 – Resultados das métricas para comparação relativa com a base de dados 'Wine'.

<i>Coeicientes \ Bases</i>	<i>k-médias</i>	<i>Agglomerativo</i>	<i>MeanShift</i>
Silhouette	0,65685	<b>0,65873</b>	0,50249
Calinski	505,42931	483,11286	454,05894
Davies	0,47878	<b>0,45862</b>	0,55615
Dunn	0,02282	<b>0,07159</b>	0,01952
<i>Coeicientes \ Bases</i>	<i>c-médias</i>	<i>Gustafson</i>	<i>possibilístico</i>
Silhouette	0,55467	-0,04799	0,55362
Calinski	<b>701,80068</b>	10,95568	697,01927
Davies	0,54621	2,26418	0,54545
Dunn	0,02254	0,00388	0,0246
Part. Coef.	<b>0,78298</b>	0,52032	0,77169
Hypervolume	0,3881	<b>2,11444</b>	0,65441
Fukuyama	<b>-14389231,57</b>	8962342,348	-14036973,47
Xie Beni	<b>0,09423</b>	7,03643	0,10139

Sem muitos destaques, o modelo de Gustafson Kessel foi quem gerou as piores respostas aos coeficiente. Observando todas as tabelas de respostas apresentadas até aqui, o modelo de Gustafson vem mostrando bons resultados a alguns dos coeficientes propriamente *fuzzy*, mas não se nota eficiência para os demais casos.

#### 5.2.2.4 Boston

Com os resultados da Tabela 18, nessa base é notável a dificuldade dos modelos *fuzzy* em realizar o agrupamento, principalmente por parte de Gustafson Kessel que gerou seu erro por correlação linear (vide Seção 3.4.5), incapacitando-o de agrupar esta base de dados. Assim, com exceção da métrica de Calinski, que se mostrou favorável aos algoritmos *fuzzy*. Os melhores resultados foram observados para os modelos clássicos e estão destacados em negrito.

Tabela 18 – Resultados das métricas para comparação relativa com a base de dados 'Boston'.

<i>Coeicientes \ Bases</i>	<i>k-médias</i>	<i>Agglomerativo</i>	<i>MeanShift</i>
Silhouette	<b>0,72343</b>	0,72086	<b>0,72343</b>
Calinski	1353,23688	1296,91003	1353,23688
Davies	0,32206	<b>0,2951</b>	0,32206
Dunn	0,10343	<b>0,12433</b>	0,10343
<i>Coeicientes \ Bases</i>	<i>c-médias</i>	<i>Gustafson</i>	<i>possibilístico</i>
Silhouette	0,49846	–	0,41797
Calinski	848,35983	–	<b>1452,90379</b>
Davies	0,80573	–	0,90139
Dunn	0,04682	–	0,01348
Part. Coef.	<b>0,75967</b>	–	0,64404
Hypervolume	2797406,012	–	<b>508371,959</b>
Fukuyama	-8464299,264	–	<b>-13578053,08</b>
Xie Beni	<b>0,3183</b>	–	0,50485

Até este ponto da pesquisa podem ser destacados os seguinte pontos:

- Os modelos baseados na lógica *fuzzy* geraram maior número de grupos maiores e;
- o tempo de execução dos modelos *fuzzy* foi comparável ao tempo para os modelos clássicos.

### 5.3 EXPERIMENTOS EM OUTRAS BASES DE DADOS

Neste conjunto de experimentos, o diferencial está nas bases de dados, selecionadas aleatoriamente na internet, descritas na Seção 4.2.4. E ainda, buscou-se extrair informações das bases de dados e não somente validar os agrupamentos a partir dos coeficientes empregados.

Na Tabela 19 podem ser observados os agrupamentos gerados para cada modelos nas duas bases de dados consideradas, O parenteses ao lado do método aglomerativo representa a ligação usada para agrupar.

Tabela 19 – Tabela com o número de grupos para cada um dos algoritmos referentes a cada base prática.

Banco	<i>k</i> -médias	Aglomerativo	<i>c</i> -médias	Gustafson	possibilístico
Seguro Clínico	2	2(simples)	7	6	7
Vendas Online	3	3(ward)	3	5	3

Com mais esta etapa observa-se novamente a tendência dos modelos *fuzzy* apresentarem resultados com um maior número de grupos.

Buscando uma análise mais detalhada dos resultados foram gerados gráficos 2D a partir dos atributos de cada base de dados numa combinação 2 a 2.

#### 5.3.1 SEGURO CLÍNICO

Tabela 20 – Resultados das métricas para comparação relativa com a base de dados 'Seguro Clínico'

Coeficientes \ Bases	<i>k</i> -médias	Aglomerativo	MeanShift
	Silhouette	0,82622	<b>0,86286</b>
Coeficientes \ Bases	<i>c</i> -médias	Gustafson	possibilístico
	Silhouette	0,72494	0,35569
Dunn	0,00017	<b>0,00016</b>	0,00062
Calinski	<b>9549,03161</b>	3826,97096	9394,07659
Davies	0,46952	0,83494	0,47192
Dunn	0,00045	0	0,00045
Part. Coef.	<b>0,87241</b>	0,61989	0,85109
Hypervolume	10026591410	<b>18421331316</b>	25762798688
Fukuyama	<b>-5,26E+21</b>	-1,35E+21	-5,08E+21
Xie Beni	<b>0,09126</b>	10,89377	0,09889

Na Tabela 20 os maiores destaques vão para o K-médias e o Aglomerativo, mas mantendo uma resposta parecida entre todos os modelos. Para o modelo Aglomerativo, apesar das melhores respostas aos coeficientes, com análise de seus grupos observados na Figura 27, observa-se que todos os dados, com exceção de um ponto, foram alocados para um mesmo cluster. Esse caso pode gerar boas respostas aos coeficientes uma vez que o dado isolado é realmente discrepante em relação aos demais, outra informação que foi analisada nas anotações que se seguem.

Observa-se nos gráficos das Figuras 29 a 49 o agrupamento realizado pelo FCM. No decorrer desta Seção são gerados gráficos deste formato que permitem uma análise mais detalhada dos resultados.

Com os resultados dos coeficientes e adicionando as informações dos gráficos, é possível notar de início que apesar das melhores medidas *crisp* terem sido estudadas em respostas com um número menor de grupos, os modelos não demonstraram muitas informações para essa base, o que prejudicou o desempenho desses algoritmos. Já os métodos *fuzzy*, apesar da análise mais complexa, demonstram facilidade em indicar números maiores de grupos, sem prejudicar suas partições, tendo resultados muito próximos a teoria clássica nas métricas comparativas.

A principal característica observada em todos algoritmos é a tendência em separar seus valores principalmente pelos custos, como observado pelas separações horizontais feitas nas figuras relativas a esse atributo (Figuras 35, 38, 41, 42 e 46). Isso provavelmente se deve pelo fato que essa é a característica com maior limite de vaiação das respostas, um fato interessante já que esse é o atributo mais observado em muitos trabalhos práticos, gerando respostas mais ricas em informações.

Nas Figuras 27 e 28, mostra-se como os modelos aglomerativo e k-médias realizaram seus agrupamentos respectivamente, utilizando a visão custos x idade para melhor visualização. Observa-se como esses dois algoritmos, apesar de indicarem um dos melhores resultados a partir dos coeficientes, não são muito ricos em informações e quase não é possível a retirada de conhecimento. A única característica observada foi ressaltada pelo método aglomerativo, que formou um grupo com apenas um elemento. Este grupo mostrou que os custos desse ponto específico realmente são discrepantes dos demais, uma vez que mesmo o indivíduo referenciado possuindo características comuns ao conjunto (mulher de 18 anos, não fumante e sem filhos), pagou um seguro abusivo comparado com os outros. Ainda que isso seja um claro conflito de ideias, ressalta-se que essa base de dados apresenta apenas informações básicas de cada cliente, sem constar históricos de doenças, tendências genéticas e outros pontos relevantes na hora de calcular um plano de saúde.

Dos próximos destaques mostra a análise dos resultados do C-Médias, o método Possibilístico e o Mean Shift, que agruparam também com uma divisão de custos, mas com 7 a 8 níveis de preços. As principais diferenças entre os resultados dos métodos são duas: a primeira foi a capacidade do Mean Shift de categorizar um oitavo grupo para demonstrar o ponto discrepante já observado com o aglomerativo, e a segunda é a divisão mais "equilibrada" dos agrupamentos

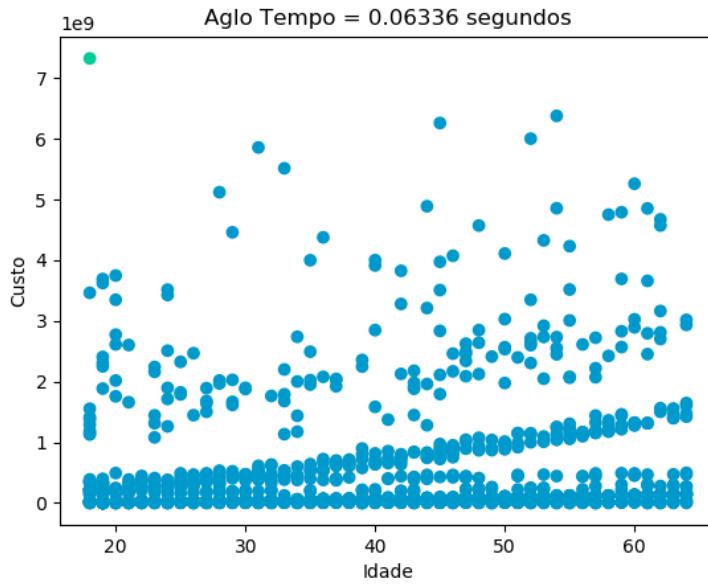


Figura 27 – Resultado gráfico representativo do método aglomerativo agrupando o ponto isolado da base de seguros clínicos

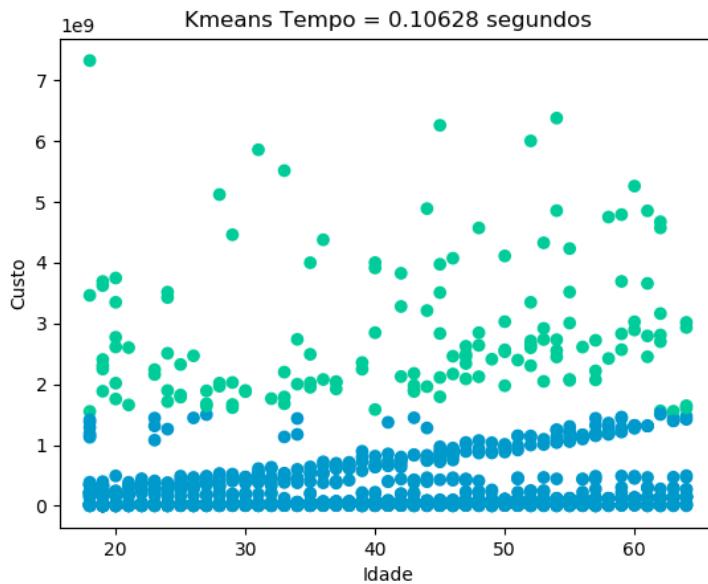


Figura 28 – Resultado gráfico representativo do método do k-médias agrupando os custos da base de seguros clínicos

*fuzzy*. Ou seja, os custos foram separados em grupos de diferença de valores mais simétricos (de 0 a 500, 500 a 1000 por exemplo), o que em alguns casos possibilitou mostrar que o número de filhos não influência fortemente no preço, algo não demonstrado pelo Mean Shift.

