
Universidade Federal do Piauí
Centro de Ciências da Natureza
Programa de Pós-Graduação em Ciência da Computação

# Rotulação com Algoritmos Supervisionados

**Tarcísio Franco Jaime**

**Número de Ordem PPGCC: M001**

**Teresina-PI, Janeiro de 2017**

Tarcísio Franco Jaime

# Rotulação com Algoritmos Supervisionados

**Qualificação de Mestrado** apresentada ao Programa de Pós-Graduação em Ciência da Computação da UFPI (área de concentração: Sistemas de Computação), como parte dos requisitos necessários para a obtenção do Título de Mestre em Ciência da Computação.

Universidade Federal do Piauí – UFPI

Centro de Ciências da Natureza

Programa de Pós-Graduação em Ciência da Computação

Orientador: Vinicius Ponte Machado

Teresina-PI

Janeiro de 2017

Tarcísio Franco Jaime
Rotulação com Algoritmos Supervisionados/ Tarcísio Franco Jaime. – Teresina-PI, Janeiro de 2017-
40 p. : il. (algumas color.) ; 30 cm.

Orientador: Vinicius Ponte Machado

Qualificação (Mestrado) – Universidade Federal do Piauí – UFPI
Centro de Ciências da Natureza
Programa de Pós-Graduação em Ciência da Computação, Janeiro de 2017.

1. Rotulação. 2. Algoritmos Supervisionados. 3. CART. 4. Naive Bayes. I. Vinicius Ponte Machado. II. Universidade Federal do Piauí. III. Rotulação com Algoritmos Supervisionados.

CDU 02:141:005.7

Tarcísio Franco Jaime

# Rotulação com Algoritmos Supervisionados

**Qualificação de Mestrado** apresentada ao Programa de Pós-Graduação em Ciência da Computação da UFPI (área de concentração: Sistemas de Computação), como parte dos requisitos necessários para a obtenção do Título de Mestre em Ciência da Computação.

Trabalho aprovado. Teresina-PI, 01 de janeiro de 2018:

---

**Vinicius Ponte Machado**
Orientador

---

Co-Orientador

---

**Professor**
Convidado 1

---

**Professor**
Convidado 2

---

**Professor**
Convidado 3

Teresina-PI
Janeiro de 2017

*Aos meus pais XXXXXXXX e YYYYYYY,*
*por sempre estarem comigo em todos os momentos.*

# Agradecimentos

Agradeço a Deus.

Agradeço aos meus pais, XXXXX e YYYYY, por ...

Aos meus irmãos, por.....

Agradeço ao meu orientador, XXXXXXXXX, por todos os conselhos, pela paciência e ajuda nesse período.

Aos meus amigos ...

Aos professores ...

À XXXXXX pelo apoio financeiro para realização deste trabalho de pesquisa.

*"Não sei o que,*
*não sei o que,*
*não sei o que lá."*
*(Autor Desconhecido)*

# Resumo

Frente ao aumento do tráfego de dados em consequência de novas tecnologias, como também a necessidade de mais equipamentos conectados à rede pedindo por processamento de dados, cada vez mais algoritmos de aprendizado de máquina não-supervisionados estão sendo estudados para obterem bons resultados, na criação de grupos (cluster), em face de obteção de informações úteis desses grupos. A partir desse problema de agrupamento, em grandes volumes de dados, tem-se um grau de dificuldade diretamente proporcional ao crescimeto desse volume. É nesse tema onde este trabalho atua, muito embora a importância desta proposta de mestrado esteja na interpretação, no entendimento dos grupos e não na criação dos mesmos. Diante o entendimento desses grupos esta pesquisa realiza de forma empírica, ou seja, através de experimentos e testes, a identificação de atributos mais significativos no grupo, junto com faixa de valores que mais se repete a ponto de representá-lo (rotulação). Dessa forma para a realização da rotulação de grupos de dados a proposta desta pesquisa é utilizar dois algoritmos supervisionados, cada um, com paradigmas diferentes: Naive Bayes (estatístico) e CART (simbólico). E a partir dos testes demonstrar que a rotulação é capaz de representar o grupo possuindo uma acurácia acima de 70% de acerto dos valores representados pelo rótulo escolhido.

**Palavras-chaves**: cluster. rotulação. aprendizado supervisionado.

# Abstract

This is the english abstract.

**Keywords**: cluster. rotulação.

# Lista de ilustrações

# Lista de tabelas

# Lista de abreviaturas e siglas

| | |
|---|---|
| EWD | Discretização por Larguras Iguais |
| EFD | Discretização por Frequências Iguais |
| CART | Classification and Regression Trees |

# Sumário

# 1 Introdução

Com a popularização da internet e mídias sociais, cada vez mais dados são processados, transportados e produzidos. E hoje, termo como, Big Data, faz parte do cotidiano de empresas e pessoas. De acordo com o autor Montgomery (2013) Big Data são os dados que excedem a capacidade de sistemas de banco de dados. É nesse cenário, com grandes volumes de dados, que não só a formação de grupos ganha importância, mas também a compreensão dos mesmos, pois a interpretação dos grupos fornecerá informações úteis para análises desses clusters.

Agrupamento de dados, ou clustering, é o termo que se usa para identificar dois ou mais objetos pertencentes ao mesmo grupo que compartilham um conceito em comum (KUMAR; ANDU; THANAMANI, 2013). Cluster é um termo bastante pesquisado no aprendizado não-supervisionado (subárea do aprendizado de máquina) e aplicada em vários contextos como segmentação de imagens, recuperação de informação e reconhecimento de objetos. Os algoritmos de agrupamento, conforme Kumar, Andu e Thanamani (2013), são aplicados em diferentes campos: Biologia (classificação de plantas e animais), Marketing (encontrar grupos de clientes com comportamentos semelhantes), planejamento de cidades (identificação de casas de acordo com seu tipo, valor e localização geográfica), entre outros.

O grau de escalabilidade dos dados gradativamente aumenta no decorrer dos anos, e embora os estudos sobre o problema de agrupamento de dados estejam avançados, fica cada vez mais complexo o entendimento dos clusters formados, pela razão do número crescentes de grupos criados. Quanto maiores são os números de grupos produzidos mais difícil são suas interpretações.

Diante desse contexto é que se extrai a temática desta proposta de mestrado qual seja - ”Rotulação automática de grupos através de algoritimos supervisionados baseados em árvores e estatísticos” - o estudo em questão dedica-se na aplicabilidade de dois algoritimos supervisionados, com paradigmas diferentes e bases de dados distintas, a fim de definir a tupla atributo/valor de maior importância nos clusters, determinando um significado para estes clusters (rotulação).

A formação do problema desta pesquisa nasce a partir do trabalho realizado por LOPES (2014), que se dedicou a estudar a possibilidade de realização de rotulação automática de grupos utilizando algoritmo não-supervisionado (K-means) para formação de grupos e algoritmo supervisionado (Redes Neurais) para a rotulação. Assim, partindo deste estudo já realizado, este trabalho questiona-se: É possível realizar rotulação de grupos de dados a partir de outros algoritmos supervisionados não testados, em específico Naive Bayes e CART?

Acredita-se que o resultado de tal problemática será positivo, considerando que os algoritmos Naives Bayes e CART também são categorizados como supervisionados. Além disso, é necessário mensurar ainda, a acurácia de cada resultado através do percentual de acertos dos atributos que são representados pelos rótulos gerados. Assim, se for possível demonstrar que a acurácia é de pelo menos 60% ficará então comprovada a possibilidade de se fazer rotulação dos grupos de dados utilizando os algoritimos supervisionados Naive Bayes e CART, pois percentuais menores do rótulo não representariam o grupo. Importante destacar, que este trabalho não se preocupa em criar grupos, mas dar maior relevância à rotulação dos mesmos, isto é, compreender os grupos de dados já formados.

O termo rotulação, neste trabalho, segue a definição conforme LOPES (2014):

**Definição 1** *Dado um conjunto de clusters* $C = \{c_1, ..., c_k | K \geqslant 1\}$*, de modo que cada cluster contém um conjunto de elementos* $c_i = \{\vec{e}_1, .., \vec{e}_{n^{(c_i)}} | n^{(c_i)} \geqslant 1\}$ *que podem ser representados por um vetor de atributos definidos em* $\mathbb{R}^m$ *e expresso por* $\vec{e}^{c_i} = (a_1, .., a_m)$ *e ainda que com* $c_i \cap c_{i'} = \{\emptyset\}$ *com* $1 \leqslant i, i \leqslant K$ *e* $i \neq i'$*; o objetivo consite em apresentar um conjunto de rótulos* $R = \{r_{c1}, ..., r_{ck}\}$*, no qual cada rótulo específico é dados por um conjunto de pares de valores, atributo e seu respectivo intervalo,* $r_{ci} = \{(a_1, [p_1, q_1]), ..., (a_{m^{(c_i)}}, ]p_{m^{(c_i)}}, q_{m^{(c_i)}}])\}$ *capaz de melhor expressar o cluster* $c_i$ *associado.*

- $K$ *é o número de clusters;*
- $c_i$ *é o i-ésimo cluster qualquer;*
- $n^{c_i}$ *é o número de elementos do cluster* $c_i$*;*
- $\vec{e}_{n^{(c_i)}}$ *se refere ao j-ésimo elemento pertencente ao cluster* $c_i$*;*
- $m$ *é a dimensão do problema;*
- $r_{c_i}$ *é o rótulo referente ao cluster* $c_i$*;*
- $]p_{m^{(c_i)}}, q_{m^{(c_i)}}]$ *representa o intervalo de valores do atributo* $a_{m^{(c_i)}}$*, onde* $p_{m^{(c_i)}}$ *é o limite inferior e* $q_{m^{(c_i)}}$ *é o limite superior;*
- $m$ *é a dimensão do problema;*

Em um exemplo, no qual a base de dados possui classes já definidas: macho, fêmea ou raça X, Y, Z, etc. E que ao criar esses grupos sabe-se que existe uma correlação das características dos grupos, acabando por não deixar visível qual característica se apresenta mais significativa dentro desses grupos. Tem-se na rotulação a intenção de definir algum significado para estes grupos, gerando um tipo de rótulo, $R = \{r_{c1}, ..., r_{ck}\}$, para melhor expressar o cluster $c_i$ associado (Definição 1).

Tecnicamente a informação do rótulo aplicada no cluster pode ajudar na tomada de decisão em algum contexto. A exemplo disso, supõe-se uma situação empregada na área urbana, onde pessoas circulam na cidade e imagina-se que os dados de controle de seus celulares estão sendo capturados pelas células das torres, e gravados em uma base

de dados pelas operadoras. Uma vez em posse desses dados, são criados clusters podendo ser aplicado rotulação nestes grupos. E através dos rótulos pode-se personalizar alguns serviços para esses grupos já formados.

Seguindo o exemplo dos dados capturados do celular, caso o rótulo ($r_{c_i}$) de um cluster ($c_i$) fosse o atributo localização, e os valores desse atributo escolhido para compor o rótulo, fossem as coordenadas geográficas, o qual definiriam o tipo de localização. Logo percebe-se que os participantes desse grupo possuem característica de frequentar alguma localização em comum. A interpretação deste rótulo poderá implicar em uma tomada de decisão personalizada para este grupo, objetivando otimizar um problema.

O trabalho em questão tem como objetivo principal demonstrar a possibilidade de fazer rotulação de dados, em grupos já formados, utilizando dois algoritimos supervisionados distintos com paradigmas diferentes. Sendo este um algoritmo com paradigma estatístico - Naive Bayes - e outro com paradigma simbólico - Classification And Regression Tree (CART).

Para alcançar tal objetivo é necessário ???? (falar do objetivo de cada capítulo)foi estruturada mediante a codificação por intermédio de uma linguagem de natureza técnica, onde fez uso de módulos de aprendizado de máquina, atuando em bases de dados e obtendo como saída deste programa, os rótulos dos grupos. Uma vez que estes grupos já possuem informações do provedor das bases de dados de como foram criados.

Esta pesquisa é eminentemente quantitativa, pois se utiliza de algoritmos supervisionados para selecionar os atributos de maior relevância nos clusters, através de um percentual de correlação entre atributos, isto é, quanto maior esse percentual maior será a relevância desse atributo em relação aos outros. Além disso faz uma análise da base de dados de forma subjetiva para definir o número de faixas que serão divididos os valores, para realização da discretização. Uma vez escolhido o atributo de maior relevância e selecionada a faixa de valor que mais se repete nesse atributo, o resultado será o rótulo composto pela tupla: atributo mais importante e faixa selecionada.

O trabalho será disposto em cinco capítulos já incluso a Introdução e Conclusão, capítulos 1 e 5 respectivamente.

O Referencial Teórico abordado no capítulo 2 é responsável em esclarecer as tecnologias utilizadas nesta pesquisa e dividida em três seções. Inicialmente na seção 2.1, tem-se uma explanação sobre aprendizado de máquina e quais os aprendizados indutivos são mais relevantes para este trabalho, ademais, a explicação dos dois algoritmos supervisionados utilizados para fazer rotulação de dados. Já na seção 2.2 é realizado a divisão das faixas de valores de cada atributo, chamada de discretização. E logo na seção 2.3 são apresentas pesquisas já consolidadas referentes ao assunto de rotulação de clusters.

Na capítulo 3 é abordada a definição do problema da pesquisa. A partir dessa

definição um modelo de resolução é definido e apresentado um fluxograma exibindo os processos a serem seguidos. Logo na seção 3.3 é demonstrado o funcionamento da técnica de correlação entre atributos. E na seção 3.4 uma base de dados fictícia é utilizada para exemplificar a execução dos processos do modelo de resolução: discretização da base de dados no Processos (I), no Processo (II) é aplicado o algoritmo supervisionado e no Processo (III) o resultado da rotulação.