Outras informações que a divisão desses algoritmos podem apresentar podem ser observadas na Figura 41, onde ser fumante influencia nos custos, já que os fumantes apresentam níveis mais altos nos gastos. Ainda pode-se também observar com a Figura 42 uma maior variedade de preços nas regiões 3 e 4, o que pode indicar que são lugares com custos mais altos

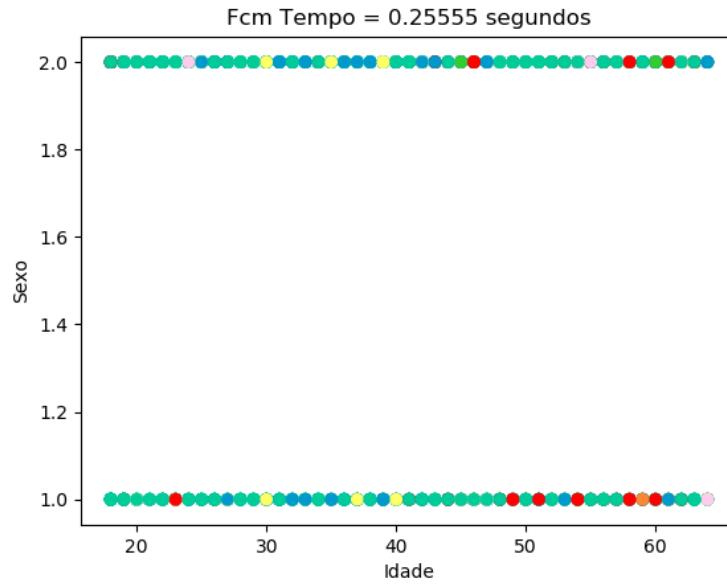


Figura 29 – Agrupamento FCM para Gênero X Idade do seguro clínico

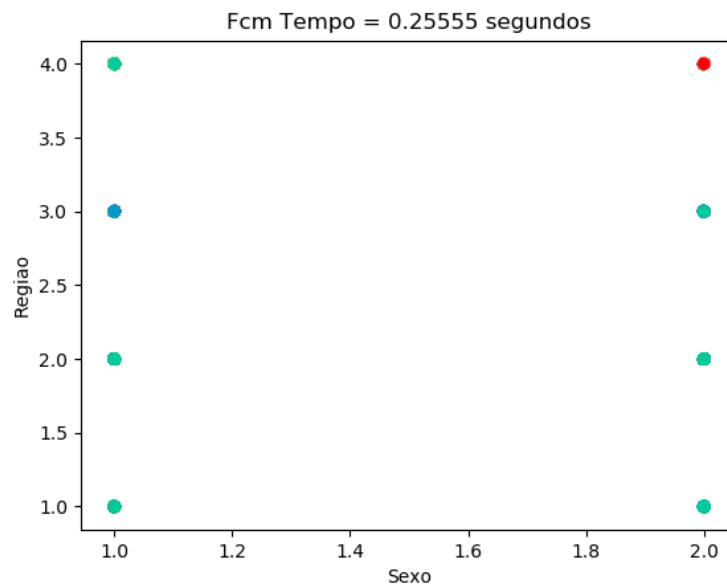


Figura 30 – Agrupamento FCM para Região X Gênero do seguro clínico

de vida.

A abordagem empregada para exploração das bases de dados neste terceiro conjunto de experimentos demonstrou como é possível confirmar hipóteses e descobrir regras a partir dos modelos de agrupamento.

Por fim, ainda aborda-se o que foi feito pelo agrupamento do Gustafson Kessel (Figura 50), o único algoritmo que realizou seu agrupamento não só por grande influência dos custos, mas demonstrou ter levado em consideração idades altas e baixas em alguns casos. Essa facilidade em usufruir melhor das outras características dos indivíduos pode ser muito útil em algumas

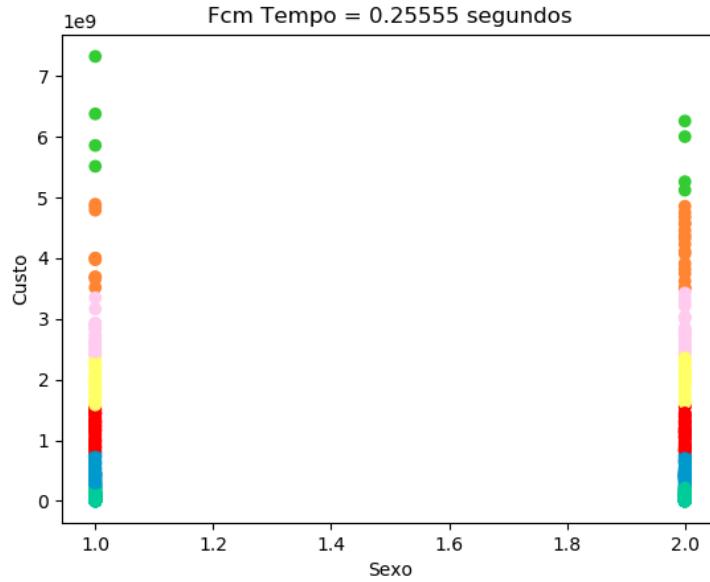


Figura 31 – Agrupamento FCM para Custo X Gênero do seguro clínico

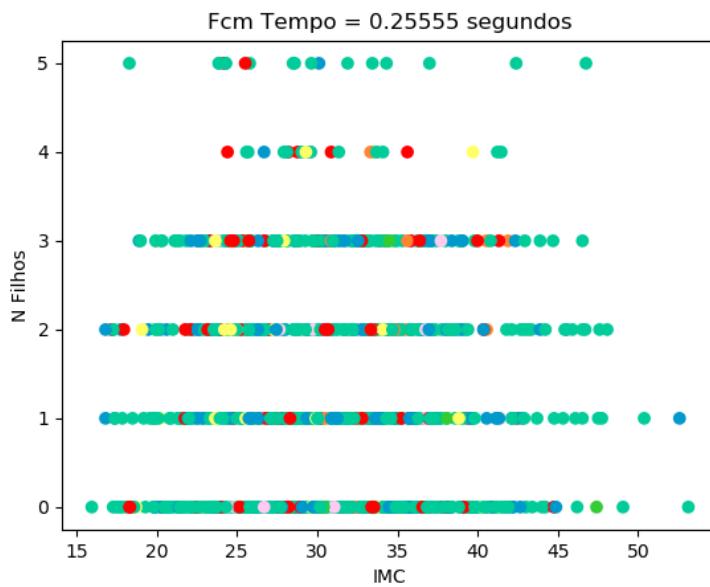


Figura 32 – Agrupamento FCM para n° filhos X IMC do seguro clínico

ocasiões, mas isso é um fator que varia para cada proposta e nesse caso acabou quebrando o potencial que o agrupamento tinha, dificultando sua análise.

Com relação ao tempo de processamento, novamente esse fator demonstrou-se favorável ao método Aglomerativo. O segundo mais rápido foi o K-médias, levemente melhor que o FCM que obteve resultados muito melhores e com um maior número de divisões.

O método Possibilístico, que alcançou o aglomerativo em diversas outras execuções, e sem muita discrepância de tempo, conseguiu também resultado tão bons quanto o FCM e o Mean Shift. Por fim, o Mean Shift, que também apresentou-se muito informativo, todavia gerou o valor

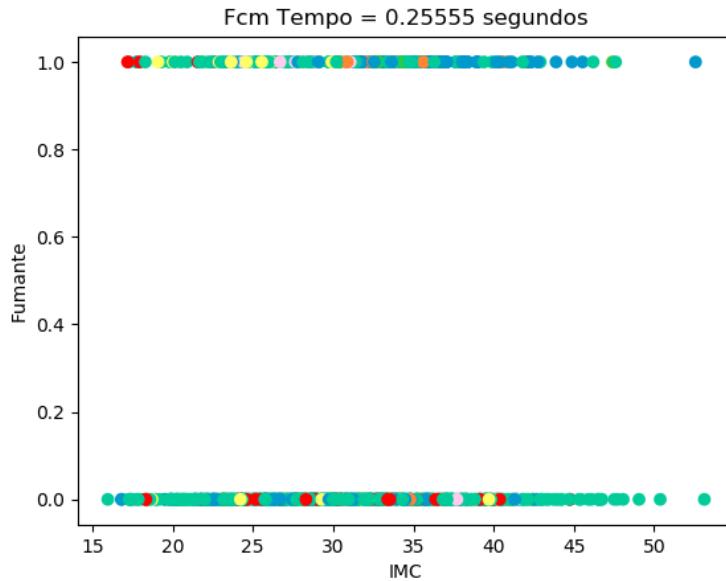


Figura 33 – Agrupamento FCM para Fumante X IMC do seguro clínico

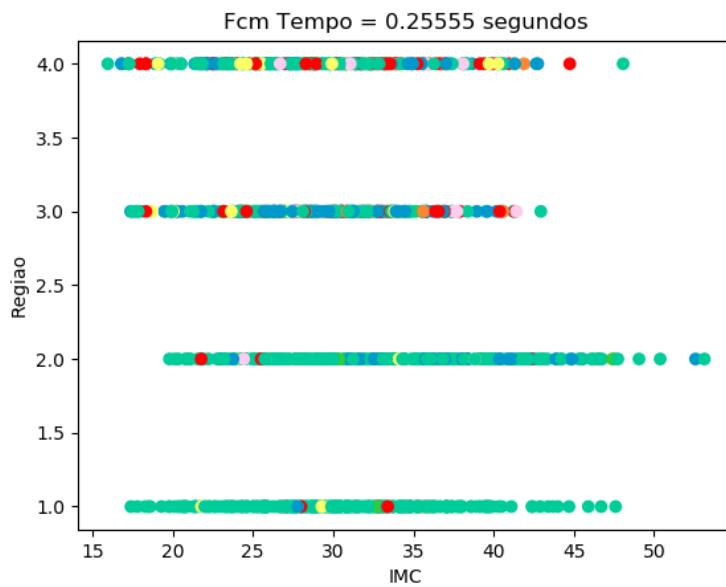


Figura 34 – Agrupamento FCM para Região X IMC do seguro clínico

mais alto de tempo, possivelmente derivado da sua maior complexidade e do acréscimo de tempo para cálculo do raio de vizinhança médio, mas com a vantagem de não necessitar selecionar um número de partições.