No capítulo 4 os resultados são apresentados separados por cada base de dados. Sendo que em cada algoritmo testado o resultado é dividio em cluster, atributo rótulo desse cluster, faixa de valores compondo o rótulo e mais dois campos expondo o grau de relevância, em porcentagem, de cada atributo em relação aos outros, junto com o número de elementos que não são representados pelo rótulo escolhido. A partir destas informações é retirado o rótulo o qual representará o cluster.

Diante de todo o exposto fica claro que esta pesquisa além de dar continuidade a um tema específico aplicado na interpretação de agrupamento de dados, também serve como ponto de partida para outra pesquisa mais aprofundada, onde poderá esta tentar comprovar a possibilidade de fazer rotulação de dados utilizando qualquer algoritmo supervisionado.

# 2 Referencial Teórico

Será abordado neste capítulo o conteúdo base para compreensão deste trabalho dividido em 3 seções: Aprendizado de Máquina, Discretização e Trabalhos Correlatos.

Na seção 2.1 contempla os principais tipos de aprendizados indutivos dando ênfase a aprendizagem supervisionada, foco da proposta deste mestrado. A indução é um tipo de inferência lógica, onde a partir de um pequeno número de observações pode-se ter uma conclusão geral de todo o conjunto. Então, caso a pequena amostra não tenha dados suficientes ou os dados da amostra não forem relevantes, o conhecimento induzido acaba sendo prejudicado por generalizar o conhecimento adquirido, dessa amostra, para todo o grupo de dados.

Já na seção 2.2 dissertará sobre a técnica de discretização adotada nesta pesquisa. Possuindo grande contribuição para os resultados gerados, e ganhando assim uma seção própria para explanação de como funciona essa técnica. E na seção 2.3, serão abordados trabalhos que possuam mesmas características desta pesquisa adicionando conhecimento ao tema.

## 2.1 Aprendizado de Máquina

A aprendizagem de máquina, diferente das metodologias tradicionais de implementação, utiliza sua experiência anterior, para melhorar suas respostas a partir de problemas em determinadas áreas. Segue sua definição segundo Mitchell (1997):

**Definição 2** *Um programa de computador aprende com a experiência E em relação a alguma classe de tarefas T e medida de desempenho P, se seu desempenho em tarefas em T, conforme medido por P, melhora com a experiência E*

Um exemplo, seria a realização do reconhecimento facial de uma pessoa utilizando aprendizado de máquina. Seria inserida várias fotos tituladas de uma certa pessoa no banco de dados, e após vários exemplos, o programa de computador seria capaz de predizer se uma nova foto é de uma determinada pessoa através de aprendizado anterior, ou melhor, fotos anteriormente inseridas.

Alguns motivos justificam que não é possível simplismente exigir do projetista implementar melhorias no sistema, de forma que ele esteja robusto bastante para lidar com todas as situações (RUSSEL; NORVIG, 2013). Um desses motivos seria a incapacidade da antecipação de todas as situações possíveis de implementação por parte do programador.

Fazendo um resumo, aprendizado de máquina seriam algoritmos capazes de aprender automaticamente através de determinados exemplos, ou comportamentos.

A partir desta síntese, tem-se uma observação. A classificação de dados no contexto de aprendizado de máquina, são compostos por dois pilares. Um, seriam os **dados** a serem classificados, e outro, o **algoritmo** que irá atuar nessa base de dados. Existem vários algoritmos como exemplo: redes neurais, árvores de decisão, Suport Vector Machine – SVM, etc. Qualquer um destes algoritmos são utilizados para encontrar um classificador. E a escolha apropriada se dará através de métricas que avaliarão o desempenho de cada um, e a melhor métrica, será o algoritmo apropriado para aquele problema de classificação de dados.

### 2.1.1 Aprendizado Supervisionado

Nesta seção será abordado um método que através de uma base de dados classificada, será realizado uma predição de novos registros com base em vários desses exemplos já classificados, ou seja, é quando existir casos que possuem uma classificador disponível para determinados conjunto de dados (conjunto de treinamento), mas precisa ser previstos para outras instâncias. Os responsáveis por essas predições de novos registros são algoritmos de aprendizado supervisionados projetados para determinados fins.

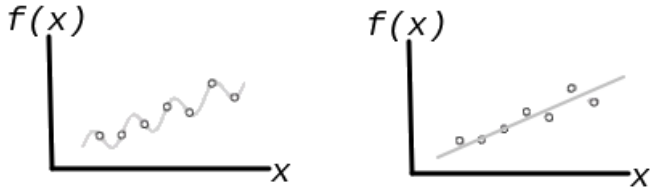
O termo ”Supervisionado” indica uma correlação entre os dados de entrada com a saída desejada (classe). Considerando uma base de dados de imagens de rostos, onde cada imagen possui uma saída representada por uma classe: masculino ou feminino. A tarefa seria criar um preditor capaz de acertar a cada novo registro se a imagem é masculina ou feminina. Seria difícil implementar de maneira tradicional, uma vez que são inúmeras as diferenças das faces masculinas e femininas. Embora haja uma dificuldade de distinção entre as faces, uma alternativa seria dar exemplos de rostos classificados, masculino ou feminino, e através desses exemplos aplicar o algoritmo que automaticamente faça a máquina ”aprender” uma regra para predizer qual sexo pertence cada rosto (BARBER, 2011).

Em (RUSSEL; NORVIG, 2013) os autores fazem uma apresentação formal do funcionamento da aprendizagem supervisionada. Dado um conjunto de treinamento

$$(x_1, y_2), (x_2, y_2), ...(x_n, y_n), \tag{2.1}$$

onde cada $y_j$ foi gerado por $y = f(x)$ desconhecida. Encontrar uma função $h$ que se aproxime da função $f$ real.

A função $h$ é uma hipótese onde prevê um melhor desempenho entre as hipóteses possíveis através dos conjuntos de dados, que são diferentes do conjunto de treinamento equação 2.1.




(a) Ajuste polinomial de grau 6

(b) Hipótese linear

Figura 1 – Hipóteses ajustadas

O exemplo da figura 1a mostra uma função de grau 6 onde acontece um sobreajuste (overfitting) no conjunto de dados de treinamento. Esse modelo acabou exibindo uma função mais complexa para atender todo o conjunto de dados do gráfico, ficando especifico para essa amostra.

Já na figura 1b o ajuste da função se torna mais simples e mesmo o gráfico não passando por todos os pontos, acabou por generalizar melhor o conjunto de treinamento, tornando talvez, um melhor resultado da predição de novos valores.

Em análise da figura 1 é apresentado duas hipóteses que tentam se aproximar ao máximo da função verdadeira ($h$), que é desconhecida. Mesmo parecendo que na figura 1a obteve-se melhor resultado, pois todos os pontos são atingidos pelo gráfico da função, este modelo acabou se ajustando muito bem na amostra de dados deixando a função $h$ muito específica, não retratando os dados em um mundo real. Então, apesar de parecer que a 1a por ser mais específica é a melhor função, não é a opção correta. Quanto mais generalizado for modelo, melhor será para predizer os valores de $y$ para novos conjuntos de dados.

#### 2.1.1.1 Algoritmo Classification and Regression Trees - CART

Esse algoritmo constroi modelos de previsão a partir de dados de treinamento onde seus resultados podem ser reprensentados em uma árvore de decisão. A árvore de decisão é uma ferramenta que da suporte à decisão utilizando como modelo um fluxograma semelhante a uma árvore, onde a cada nó folha é feita uma pergunta para tomada de decisão, permitindo uma abordagem do problema de forma estruturada e sistemática até chegar a uma conclusão lógica.

O algoritmo CART se torna um classificador se a saída da árvore de decisão for uma classe, definindo como árvore de classificação. Por exemplo, em um conjunto de dados de um paciente onde tenta prever se o mesmo possuirá câncer. A classe seria ”Terá Câncer” ou ”Não terá Câncer”. Mas ao contrário de uma árvore de classificação que prediz uma classe, o CART também pode assumir uma árvore de regressão, onde poderá prever um valor numérico ou contínuo, como período de tempo de internação do paciente, preço de uma cirurgia ou quantidade de água ingerida.

No caso de não ser probabilístico o grau de confiança em seu modelo de predição será embasada em respostas semelhantes em outras circunstâncias antes analisadas.

Inicialmente todas as amostras se concentram no nó raiz, e a partir daí é apresentado uma questão, onde a intenção é separar o nó raiz em dois grupos mais homeogêneos. Dependendo da questão as amostras irão para a folha esquerda ou direita do nó raiz. O CART faz essa divisão em função da regra Gini de Impureza[1] (**??**), e o índice Gini varia de 0 a 1, definindo o grau de pureza do nó.

$$Gini(S) = 1 - \sum p^2(j/t) \tag{2.2}$$

Onde: $p(j/t)$ é probabilidade a priori da classe $j$ se formar no nó $t$. E $S$ é um conjunto de dados que contém exemplos de n classes

Para construção de uma árvore existem três componente importantes (**??**):

- Um conjunto de perguntas que servirá de base para fazer uma divisão;
- Regras de divisão para julgar o quanto é boa esta divisão;
- Regras para atribuir uma classe a cada nó;

**Algorithm 1:** Rotina de funcionamento do CART com critério Gini (RAIMUNDO; MATTOS; WALESKA, 2008)

```
1  melhorGini; /* cria a variável                                   */
2  divisaoCorrente ← 4.9;/* Ex. recebe o 1° valor do atributo       */
3  direita ← 0;
4  esquerda ← 6;/* Ex. recebe o total de dados existentes para o
       atributo                                                     */
5  while existirem dados do
6      if 1ª Dado Lista do Atributo MAIOR divisaoCorrente then
7          valorGini ← calculaGini(divisaoCorrente);
8      else
9          valorGini ← calculaGini(1ªDadoLista);
10     if  Primeiro Gini encontrado then
11         melhorGini ← valorGini;
12     else
13         if valorGini > melhorGini then
14             melhorGini ← valorGini
15     divisaoCorrente ← 5.4; /* recebe o próximo dado do atributo   */
16     direita recebe o que possui +1 e esqueda o -1;
17     (valorGini + divisaoCorrente)/2;/* encontrar ponto de divisão */
```

[1] O CART pode utilizar outros critérios de divisão de dados como: entropia e critério de Twoing

#### 2.1.1.2 Algoritmo Naive Bayes

É um modelo probabilístico que pode ser calculado diretamente entre seus dados de treinamento. Depois de calculado, o modelo pode ser utilizado para fazer previsões de novos dados através do teorema de Bayes. Esse teorema utiliza uma teoria estatística e probabilística para previsão de acontecimento de um evento, sendo este evento relacionado a condição da probabilidade de ocorrência anteriores do mesmo. É nesse seguimento que o algoritmo Naive Bayes funciona. Criando classificadores probabilístico baseados no teorema de Bayes.

Pode-se citar como exemplo desse evento, a descoberta do câncer em uma pessoa, pois se essa doença estiver relacionada ao sexo, então, utilizando o teorema de Bayes, o sexo de uma pessoa pode ser utilizada para da maior precisão a probabilidade de câncer, ao invés de fazer uma avaliação de probabilidade sem a utilização do sexo da pessoa.

O Naive Bayes utiliza uma técnica de independência dos atributos, onde cada variável de entrada não depende de recursos de outras. Essa independência condicionada entre os atributos, os quais nem sempre ocorrem nos problemas reais, acabou deixando conhecida por Bayes ingênuo, ou Naive Bayes.

Naive Bayes como classificador estatístico possue um modelo de simples construção, e ficou conhecido por ter bons resultados em relação a algoritmos mais sofisticados, mesmo trabalhando com grandes quantidades de dados. Ele agrupa objetos de uma certa classe em razão da probabilidade do objeto pertencer a esta classe.

$$P(c/x) = \frac{P(x/c)P(c)}{P(x)} \tag{2.3}$$

$$P(c/x) = P(x_1|c) * P(x_2|c) * ... * P(x_n|c) * P(c) \tag{2.4}$$

- $P(c/x)$ probabilidade posterior da classe $c, alvo$ dada preditor $x, atributos$.
- $P(c)$ é a probabilidade original da classe.
- $P(x|c)$ é a probabilidade que representa a probabilidade de preditor dada a classe.
- $P(x)$ é a probabilidade original do preditor.

A utilização do algoritmo Naive Bayes já é bem difundida, e está presente em vários trabalhos, como classificação de textos, filtro de SPAM, analisador de sentimentos, entre outros (MADUREIRA, 2017; LUCCA et al., 2013; WU et al., 2008; MCCALLUM; NIGAM, 1997). Mas mesmo atingido boa popularidade possui pontos negativos. A suposição de ter preditores independentes não acontece muito na vida real, pois acaba sendo difícil ter uma amostra de dados que sejam inteiramentes independentes.

### 2.1.2 Aprendizado Não-Supervisionado

No Aprendizado Não-Supervisionado, não existe uma tentativa de se encontrar uma função que se aproxime da real. Logo porque os registros não são classificados, então o conjunto de treinamento não possui informação da saída sobre determinada entrada. Desta forma os algoritmos procuram algum grau de similaridade entre os registros e tenta agrupá-los de forma a ter algum sentido deles estarem juntos.

Quando o algoritmo encontra dados com mesma similaridade ele os agrupa formando clusters. Os números de clusters encontrados irão depender de como os algoritmos funcionam, junto com o grau de dissimilaridade entre elementos de grupos diferentes. Como não existe uma variável classe no Aprendizado Não-Supervisionado, então (BARBER, 2011) diz que o maior interesse seria em uma perspectiva probabilística de ditribuição $p(x)$ de um determinado conjunto de dados.

$$D = \{x_n, n = 1, ..., N\} \tag{2.5}$$

Uma vez que no conjunto (2.5), não existe classe $y$, encontrado em um conjunto de treinamento, equação 2.1, o algoritmo precisa encontrar padrões nos atributos para fazer os agrupamentos.

## 2.2 Discretização

O método de discretização faz a conversão de valores contínuos em valores discretos. A partir de um atributo com valores contínuos, a discretização irá forçar um ponto inicial e final definindo um intervalo e designando uma faixa para cada intervalo. Assim, ao invés de valores contínuos em cada atributo, será relacionado a faixa que aquele atributo pertence, definindo assim seu novo valor.