### 5.3.2 CONSUMO ONLINE

Na última base de dados trabalhada observa-se pouca diferença entre o número de partições gerados, sendo três grupos o mais bem indicado pela maioria, as únicas exceções são Gustafson Kessel e o Mean Shift que empregam 5 grupos. Observando os coeficientes já é

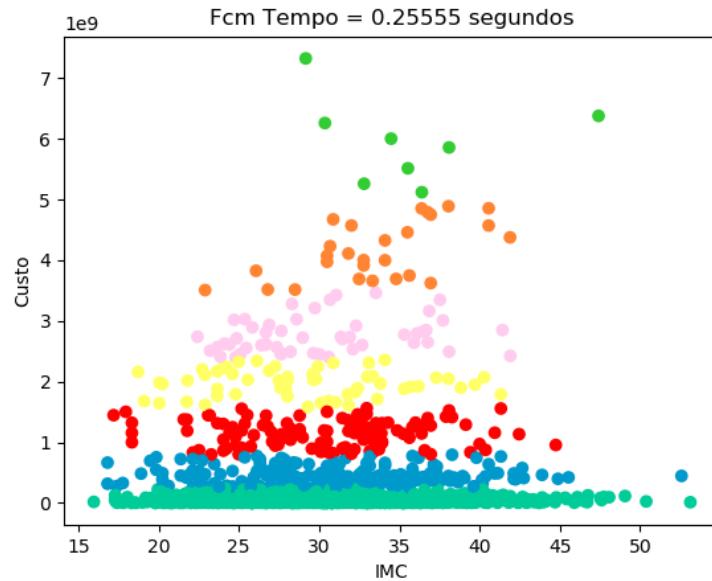


Figura 35 – Agrupamento FCM para Custo X IMC do seguro clínico

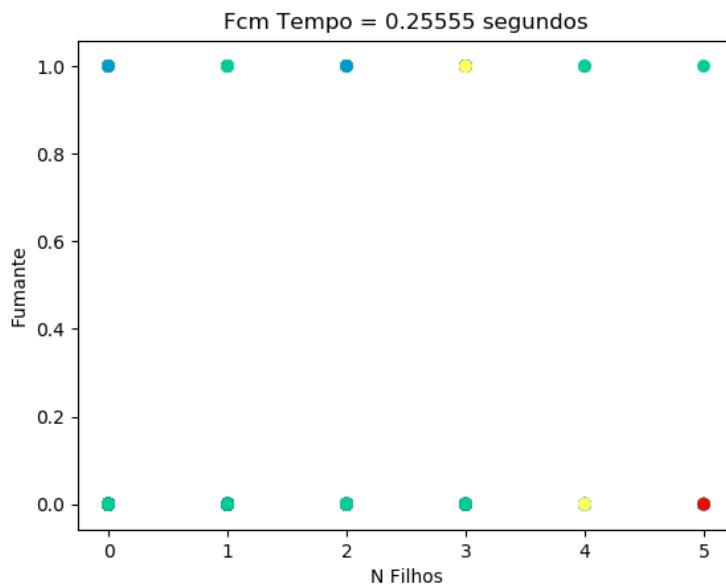


Figura 36 – Agrupamento FCM para Fumante X nº filhos do seguro clínico

possível perceber que os algoritmos C-médias, Possibilístico e K-médias obtiveram os mesmos resultados.

Na Tabela 21 observa-se os destaques dos coeficientes para a base do consumo online. Estruturalmente esta base possui uma quantia maior de dados que a base de dados anterior (Seguro clínico), porém com menor número de variáveis. Os modelo *fuzzy*, C-médias e Possibilístico, continuaram entre os destaques do agrupamento.

Nas Figuras 51 a 56 mostra-se os resultados obtidos pela base K-médias, C-médias e Possibilístico que fizeram o mesmo agrupamento. Assim como o Seguro Clínico, a divisão

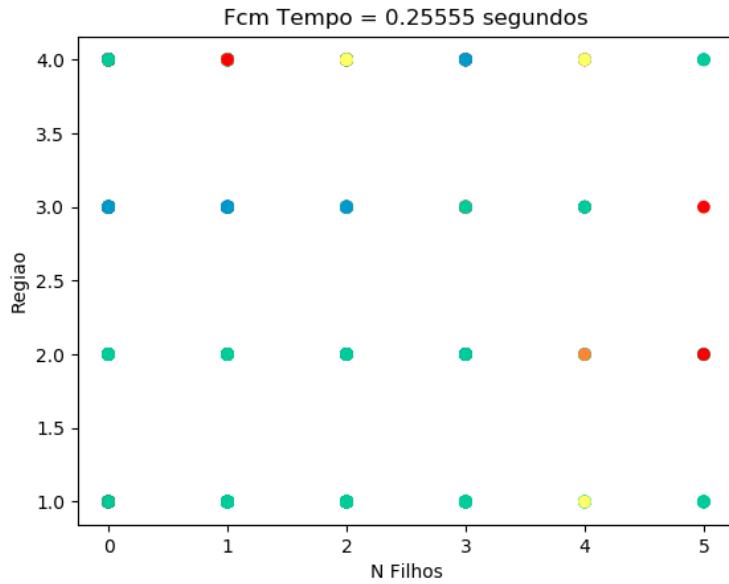


Figura 37 – Agrupamento FCM para Região X nº filhos do seguro clínico

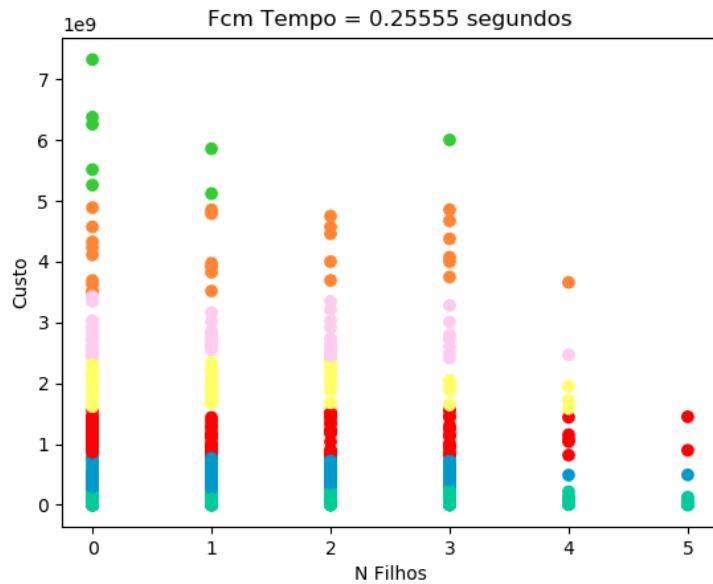


Figura 38 – Agrupamento FCM para Custo X nº filhos do seguro clínico

dessa base sofre maior influência do atributo com maior limite de variação das respostas, no caso os identificadores dos clientes. Observa-se a partir das Figuras 56, 54 e 52 que todas as divisões seguem o eixo do identificador de clientes, levando a essa conclusão. Esse tipo de agrupamento pode ser útil para demonstrar grupos de interesses similares ou outros aspectos sobre determinados clientes.

Assumindo que compras de um mesmo produto em altas quantidades ( $> 100$ ) são geralmente realizadas por empresas para reposição de estoque e adereços, a principal conclusão que pode-ser retirada dos resultados é que a região dos clientes demarcados em azul com identificadores (<

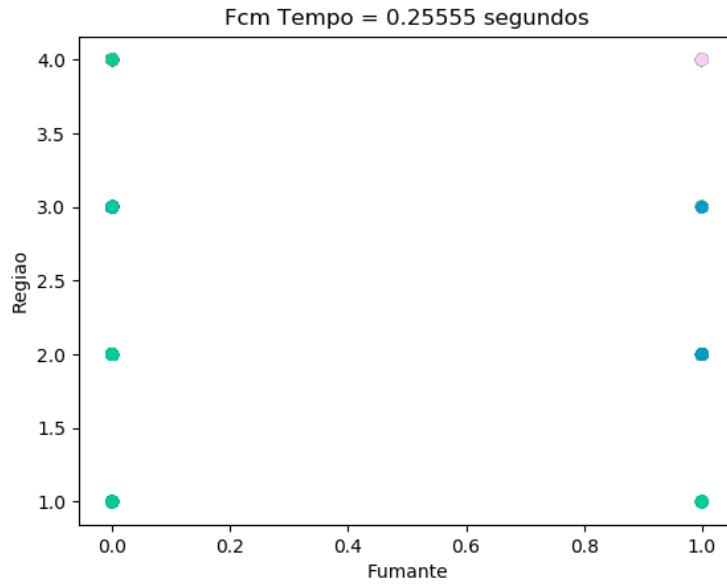


Figura 39 – Agrupamento FCM para Região X Fumante do seguro clínico

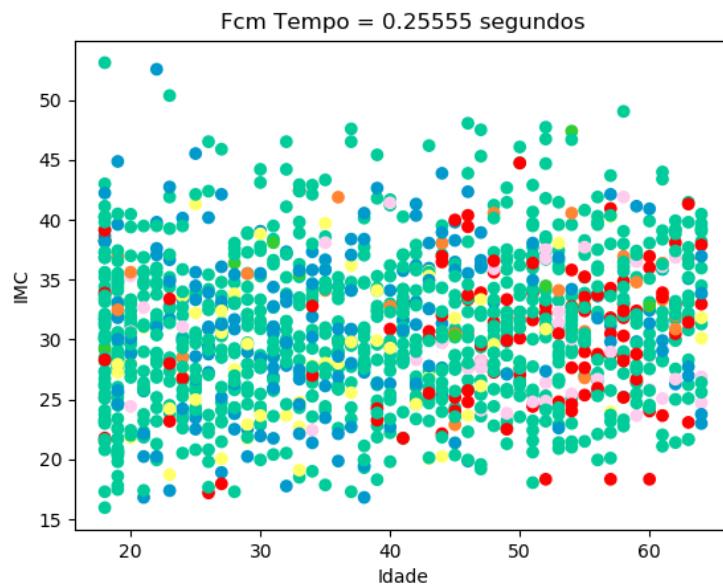


Figura 40 – Agrupamento FCM para IMC X Idade do seguro clínico

13.000) possui uma maior concentração de industrias, sendo também a que fez compras de mais alto custo (Figuras 52 e 54 respectivamente). Seguindo a mesma técnica, ao observar a Figura 53, é possível definir alguns países como mais industrializados que outro, mas sendo essa informação não destacada pelos algoritmos, não existe muitas outras conclusões a se retirar sobre a base.

Para os demais algoritmos, destaca-se novamente a tendência do modelo Gustafson Kessel de receber influências de outros atributos além do maior limite de variação. Na Figura 57 observa-se que o agrupamento não segue mais apenas o eixo do identificador dos clientes.