Segundo alguns autores (**??**HWANG; LI, 2002) a discretização melhora a precisão e deixa um modelo mais rápido em seu conjunto de treinamento. Os métodos de discretização mais comumente utilizados no âmbito dos métodos não-supervisionados de acordo com (KOTSIANTIS; KANELLOPOULOS, 2006; DOUGHERTY; KOHAVI; SAHAMI, 1995) são os métodos de Discretização por Larguras Iguais(EWD) e Discretização por Frequências Iguais (EFD).

### 2.2.1 Discretização por Larguras Iguais - EWD

O método de Discretização por Larguras Iguais (EWD) faz a discretização de um intervalo, entre valores contínuos, dividindo em faixas de tamanhos iguais. Logo se existir um intervalo com valores contínuos [a,b], e deseja particionar em $R$ faixas de tamanhos iguais serão necessários $R - 1$ pontos de corte, figura 2.

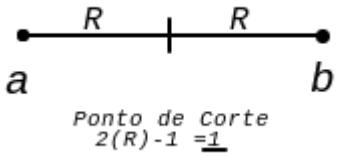



Figura 2 – Ponto de Corte (R-1)

Para haver o ponto de corte antes tem que ser realizado a ordenação dos dados. A largura de cada faixa $r_1, ..., r_R$ na equação 2.6 é representada por $w$, que é calculada pela diferença entre os limites superior e inferior do intervalo, dividido pela quantidade $R$ de valores a serem gerados.

$$w = \frac{b - a}{R} \tag{2.6}$$

A variável $w$ determina os pontos de corte $(c_1, ..., c_{R-1})$ que irão delimitar o tamanho das faixas de valores. O primeiro ponto de corte, $c_1$, é obtido através da soma do limite inferior $a$ com a tamanho de $w$. E os pontos de corte seguintes são calculados pela soma do ponto de corte anterior com $w$.

O valor de cada faixa será representado por $i$, onde $i$ é o índice indicando a faixa. De acordo com a figura 3 para dividir o intervalo $[a, b]$ em $R$ faixas será necessário de $R - 1$ pontos de corte.

$$c_i = \begin{cases} a + w, & se\, i = 1 \\ c_{i-1} + w, & caso\, contrário \end{cases} \tag{2.7}$$

O valor da faixa do intervalo $[a, c_1]$ será o valor discreto igual ao índice de sua faixa $r_1$. Então, um valor na faixa $r_1$ terá o valor reprensentado por $1(um)$, pois $i = 1$ é o limite inferior mais largura da faixa, equação 2.7. E seguindo o mesmo raciocínio o valor da faixa $r_2 = ]c_1, c_2]$ é representado por $2(dois)$, e consequentemente o valor que se encontra em uma faixa qualquer $r_i$ será reprensentado por $i$.

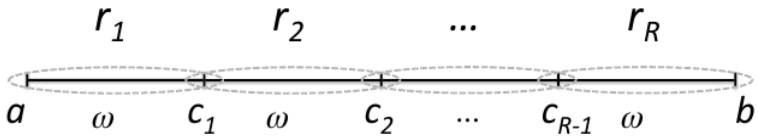



Figura 3 – Discretização EWD baseada em (LOPES, 2014)

### 2.2.2 Discretização por Frequência Iguais - EFD

Esse outro método de discretização já possue uma abordagem diferente a do EWD, pois a idéia é manter a quantidade de elementos distintos, entre os pontos de corte, com o

mesmo número. Dado um intervalo $[a, b]$ o número de faixas $R$ e a quantidade de valores distintos $\xi$ , onde $\xi \geqslant R$ o método EFD irá segmentar em $R$ faixas de valores que possuem a mesma quantidade de elementos distintos $\lambda$. Então serão realizados $R - 1$ pontos de corte gerando $R$ faixas de valores, $(r_1, ..., r_R)$, com a mesma quantidade de elementos distintos $\lambda$. Para encontrar $\lambda$ calcula-se o valor inteiro da divisão entre a quantidade de elementos distintos $\xi$ pela quantidade de faixas de valores $R$, obtendo o número de elementos da faixa 2.8.

$$\lambda = \frac{\xi}{R} \tag{2.8}$$

Uma observação nesse método é a ocorrência em amostras que possuem uma má distribuição de valores de um dado atributo. Como um número significativo de repetições, causando um desiquilíbrio nas distribuições dos elementos.

Uma vez no intervalo $[a, b]$ de elemetos ordenado e calculado $\lambda$ contendo $R$ elementos $v_{[R]}$ pode-se determinar os pontos de corte $(c_1, ..., c_{R-1})$ que são os delimitadores das faixas. Cada ponto de corte $c_i$ pode ser calculado por $v_{i\lambda} - \acute{e}simo$ elemento, 2.9.

$$c_i = v_{[i\lambda]} \tag{2.9}$$

Como na seção anterior do método EWD o valor que estiver no intervalo $[a, c_1]$ terá seu valor associado a um valor discreto igual ao índice $i$ de sua faixa $r_i$ conforme figura 4. Então, caso o valor esteja na faixa $r_2$ ele passará a ter o valor de seu índice $i$ igual a $2(dois)$. De maneira consecutiva os valores que estiverem na faixa $r_3 = ]c_2, c_3]$ terão valor $3(três)$. Uma outra observação desse método é que diferente do EWD, as faixas podem assumir faixas com tamanhos diferentes.

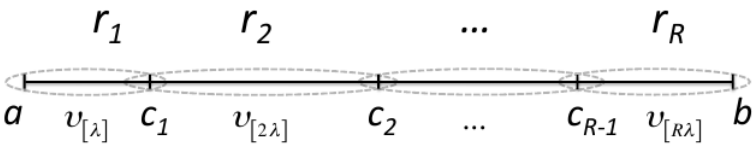



Figura 4 – Discretização EFD[2]

## 2.3 Trabalhos Correlatos

Esta seção propõe relacionar outros trabalhos servindo de complemento teórico para entender a variedade de aplicações referente ao assunto de rotulação de dados.

O trabalho escrito por LOPES (2014) fez um estudo abordando o tema de rotulação de dados, tema este, proposto também por esta pesquisa, mas com abrangência e execução

[2] Figura extraída de (LOPES, 2014)

diferentes do modelo da figura 5 . Nesse trabalho foi utilizado como entrada um conjunto de dados onde foi feito o agrupamento automático, com algoritmos não-supervisionados formando clusters. Logo após é utilizado algoritmos supervisionados nos grupos de dados, e apresentado como saída um rótulo específico que melhor define o grupo formado. Esses rótulos são formados pela faixa de valor, que mais se repetem, em conjunto com os atributos mais relevantes.

Pode-se verificar na figura 5 que na parte onde é aplicado os algoritmos supervisionados (processo III) é o local exato que esta pesquisa utiliza para testar outros algoritmos supervisionados, servindo para comprovar a hipótese desta proposta de mestrado.

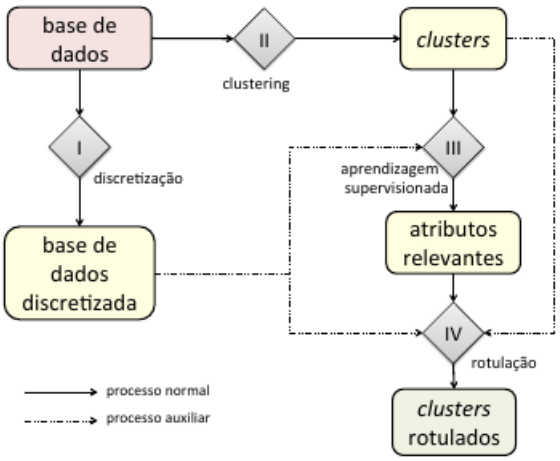



Figura 5 – Modelo (LOPES, 2014)

Em (LIMA, 2015) o problema em questão é fazer classificação e rotulação em uma base que possuem poucos elementos classificados utilizando método semi-supervisionado. O método inicia com uma base dividida em elementos classificados(L) e não classificados(U). Após cada iteração o grupo L vai crescendo e automaticamente diminuindo o grupo U até que não tenha mais nenhum elemento em U, figura 6. Após isso é realizado uma etapa de agrupamento, sem levar em consideração os dados classificados anteriormente. Terminada essa etapa é feito uma validação para saber quais os rótulos foram considerados corretos.

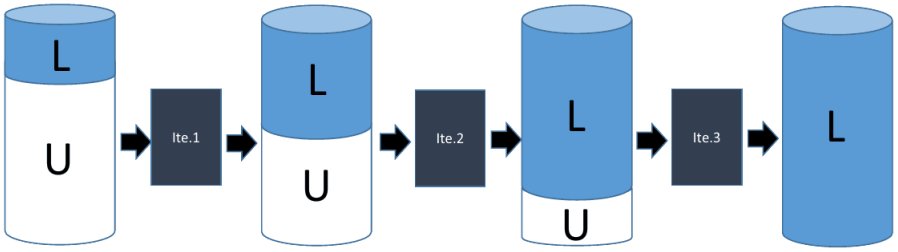



Figura 6 – Comportamento da base de dados a cada iteração. Método (LIMA, 2015)

O método proposto é uma combinação de um classificador com um método de agrupamento, onde a rotulação de um conjunto de dados é feita com conhecimento prévio

de um outro conjunto menor rotulado. O classificador treina com a parte de dados rotulada e classifica os dados não-rotulados.

Outra pesquisa sore rotulação está em (FILHO, 2015) onde aborda o mesmo Problema de Rotulação, mas a atuação é diferenciada, pois o modelo, figura 7, procura diferenças existentes em cada grupo através da seleção dos elementos que representam o grupo, e depois é construído a faixa de valores. Os grupos são formados pelo algoritmo Fuzzy C-Means e após isso que é selecionado os atributos.

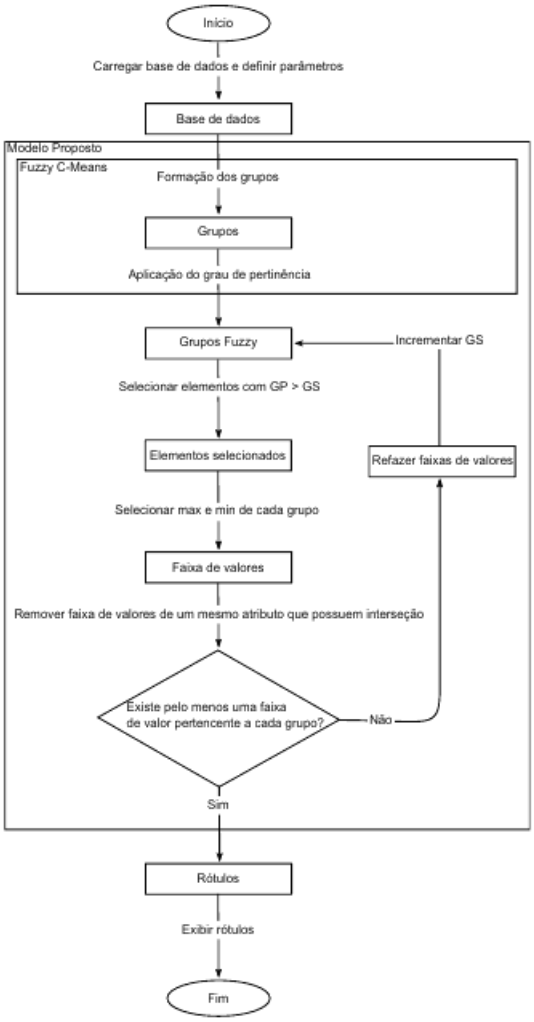



Figura 7 – Modelo (FILHO, 2015)

# 3 Metodologia

Esse capítulo abordará o problema proposto por esse trabalho, e logo em seguida, será apresentado um modelo de resolução para rotulação de dados utilizando dois algoritmos supervisionados baseados em paradigmas simbólico e estatístico . O objetivo ao final deste capítulo é poder resolucionar o problema de rotulação, e atribuir a qualquer outro pesquisador todo o conhecimento necessário para replicar este trabalho através das informações produzidas aqui.

## 3.1 Considerações do Problema

A abordagem do problema referente a essa proposta de mestrado segue uma linha já pequisada por LOPES (2014), que seria o **Problema de Rotulação**. Esse conceito, rotulação de dados, já é estudado na literatura na área de aprendizagem não-supervisionada, subseção 2.1.2, onde é comum os algoritmos lidarem com os agrupamentos dos dados, e a criação de clusters a partir dos graus de similaridade entre os elementos.

Muitas pesquisas realizadas na área de rotulação fazem referencia, de fato, a classificação do dados e não da rotulação nos termos desse trabalho. Ao agrupar um conjunto de elementos por um derterminado critério, esta havendo uma classificação desses elementos escolhidos, mas pouco se sabe, qual é a compreensão desses grupos já classificados.

Existe uma importância na criação dos clusters, contudo para o espectador é interessante existir um rótulo, desse grupo formado, oferecendo elementos em alguma tomada de decisão em razão de seu significado (rótulo).

Tem-se então o real problema de rotulação, contudo é necessário existir algum elemento definindo o porquê daquele grupo formado. O elemento é um rótulo composto por um, ou vários, atributos de maior relevância no cluster, junto com uma faixa de valores.

Essa faixa, é um intervalo de valores definido pela discretização seção 2.2, onde o intervalo escolhido, seria a faixa que apresenta os valores que se repetem com a maior frequência. A exemplo, tem-se um vetor de elementos já discretizados, $\vec{e}_{(c_i)} = \{1, 1, 1, 2, 2, 2, 2, 3, 3\}$. Neste vetor o valor que mais se repete é o número 2, então, essa faixa é a escolhida junto com o atributo mais relevante para compor o rótulo.