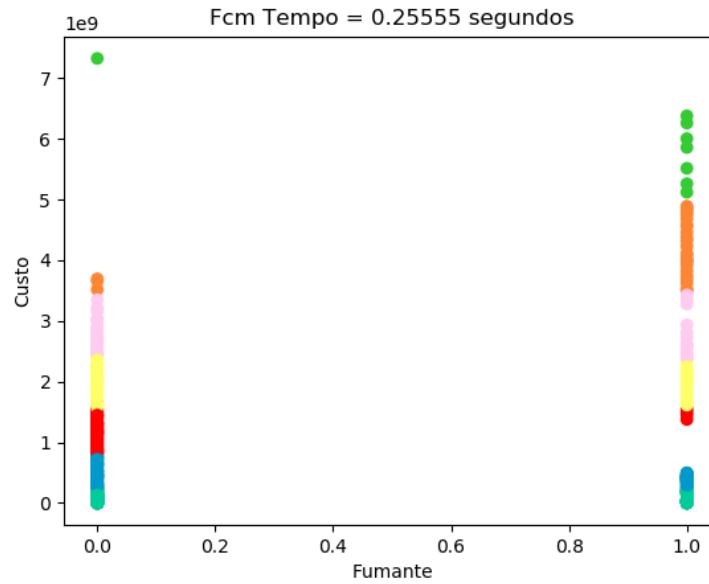


Figura 41 – Agrupamento FCM para Custo X Fumante do seguro clínico

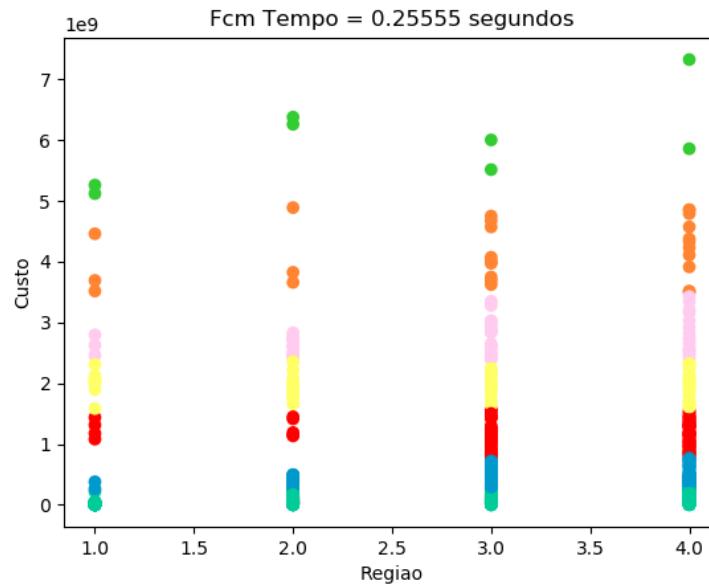


Figura 42 – Agrupamento FCM para Custo X Região do seguro clínico

#### 5.4 CONSIDERAÇÕES FINAIS

Nesta seção é apresentado o resumo da principais observações acerca dos experimentos realizados.

Para os diferentes experimentos destaca-se a tendência dos modelos *fuzzy* gerarem agrupamentos com um maior número de grupos, quando comparados aos agrupamentos gerados pelos modelos clássicos. Para bases com mais dados e maior dispersão essa tendência demonstra maior influência do que nos casos contrários.

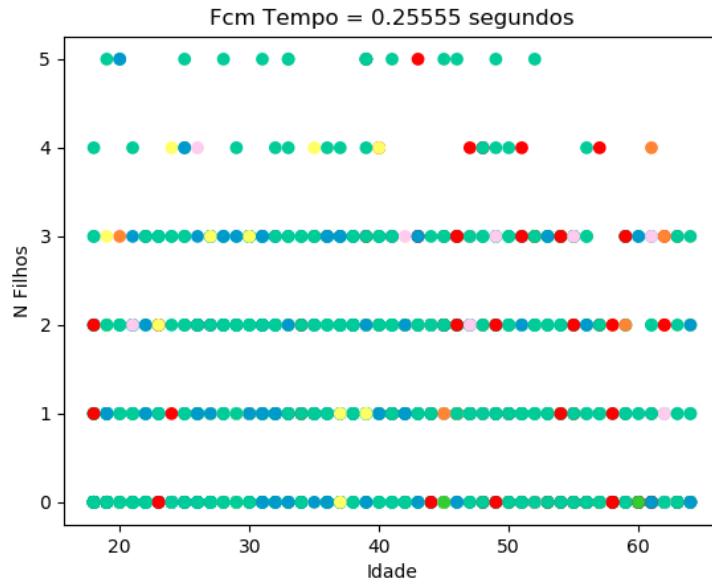


Figura 43 – Agrupamento FCM para nº filhos X Idade do seguro clínico

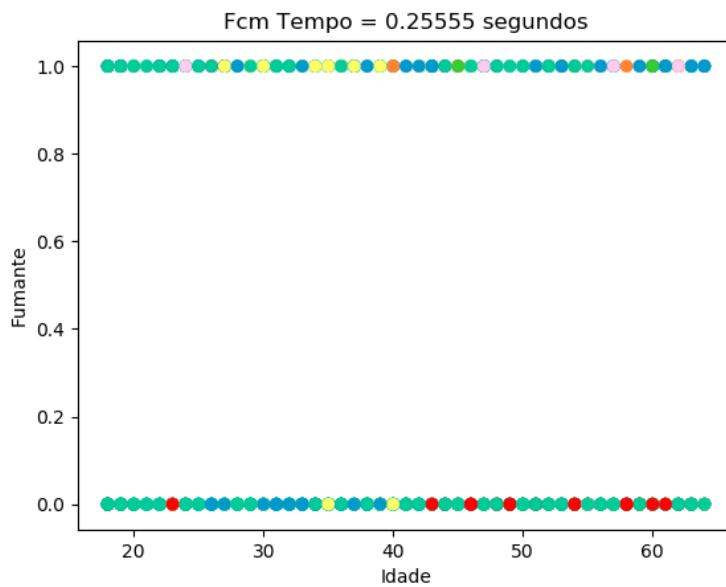


Figura 44 – Agrupamento FCM para Fumante X Idade do seguro clínico

Para os experimentos propostos, os agrupamentos obtidos foram similares para técnicas diferentes em diversos casos. E a análise através dos coeficientes mostrou que os modelos *fuzzy* apresentavam desempenho ruim, mas que os agrupamentos obtidos foram adequados.

Não houve resposta uniforme entre os modelos para indicar que os modelos *fuzzy* ou os modelos clássicos são as melhores técnicas. Respostas ora muito boas, ora ruins, podem indicar que os conjuntos de dados apresentam disposição espacial distinta e, confirmam a dependência da adequação da técnica com a disposição do conjunto de dados.

Ressalta-se ainda outros pontos específicos para cada modelo estudado. O método

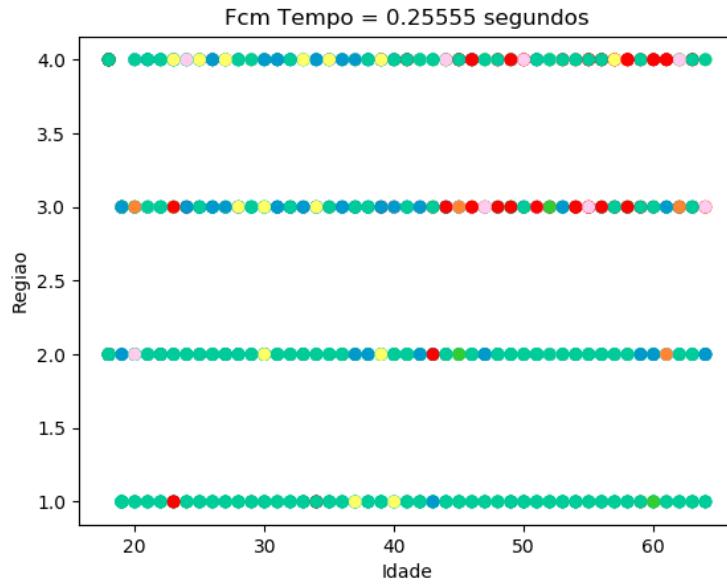


Figura 45 – Agrupamento FCM para Região X Idade do seguro clínico

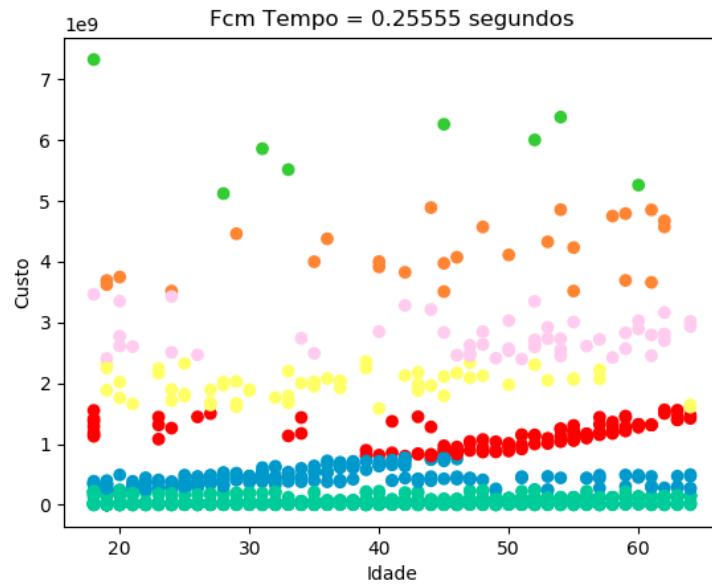


Figura 46 – Agrupamento FCM para Custo X Idade do seguro clínico

aglomerativo se apresentou como o melhor diversas vezes pela análise aos coeficientes e na maioria dos casos obteve também o melhor tempo de execução, porém é o mais incapaz de agrupar bases muito grandes. O Mean Shift foi o método clássico que obteve mais variadas quantidades de grupo, se aproximando muito dos resultados *fuzzy* em várias ocasiões e sendo destaque diversas vezes ao longo da pesquisa, porém a decisão de um raio de vizinhança se mostrou complexa e a escolha matemática exigiu um grande aumento do tempo de execução. O algoritmo Gustafson Kessel demonstrou que com seu método diferente de cálculo de similaridade tem maiores tendências a sofrer influência de todos os atributos de uma base de dados, sendo esse

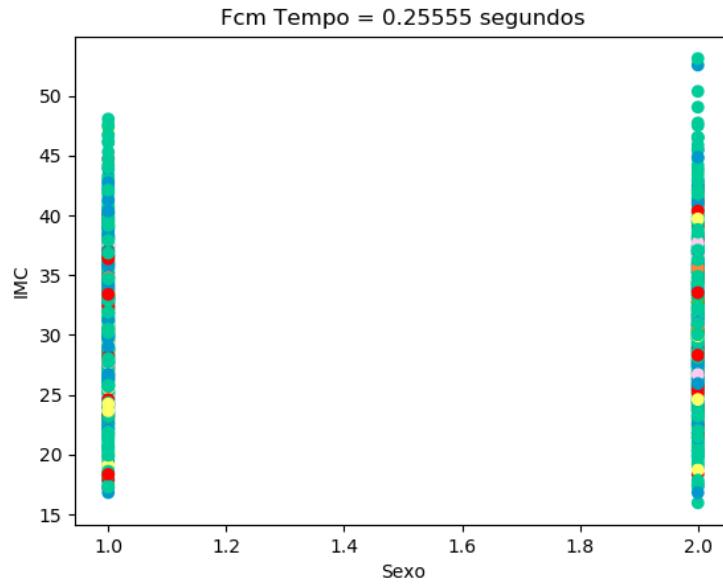


Figura 47 – Agrupamento FCM para IMC X Gênero do seguro clínico

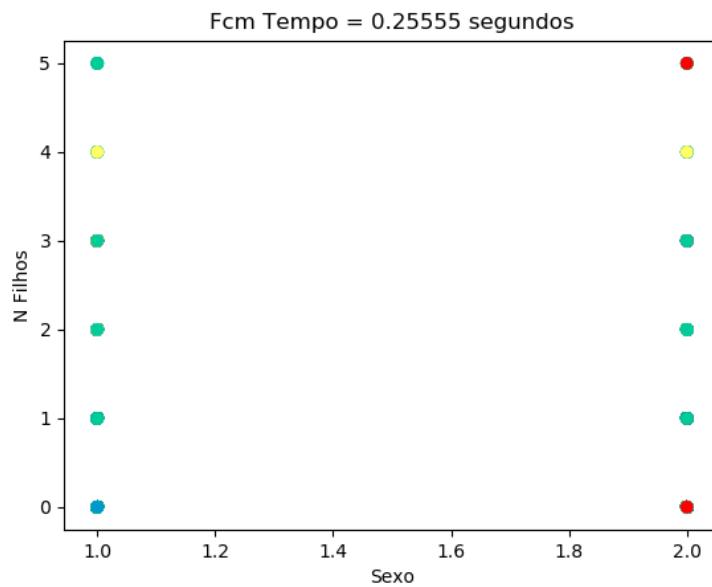


Figura 48 – Agrupamento FCM para nº filhos X gênero do seguro clínico

outro grande ponto para novos estudos. O modelo Possibilístico apresentou diversos resultados com o melhor tempo de execução.

## 5.5 ENSEMBLE POR VOTAÇÃO

A última fase de pesquisa foi a aplicação do modelo de *ensemble* por votação a partir dos algoritmos clássicos e *fuzzy*. Nessa etapa são mostrados os melhores resultados alcançados, análises e principais conclusões a partir dos combinadores.

A combinação de modelos (*ensemble*) é apontada na literatura como um recurso que

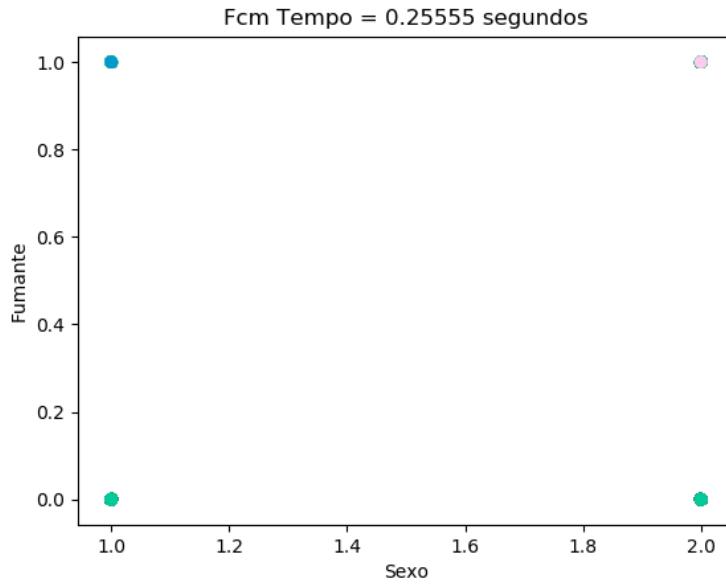


Figura 49 – Agrupamento FCM para Fumante X Gênero do seguro clínico

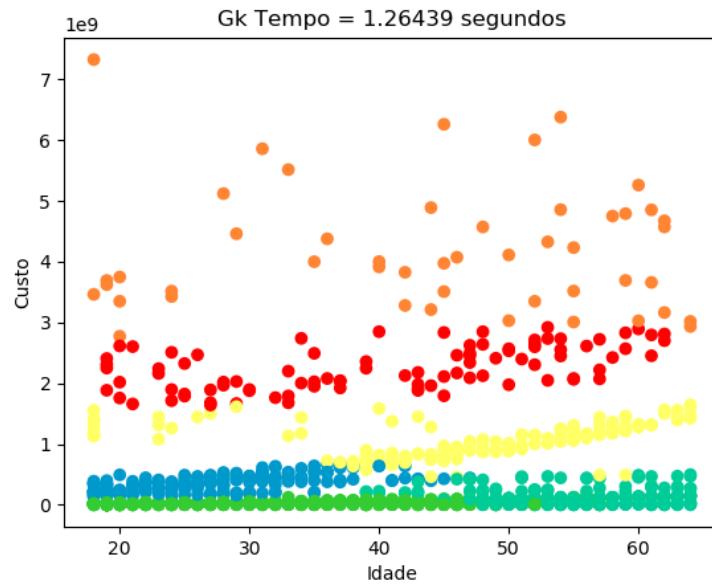


Figura 50 – Resultado gráfico representativo do método de Gustafson Kessel agrupando a base de seguros clínicos segundo idade x custos

pode auxiliar na melhoria de um modelo. O objetivo desta fase é investigar o comportamento dos combinadores para as técnicas estudadas.

### 5.5.1 BASES DE DADOS DE VALIDAÇÃO

A partir das bases de dados sintéticas os resultados obtidos podem ser observados nos gráficos das Figuras 58 e 59. É importante ressaltar que a análise do número de partições se tornou muito mais complexa nesse estágio, principalmente para os modelos *fuzzy*, sendo um grande exemplo o algoritmo probabilístico, cuja a discrepância que pode existir entre valores,

Tabela 21 – Resultados das métricas para comparação relativa com a base de dados 'Consumo Online'

<i>Coeficientes \ Bases</i>	<i>k-médias</i>	<i>Aglomerativo</i>	<i>MeanShift</i>
Silhouette	<b>0,87236</b>	0,87197	0,83326
Calinski	<b>56403,20845</b>	55846,03904	25283,02104
Davies	<b>0,23572</b>	0,23343	0,27707
Dunn	0,08271	0,11115	<b>0,13393</b>
<i>Coeficientes \ Bases</i>	<i>c-médias</i>	<i>Gustafson</i>	<i>possibilístico</i>
Silhouette	<b>0,87236</b>	0,24054	<b>0,87236</b>
Calinski	<b>56403,20845</b>	2451,27782	<b>56403,20845</b>
Davies	<b>0,23572</b>	5,54653	<b>0,23572</b>
Dunn	0,08271	4,00E-05	0,08271
Part. Coef.	<b>0,96396</b>	0,68675	0,96001
Hypervolume	28012351,25	<b>2151113,536</b>	54914330,06
Fukuyama	<b>-14837542703</b>	-5171275322	-14422315961
Xie Beni	<b>0,01183</b>	6,44983	0,01227

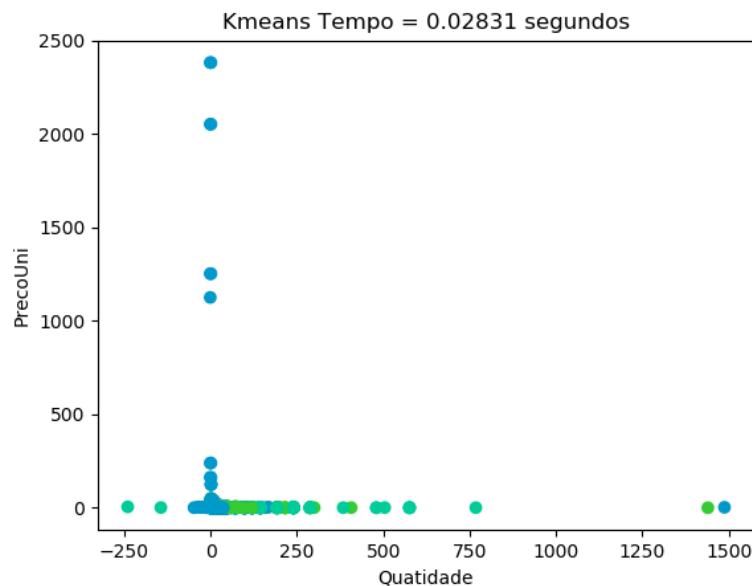


Figura 51 – Agrupamento K-médias para Preço X Quantidade da base de consumos online

mesmo com uma normalização, é muito alta e afeta na análise dos gráficos.

Inicialmente, observa-se o efeito, antes visto das métricas *fuzzy*, que tendem a indicar um número maior de clusters para o agrupamento, não foi demonstrado para esses casos. Tanto para os modelos clássicos quanto *fuzzy* os resultados alcançados não foram melhores do que aqueles obtidos para os modelos independentes. Ainda assim, alguns pontos são ressaltados a seguir.

Comparando os resultados obtidos para as combinações foi possível observar que o combinador *fuzzy* responde melhor para a base "ruídos" do que para a base "tríade", indicando dificuldade do modelo para este formato de dado. Os resultados alcançados ainda sugerem que os combinadores trabalham melhor com grupos simétricos, horizontais e verticais.

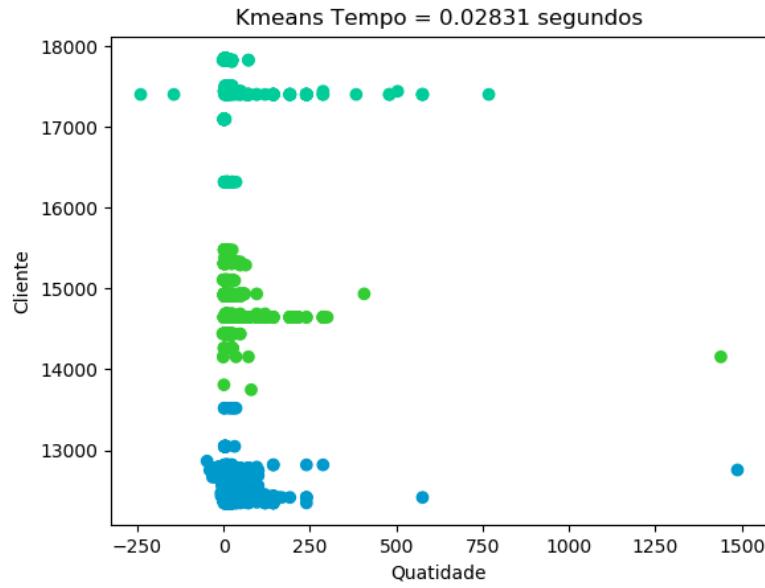


Figura 52 – Agrupamento K-médias para Cliente X Quantidade da base de consumos online

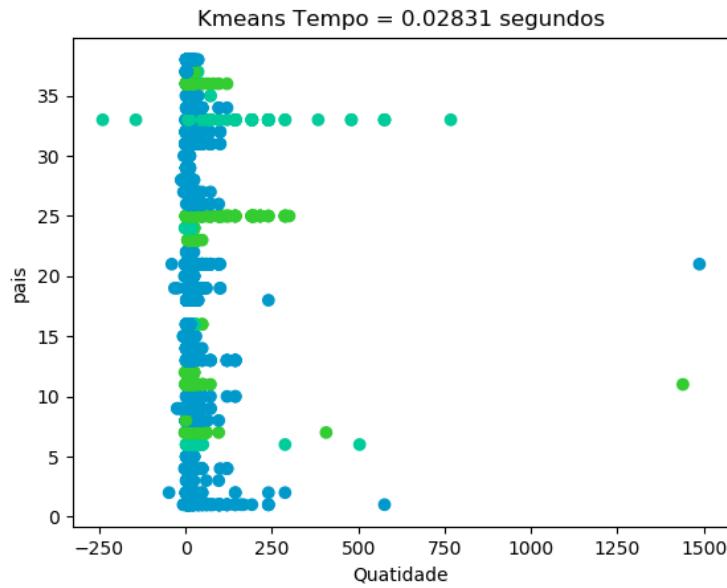


Figura 53 – Agrupamento K-médias para País X Quantidade da base de consumos online

### 5.5.2 BASES DE DADOS CLÁSSICAS

Para as bases de dados clássicas o *ensemble* também não alcançou melhor resultados do que os algoritmos isolados, diferindo principalmente no número de grupos selecionados para o agrupamento. É importante notar que o número de grupos obtidos para cada modelo base é diferente, sendo assim poderia ser interessante investigar o emprego do ensemble aceitando um número diferente de grupos para cada modelo.