O Problema de Rotulação é formalmente definido como segue abaixo:

**Definição 3** *Dado um conjunto de clusters* $C = \{c_1, ..., c_k | K \geqslant 1\}$*, de modo que cada cluster contém um conjunto de elementos* $c_i = \{\vec{e}_1, .., \vec{e}_{n^{(c_i)}} | n^{(c_i)} \geqslant 1\}$ *que podem ser re-*

*presentados por um vetor de atributos definidos em* $\mathbb{R}^m$ *e expresso por* $\vec{e}^{c_i} = (a_1, .., a_m)$ *e ainda que com* $c_i \cap c_{i'} = \{0\}$ *com* $1 \leqslant i, i \leqslant K$ *e* $i \neq i'$. [1]

- $K$ *é o número de clusters;*
- $c_i$ *é o i-ésimo cluster qualquer;*
- $n^{c_i}$ *é o número de elementos do cluster* $c_i$*;*
- $\vec{e}_{n^{(c_i)}}$ *se refere ao j-ésimo elemento pertencente ao cluster* $c_i$*;*
- $m$ *é a dimensão do problema;*

## 3.2 O Modelo de Resolução

Uma vez já conhecido a definição do problema - *Definição 3* - é possível situar a abrangência abordada aqui nessa pesquisa, pois a intenção do estudo científico desenvolvido aqui é provar a realização de rotulação de dados com os dois algoritmos supervisionados de paradigmas diferentes: Naive Bayes e CART, utilizando as técnicas abordadas neste texto.

Este modelo de resolução consiste em apresentar como saída um conjunto de rótulos, onde cada rótulo específico é dado por um conjunto de pares de valores, atributo e seus respectivos intevalos, gerados a partir das frequências dos valores repetidos neste intervalo. Segue *Definição 4* formalizando a saída do modelo:

**Definição 4** *Dado um conjunto de rótulos* $R = \{r_{c1}, ..., r_{ck}\}$*, no qual cada rótulo específico é dados por um conjunto de pares de valores, tem como saída um vetor com atributo e seu respectivo intervalo,* $r_{ci} = \{(a_1, [p_1, q_1]), ..., (a_{m^{(c_i)}}, ]p_{m^{(c_i)}}, q_{m^{(c_i)}}])\}$ *capaz de melhor expressar o cluster* $c_i$*.* [2]

- $k$ *número de rótulos;*
- $R$ *representa o conjunto de rótulos na saída do modelo;*
- $a$ *é o atributo*
- $c_i$ *é o i-ésimo cluster;*
- $r_{c_i}$ *é o rótulo referente ao cluster* $c_i$*;*
- $]p_{m^{(c_i)}}, q_{m^{(c_i)}}]$ *representa o intervalo de valores do atributo* $a_{m^{(c_i)}}$*, onde* $p_{m^{(c_i)}}$ *é o limite inferior e* $q_{m^{(c_i)}}$ *é o limite superior;*
- $m$ *é a dimensão do problema;*

Como apresentado na seção 2.3, o autor LOPES (2014) foca em rotulação automática de grupos utilizando a estratégia de aprendizagem de máquina supervisionada, com paradigma conexionista, para provar seu trabalho. Porém, nesta pesquisa foi aplicado no

[1] Adaptada de (LOPES, 2014)
[2] Adaptada de (LOPES, 2014)

modelo de resolução dois algoritmos com paradigmas de aprendizado diferente do que já havia sido testado anteriormente, provando que é possível fazer rotulação de dados com algoritmos supervisionados com paradigmas simbólico e probabilístico, CART e Naive Bayes respectivamente.

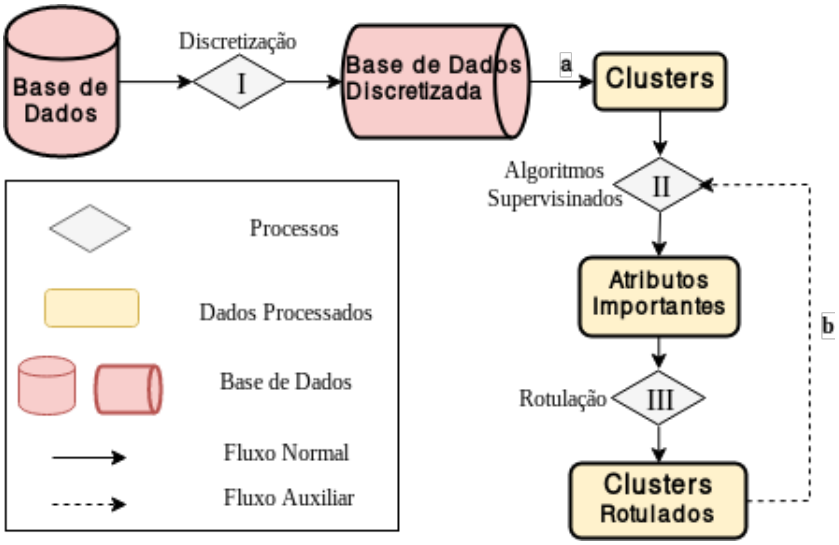



Figura 8 – Modelo de Resolução Proposto

O modelo (figura 8) inicialmente mostra a Base de Dados já classificada, pois o cerne desta pesquisa é conceder ao grupo um significado,a rotulação, através da técnica de correlação entre atributos, sessão 3.3. Essa base conterá valores contínuos, contudo, conforme modelo será necessário aplicar o método de discretização (I).

Uma vez com a base discretizada ocorre somente a separação dos clusters já classificados de acordo com a própria base de dados[3]. Isso é o funcionamento do fluxo (a), que nada mais é do que a separação da base em grupos já classificados.

No passo (II) serão executados os algoritmos de aprendizagem supervisionados, já visto nas subseções 2.1.1.1 e 2.1.1.2. Essa etapa utiliza uma técnica demonstrada na seção 3.3 sendo uma das mais importantes do método. A quantidade de vezes que O algoritmo supervisionado é aplicado irá ser o mesmo número de atributos do conjunto de dados. Utilizando a figura 9 como exemplo, o algoritmo supervisionado seria executado três vezes, sendo essa quantidade igual ao número de atributos.


Figura 9 – Exemplo da técnica aplicada ao atr1 sendo classe

[3] UCI - Machine Learning Repository. http://archive.ics.uci.edu/ml/

Seguindo para o processo (III) acontecerá a escolha do(s) atributo(s) mais relevante(s), selecionado na tabela de atributos importantes, junto com o valor mais frequente desse atributo. Após essa etapa é criado um conjunto de rotulos para cada clusters. O fluxo (b) será utilizado enquanto houver outros algoritmos para serem executados.

## 3.3 Técnica de Correlação entre Atributos

Essa técnica[4] possui um grau de processamento diretamente proporcional a quantidade de características expressa na base de dados definido em $R^m$. Ela implica em utilizar todos os atributos, menos o definido como classe, para fazer uma correlação entre eles junto ao algoritmo.

Utilizando como exemplo uma base com os seguintes atributos: **atr1, atr2, atr3, classe**. O atributo classe é retirado e a cada iteração um atributo será definido como a nova classe. Em um primeiro processamento de três, o primeiro atributo **atr1** se torna classe e executado com os outros dois atributos restantes com um algoritmo supervisionado.