As Tabelas 22, 23 e 24 apresentam os resultados dos coeficientes para cada base de dados.

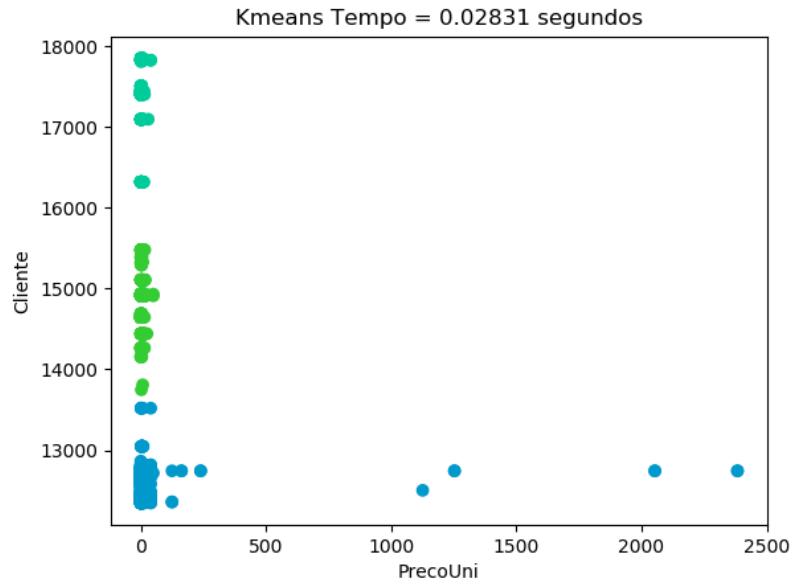


Figura 54 – Agrupamento K-médias para Cliente X Preço da base de consumos online

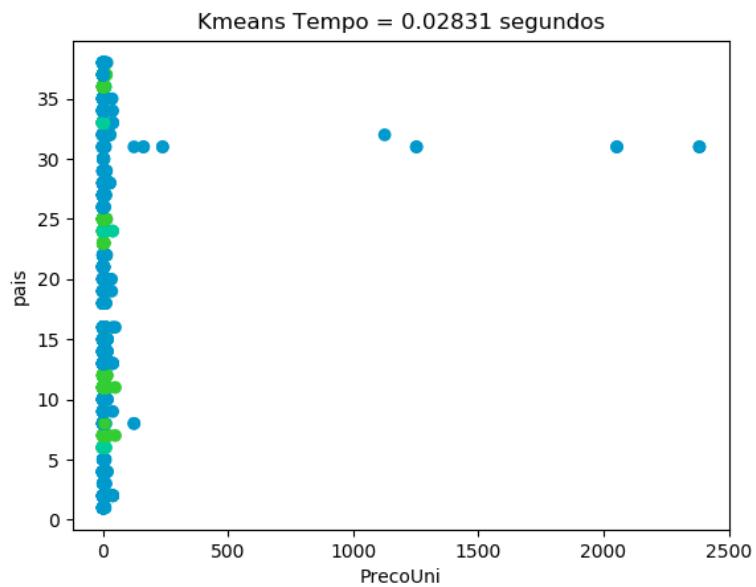


Figura 55 – Agrupamento K-médias para País X Preço da base de consumos online

Tabela 22 – Resultados dos coeficientes do ensemble da base de Íris

<i>Coeicientes \ Bases</i>	<i>Clássico</i>	<i>fuzzy</i>
Silhouette	0,68105	0,68674
Calinski	513,92455	502,82156
Davies	0,40429	0,38275
Dunn	0,07651	0,33891

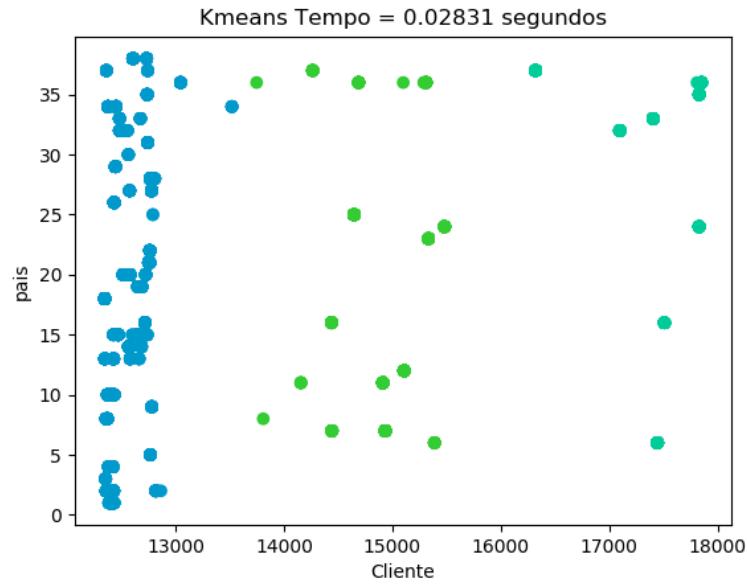


Figura 56 – Agrupamento K-médias para País X Cliente da base de consumos online

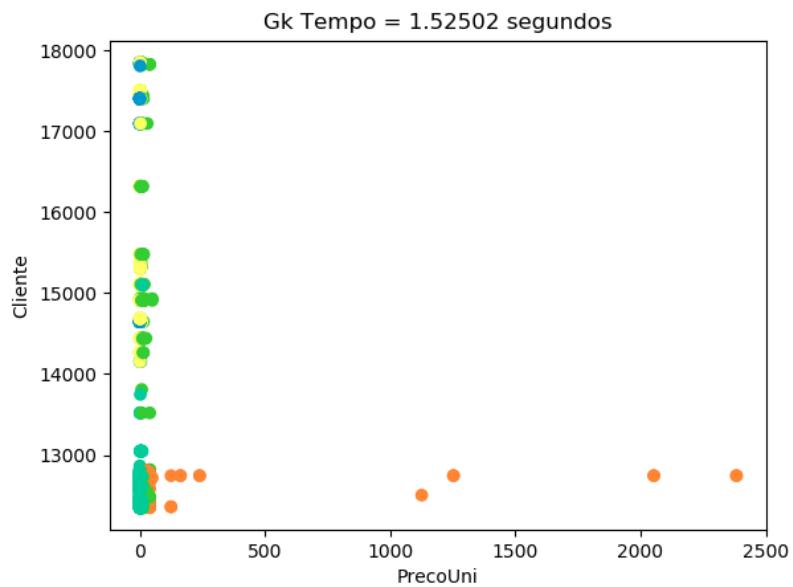


Figura 57 – Agrupamento Gustafson Kessel para Cliente X Preço da base de consumos online

Tabela 23 – Resultados dos coeficientes do ensemble da base de Wine

<i>Coeicientes \ Bases</i>	<i>Clássico</i>	<i>fuzzy</i>
Silhouette	0,65685	0,50088
Calinski	505,42931	760,21714
Davies	0,47878	0,55676
Dunn	0,02282	0,02087

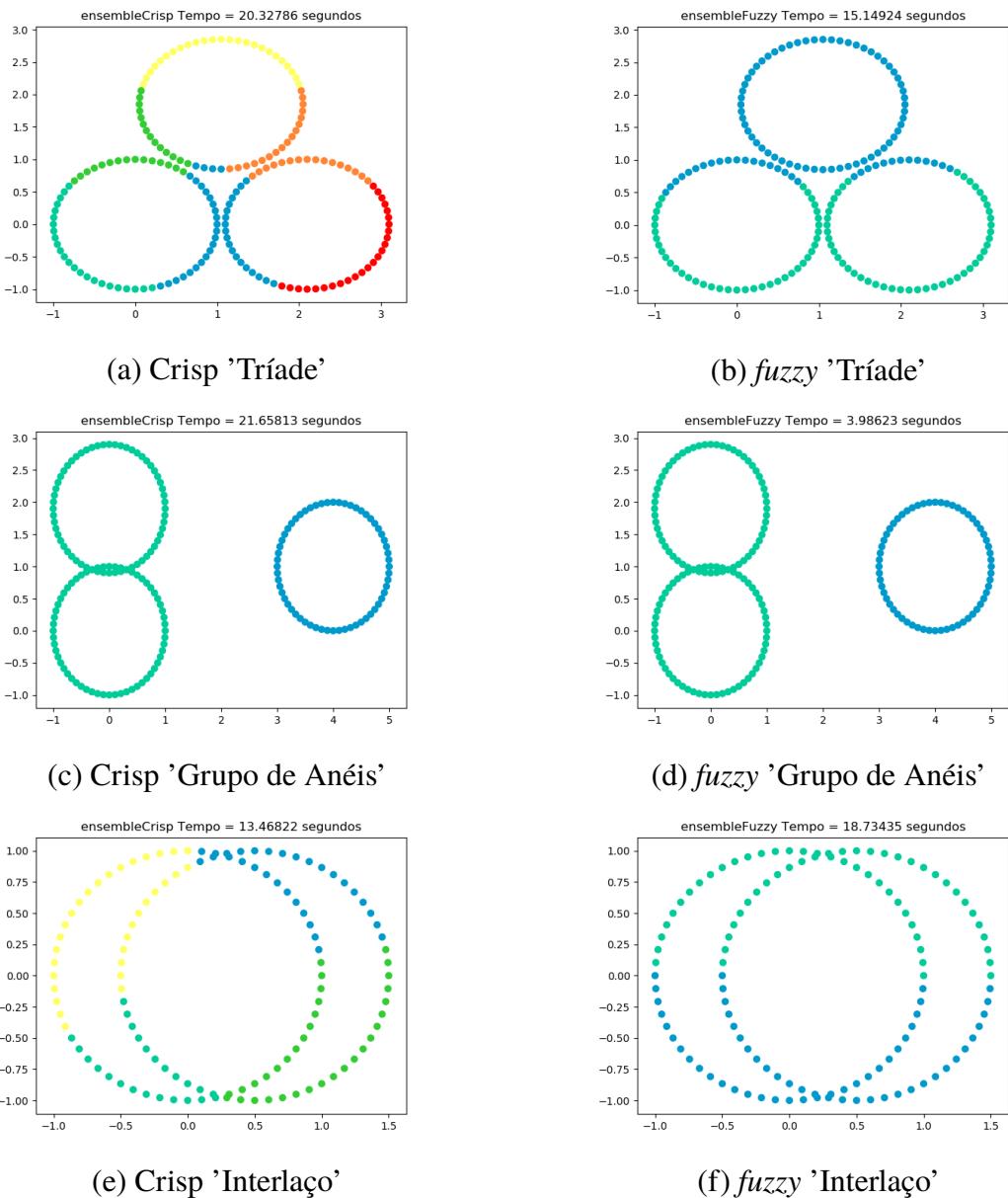


Figura 58 – Agrupamento das bases Tríade, Grupo de Anéis e Interlaço, realizado pelas técnicas de ensemble.

Tabela 24 – Resultados dos coeficientes do ensemble da base de Boston

Coefficientes \ Bases	Clássico	fuzzy
Silhouette	0,6914	0,44682
Calinski	1198,8304	1294,00106
Davies	0,54356	0,87351
Dunn	0,52481	0,02537

Observa-se ainda o aumento do tempo de execução para o *ensemble* a partir dos modelos *fuzzy*. Ao gerar os gráficos para descoberta de uma boa configuração inicial para o modelo, percebeu que a exigência de múltiplas execuções do algoritmo resultou em um grande aumento do tempo de processamento dos modelos *fuzzy*, tornando essa técnica muito menos viável para

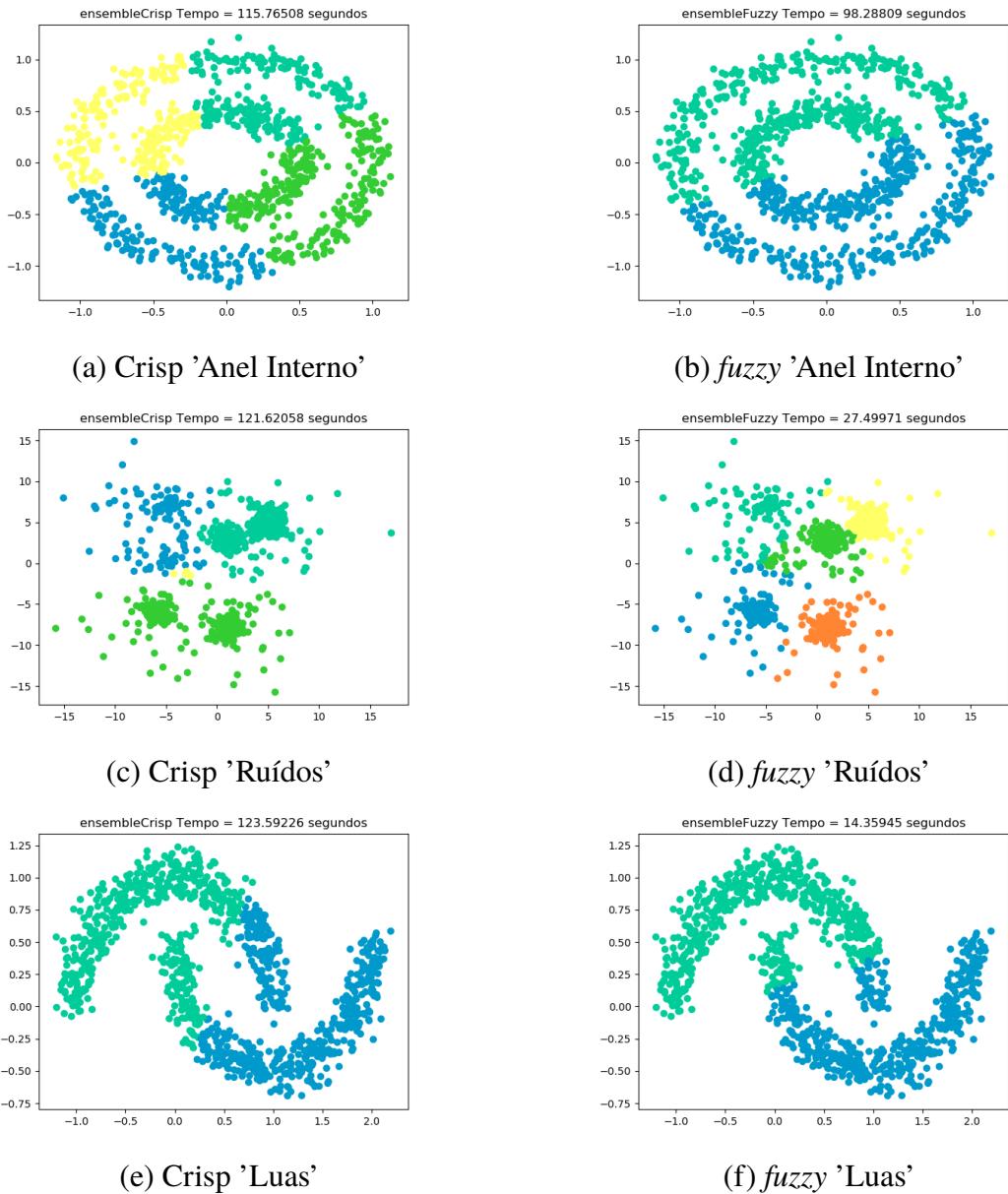


Figura 59 – Agrupamento das bases Anel Interno, Ruídos e Luas, realizado pelas técnicas de ensemble.

esse tipo de problema.

### 5.5.3 OUTRAS BASES DE DADOS

Nas bases de dados selecionadas aleatoriamente na internet, Seguro Clínico e Consumo online, a combinação foi extremamente lenta em função do tamanho dos conjuntos de dados. Além disso, não foi possível executar os testes com o modelo *fuzzy*, dado que a quantidade de valores gerava um erro no cálculo do determinante para as equações do método Possibilístico.

Em resumo, o ensemble não se apresentou adequado para o agrupamento dos conjuntos selecionados nesta pesquisa, quando comparados a suas versões isoladas. Além disso, como esta pesquisa não teve como foco a combinação de modelos, a metodologia empregada na definição dos experimentos isolados pode não ser adequada para testar *ensembles*, contribuindo para os resultados pouco significativos que foram encontrados nesta etapa.

## 6 CONCLUSÃO

Esta pesquisa apresentou um estudo de modelos de agrupamento clássicos em comparação com modelos baseados na lógica *fuzzy*. Foi proposto um conjunto de experimentos empregando bases de dados com características diversas buscando entender o quanto o emprego da lógica *fuzzy* na técnica auxilia na obtenção de um resultado adequado.

A discussão do emprego de modelos de aprendizado de máquina adequados vai de encontro à natureza dos dados. Uma técnica não é melhor, mas pode ser a mais adequada para uma aplicação específica, e sua adequação se deve a fatores como: tamanho do conjunto de dados (tanto em número de elementos quanto em número de características), a disposição dos dados no espaço e a interpretação do agrupamento obtido.

Neste trabalho todas as análises realizadas foram baseadas em um conjunto de coeficientes de avaliação de agrupamentos. A partir dos experimentos com bases de dados sintéticas, dada a simplicidade dos dados, foi possível a observação dos agrupamentos alcançados através de gráficos, constituindo uma etapa de validação. Além deste, dois outros experimentos com bases de dados clássicas e reais foram executados e analisados.

Considerando os modelos *fuzzy* duas características foram identificadas: geração de agrupamentos com maior número de grupos, cujos resultados se assemelham àqueles obtidos pelas modelos clássicos. E ainda, obtenção de grupos com mesmo formato.

Os coeficientes de validação interna empregados para análise dos modelos *fuzzy* se apresentaram menos eficientes para encontrar as configurações iniciais dos modelos, quando comparado ao conjunto dos coeficientes empregos para os modelos clássicos.

Com relação ao tempo de processamento, os modelos *fuzzy* apresentaram durante os experimentos resultados tão bons quanto a versão clássica. Contudo mostraram que para problemas que necessitam de várias execuções dos modelos, como o *ensemble*, possuem um grande aumento nesse tempo necessário para o processamento.

No geral, os demais pontos levantados não apontam tantas evidências a vantagens ou desvantagens ao emprego da lógica *fuzzy*, o que leva a necessidade de estudos em outras frentes, que possam demonstrar se há uma melhor aproximação dos modelos *fuzzy* com a interpretação humana ou, se estes modelos podem trabalhar melhor com dados dispostos em certas formas geométricas. Em outros casos, as particularidades observadas foram provenientes dos algoritmos em si e não pareciam ter ligação com o emprego ou não da teoria de conjuntos *fuzzy*. Ainda assim, a pesquisa mostrou diversos destaques descritos sobre cada algoritmo avaliado, como a maior capacidade de influências das informações sobre a divisão do GK, a capacidade e adaptação do Mean Shift, a velocidade e eficiência dos métodos aglomerativo e probabilístico, e ainda a robustez dos métodos k-médias e c-médias como algoritmos base.

Como trabalhos futuros ficam destacados: a necessidade de uma pesquisa aprofundada sobre a influência da informação sobre o GK em casos reais; os limites de cálculos para o método probabilístico; a aplicação dos modelos clássicos e *fuzzy* aqui estudados para agrupamento em imagens ou Big-data; a influência do formato gráfico dos dados sobre o agrupamento; e ainda um trabalho sobre a otimização *fuzzy* com conjuntos de dados que já possuem classificação prévia.

## REFERÊNCIAS BIBLIOGRÁFICAS

- 99991, J. W. J. S. scikit-fuzzy; twmeggs; alexsavio; Aishwarya Unnikrishnan; Guilherme Castelão; Felipe Arruda Pontes; Tobias Uelwer; pd2f; laurazh; Fernando Batista; alexbuy; Wouter Van den B. W. S. T. G. B. R. A. M. P. J. F. P. H. M. G. O. T. A. H. *JDWarner/scikit-fuzzy: Scikit-Fuzzy version 0.4.2.* [S.l.]: GitHub, 2019. Citado na página 10.
- Babuka, R.; van der Veen, P. J.; Kaymak, U. Improved covariance estimation for gustafson-kessel clustering. In: *2002 IEEE World Congress on Computational Intelligence. 2002 IEEE International Conference on Fuzzy Systems. FUZZ-IEEE'02. Proceedings (Cat. No.02CH37291)*. [S.l.: s.n.], 2002. v. 2, p. 1081–1085 vol.2. ISSN null. Citado na página 21.
- BUITINCK, L. et al. API design for machine learning software: experiences from the scikit-learn project. In: *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*. [S.l.: s.n.], 2013. p. 108–122. Citado 7 vezes nas páginas 3, 5, 10, 18, 19, 24 e 50.
- CHARRAD, M. et al. Nbclust: An r package for determining the relevant number of clusters in a data set. *Jurnal of Statistical Software*, 2014. Disponível em: <<https://www.jstatsoft.org/article/view/v061i06>>. Citado 2 vezes nas páginas 22 e 28.
- Comaniciu, D.; Meer, P. Mean shift analysis and applications. In: *Proceedings of the Seventh IEEE International Conference on Computer Vision*. [S.l.: s.n.], 1999. v. 2, p. 1197–1203 vol.2. ISSN null. Citado na página 19.
- Comaniciu, D.; Meer, P. Mean shift: a robust approach toward feature space analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 24, n. 5, p. 603–619, May 2002. ISSN 1939-3539. Citado na página 18.
- COPPIN, B. *Inteligência Artificial (Em Portuguese do Brasil)*. LTC, 2010. ISBN 8521617291. Disponível em: <<https://www.amazon.com/Intelig%C3%A1ncia-Artificial-Em-Portuguese-Brasil/dp/8521617291?SubscriptionId=AKIAIOINVZYXZQZ2U3A&tag=chimbori05-20&linkCode=xm2&camp=2025&creative=165953&creativeASIN=8521617291>>. Citado na página 14.
- DAVE, R. N. Validating fuzzy partitions obtained through c-shells clustering. *Pattern Recognition Letters*, v. 17, n. 6, p. 613 – 623, 1996. ISSN 0167-8655. Disponível em: <<http://www.sciencedirect.com/science/article/pii/0167865596000268>>. Citado 2 vezes nas páginas 11 e 29.
- Davies, D. L.; Bouldin, D. W. A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-1, n. 2, p. 224–227, 1979. Citado na página 22.
- DEMBELE, D.; KASTNER, P. Fuzzy c-means method for clustering microarray data. *bioinformatics*, Oxford University Press, v. 19, n. 8, p. 973–980, 2003. Citado na página 10.
- DEVASENA, C. L. et al. Effectiveness evaluation of rule based classifiers for the classification of iris data set. *Bonfring International Journal of Man Machine Interface, Vol. 1, Special Issue, December 2011*, v. 1, p. 05–09, 2011. Disponível em: <<http://www.journal.bonfring.org/abstract.php?id=4&archiveid=103>>. Citado na página 31.

- DONI, M. V.; OLIVEIRA, R. AnÁlise de cluster: MÉtodos hierárquicos e de particionamento. *Universidade Presbiteriana Mackenzie*, 2004. Disponível em: <<http://meusite.mackenzie.com.br/rogerio/tgi/2004Cluster.PDF>>. Citado na página 18.
- DRINEAS, P. et al. Clustering large graphs via the singular value decomposition: Theoretical advances in data clustering (guest editors: Nina mishra and rajeev motwani). *Machine Learning*, v. 56, 01 2004. Citado na página 19.
- DUA, D.; GRAFF, C. *UCI Machine Learning Repository*. 2017. Disponível em: <<http://archive.ics.uci.edu/ml>>. Citado na página 31.
- EL-DIN, A. N. J.; ALJABASINI, O. *fuzzy-clustering*. [S.l.]: GitHub, 2018. <<https://github.com/ITE-5th/fuzzy-clustering>>. Citado na página 26.
- FACELI, K. *Inteligência Artificial. Uma Abordagem de Aprendizado de Máquina (Em Portuguese do Brasil)*. LTC, 2011. ISBN 8521618808. Disponível em: <<https://www.amazon.com/Intelig%C3%A1ncia-Artificial-Abordagem-Aprendizado-Portuguese/dp/8521618808?SubscriptionId=AKIAIOINVZYXZQZ2U3A&tag=chimbori05-20&linkCode=xm2&camp=2025&creative=165953&creativeASIN=8521618808>>. Citado 3 vezes nas páginas 14, 15 e 24.
- FARBER, I. et al. On using class-labels in evaluation of clusterings. In: *MultiClust: 1st International Workshop on Discovering, Summarizing and Using Multiple Clusterings Held in Conjunction with KDD 2010, Washington, DC*. [s.n.], 2010. Disponível em: <[http://scholar.google.com.au/scholar.bib?q=info:d8mnI5RN02oJ:scholar.google.com/&output=citation&hl=en&as\\_sdt=0,5&ct=citation&cd=0](http://scholar.google.com.au/scholar.bib?q=info:d8mnI5RN02oJ:scholar.google.com/&output=citation&hl=en&as_sdt=0,5&ct=citation&cd=0)>. Citado na página 22.
- FIRMANSYAH, A. F. B.; PRAMANA, S. Ensemble based gustafson kessel fuzzy clustering. Vol 1 No 1 (2018): *Journal of Data Science and Its Applications*, 2018. Disponível em: <<https://commidis.telkomuniversity.ac.id/jdsa/index.php/jdsa/article/view/6>>. Citado na página 20.
- FISHER, R. A. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, v. 7, n. 2, p. 179–188, 1936. Disponível em: <<https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1469-1809.1936.tb02137.x>>. Citado na página 31.
- FOWLKES, E. B.; MALLOWS, C. L. A method for comparing two hierarchical clusterings. *Journal of the American Statistical Association*, Taylor & Francis, v. 78, n. 383, p. 553–569, 1983. Disponível em: <<https://amstat.tandfonline.com/doi/abs/10.1080/01621459.1983.10478008>>. Citado na página 22.
- GERMER, T. *FuzzyClustering*. [S.l.]: GitHub, 2017. <<https://github.com/99991/FuzzyClustering>>. Citado 2 vezes nas páginas 5 e 23.
- HARRISON, J. D.; RUBINFELD, D. L. Hedonic housing prices and the demand for clean air. *Journal of Environmental Economics and Management*, v. 5(1), p. 81–102, 1978. Disponível em: <<http://hdl.handle.net/2027.42/22636>>. Citado na página 31.
- JAIN, A. K. Data clustering: 50 years beyond k-means. *Pattern Recognition Letters*, v. 31, n. 8, p. 651 – 666, 2010. ISSN 0167-8655. Award winning papers from the 19th International Conference on Pattern Recognition (ICPR). Disponível em: <<http://www.sciencedirect.com/science/article/pii/S0167865509002323>>. Citado 2 vezes nas páginas 16 e 19.

- KOUTROUMBAS, K.; THEODORIDIS, S. *Pattern Recognition*. Academic Press, 2008. ISBN 9780080949123. Disponível em: <<https://www.amazon.com/Pattern-Recognition-Konstantinos-Koutroumbas-ebook/dp/B003FQM374?SubscriptionId=AKIAIOBINVZYXZQZ2U3A&tag=chimbori05-20&linkCode=xm2&camp=2025&creative=165953&creativeASIN=B003FQM374>>. Citado 2 vezes nas páginas 22 e 28.
- K.SASIREKHA; P.BABY. Agglomerative hierarchical clustering algorithm- a review. *International Journal of Scientific and Research Publications (IJSRP)*, v. 3, 2013. Disponível em: <<http://www.ijsrp.org/research-paper-0313.php?rp=P15831>>. Citado na página 17.
- KUHN, H. W. The hungarian method for the assignment problem. *Naval research logistics quarterly*, Wiley Online Library, v. 2, n. 1-2, p. 83–97, 1955. Citado na página 34.
- LI, C. sheng. The improved partition coefficient. *Procedia Engineering*, v. 24, p. 534 – 538, 2011. ISSN 1877-7058. International Conference on Advances in Engineering 2011. Disponível em: <<http://www.sciencedirect.com/science/article/pii/S1877705811055433>>. Citado na página 36.
- MALHOTRA, V. K.; KAUR, H.; ALAM, M. A. Article: An analysis of fuzzy clustering methods. *International Journal of Computer Applications*, v. 94, n. 19, p. 9–12, May 2014. Full text available. Citado 2 vezes nas páginas 10 e 20.
- MANY. *Advances in Fuzzy Clustering and its Applications*. Wiley, 2007. ISBN 0470027606. Disponível em: <<https://www.amazon.com/Advances-Fuzzy-Clustering-its-Applications/dp/0470027606?SubscriptionId=AKIAIOBINVZYXZQZ2U3A&tag=chimbori05-20&linkCode=xm2&camp=2025&creative=165953&creativeASIN=0470027606>>. Citado na página 10.
- MONARD, M. C.; BARANAUSKAS, J. A. Conceitos sobre aprendizado de máquina. In: *Sistemas Inteligentes Fundamentos e Aplicações*. 1. ed. Barueri-SP: Manole Ltda, 2003. p. 89–114. ISBN 85-204-168. Citado na página 14.
- MONTEIRO, R. B. Comparação de técnicas de aprendizado de máquina para predição da disponibilidade de bicicletas no projeto biciletar fortaleza. *Universidade Federal do Ceará, Campus de Quixadá*, 2018. Disponível em: <<http://www.repository.ufc.br/handle/riufc/34716>>. Citado 2 vezes nas páginas 3 e 15.
- Pakhira, M. K. A linear time-complexity k-means algorithm using cluster shifting. In: *2014 International Conference on Computational Intelligence and Communication Networks*. [S.l.: s.n.], 2014. p. 1047–1051. ISSN null. Citado na página 17.
- Pal, N. R.; Bezdek, J. C. On cluster validity for the fuzzy c-means model. *IEEE Transactions on Fuzzy Systems*, v. 3, n. 3, p. 370–379, Aug 1995. ISSN 1941-0034. Citado na página 19.
- PIECH, C. *K Means*. 2013. Disponível em: <<https://stanford.edu/~cpiech/cs221/handouts/kmeans.html>>. Citado 2 vezes nas páginas 3 e 17.
- ROSENBERG, A.; HIRSCHBERG, J. V-measure: A conditional entropy-based external cluster evaluation measure. In: *Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL)*. [S.l.: s.n.], 2007. p. 410–420. Citado na página 22.
- ROSSUM, G. V.; JR, F. L. D. *Python reference manual*. [S.l.]: Centrum voor Wiskunde en Informatica Amsterdam, 1995. Citado na página 10.

ROUSSEEUW, P. J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, Elsevier, v. 20, p. 53–65, 1987. Citado na página 22.

SKINNER, H. *PossibilisticCMeans*. [S.I.]: GitHub, 2018. <<https://github.com/holtskinner/PossibilisticCMeans>>. Citado na página 26.

Takumi, S.; Miyamoto, S. Top-down vs bottom-up methods of linkage for asymmetric agglomerative hierarchical clustering. In: *2012 IEEE International Conference on Granular Computing*. [S.I.: s.n.], 2012. p. 459–464. ISSN null. Citado na página 17.

YANG, K. *Ensemble-Clustering*. [S.I.]: GitHub, 2016. <<https://github.com/Kr4t0n/Ensemble-Clustering>>. Citado na página 27.

ZADEH, L. Fuzzy sets. *Information and Control*, v. 8, n. 3, p. 338 – 353, 1965. ISSN 0019-9958. Disponível em: <<http://www.sciencedirect.com/science/article/pii/S001999586590241X>>. Citado na página 13.

Thiago Silva Pereira