O resultado da correlação entre os atributos **atr2, atr3** em relação ao **atr1** (figura 9) é armazanado em uma matriz. Por conseguinte é realizado a aplicação do algoritmo com **atr2** sendo classe, e assim sucessivamente até o último atributo. Essa etapa só é finalizada quando todos os atributos tiverem a chance de ser classe e armazenado seus valores em porcentagem na tabela. No final uma tabela será formada pelos valores em porcentagem da correlação entre eles.

## 3.4 Exemplo

Para melhor esclarecer as etapas da figura 8, será utilizado a tabela 1 como exemplo no processo de modelo de resolução proposto nesta pesquisa. Uma tabela de ciquenta linhas e três atributos e um atributo classe. Logo na primeira coluna da tabela, possui o índice da linha da tabela identificando cada registro e outros campos são atributos que definem características do registro identificado pelo índice da primeira coluna, e na quinta coluna a classe de cada registro.

Seguindo a definição 3 um elemento é expresso por um vetor de dimensão $m$, com tamanho igual ao número de atributos. Um exemplo do elemento 2 da tabela 1, pode ser representado por $\vec{e_2} = (1.26, 85.03, 20.45)$.

[4] Desenvolvida também por (LOPES, 2014)

Tabela 1 – Base de Dados Modelo

| | atr1 | atr2 | atr3 | classe | | atr1 | atr2 | atr3 | classe |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 2.08 | 92.11 | 22.07 | 2 | 26 | 1.42 | 53.51 | 19.64 | 3 |
| 2 | 1.26 | 85.03 | 20.45 | 1 | 27 | 1.12 | 62.71 | 19.07 | 1 |
| 3 | 2.00 | 108.36 | 22.68 | 2 | 28 | 2.09 | 60.58 | 20.20 | 1 |
| 4 | 1.74 | 43.78 | 18.72 | 3 | 29 | 1.95 | 69.23 | 19.68 | 1 |
| 5 | 1.82 | 100.20 | 23.09 | 2 | 30 | 1.03 | 47.81 | 19.47 | 3 |
| 6 | 1.43 | 77.59 | 21.80 | 1 | 31 | 1.75 | 90.92 | 21.39 | 2 |
| 7 | 1.53 | 44.01 | 20.98 | 3 | 32 | 1.72 | 42.35 | 22.89 | 3 |
| 8 | 1.14 | 107.77 | 18.99 | 2 | 33 | 1.47 | 101.77 | 19.20 | 2 |
| 9 | 1.97 | 98.00 | 22.32 | 2 | 34 | 1.53 | 41.16 | 22.67 | 3 |
| 10 | 1.50 | 39.67 | 21.78 | 3 | 35 | 1.44 | 93.61 | 21.03 | 2 |
| 11 | 1.74 | 55.86 | 20.31 | 3 | 36 | 1.51 | 98.65 | 19.24 | 2 |
| 12 | 1.80 | 65.72 | 19.62 | 1 | 37 | 1.06 | 68.82 | 21.68 | 1 |
| 13 | 1.33 | 82.01 | 19.82 | 1 | 38 | 1.48 | 80.40 | 21.43 | 1 |
| 14 | 1.66 | 103.93 | 21.10 | 2 | 39 | 1.14 | 61.59 | 19.90 | 1 |
| 15 | 1.42 | 66.14 | 21.61 | 1 | 40 | 1.08 | 91.93 | 20.81 | 2 |
| 16 | 1.87 | 88.36 | 22.45 | 2 | 41 | 1.62 | 79.21 | 18.43 | 1 |
| 17 | 1.11 | 107.82 | 19.32 | 2 | 42 | 1.68 | 80.87 | 18.42 | 1 |
| 18 | 2.08 | 67.66 | 20.74 | 1 | 43 | 1.81 | 98.24 | 22.13 | 2 |
| 19 | 1.85 | 82.65 | 20.35 | 1 | 44 | 1.30 | 69.27 | 18.83 | 1 |
| 20 | 1.04 | 102.62 | 19.46 | 2 | 45 | 1.80 | 101.21 | 21.61 | 2 |
| 21 | 1.97 | 100.37 | 21.94 | 2 | 46 | 1.79 | 72.02 | 22.02 | 1 |
| 22 | 1.95 | 45.70 | 22.10 | 3 | 47 | 1.56 | 81.71 | 22.10 | 1 |
| 23 | 1.77 | 50.04 | 20.16 | 3 | 48 | 1.98 | 77.16 | 21.71 | 1 |
| 24 | 1.97 | 81.57 | 19.83 | 1 | 49 | 1.86 | 89.12 | 22.84 | 2 |
| 25 | 1.52 | 93.13 | 20.61 | 2 | 50 | 1.55 | 76.01 | 19.74 | 1 |

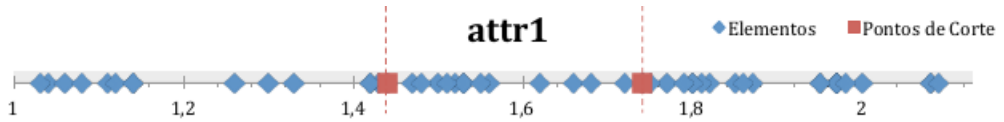
### 3.4.1 Processo (I) - Discretização

Segundo Catlett (1991), **??**) através de resultados expermentais, na conversão em atributos discretos ordenados de vários domínios constatou, que a mudança de representação da informação na maioria das vezes pode aumentar a acurácia do sistema de aprendizado. Dessa maneira a etapa de discretização ganha um papel importante no modelo, e também no processo de Rotulação (III), pois é utilizada uma inferência na faixa discretizada para encontrar o intervalo na faixa.

Utilizando como exemplo a tabela 1 será utilizada a técnica de discretização por frequências iguais - EFD - e divisão de números de faixas igual a R=3. Na figura[5] 10 poderá ser vizualizado como é feita a discretização.

Através da figura 10 fica claro o conteúdo da faixa 1, contendo o valor inicial, 1(um), até o primeiro ponto de corte. Na faixa 2, o valor inicial é o primeiro número após o primeiro ponto de corte (término da faixa 1) até o segundo ponto de corte, incluindo o

[5] Figura adaptada de (LOPES, 2014)




(a) Discretização em atr1


(b) Discretização em atr2


(c) Discretização em atr3

Figura 10 – Discretização de atributos utilizando EFD com R = 3

próprio ponto de corte. E na faixa 3 contém todos valores a partir do segundo ponto de corte.

Tabela 2 – Base de Dados Modelo Discretizada

| | atr1 | atr2 | atr3 | classe | | atr1 | atr2 | atr3 | classe |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 3 | 3 | 3 | 2 | 26 | 1 | 1 | 1 | 3 |
| 2 | 1 | 2 | 2 | 1 | 27 | 1 | 1 | 1 | 1 |
| 3 | 3 | 3 | 3 | 2 | 28 | 3 | 1 | 2 | 1 |
| 4 | 2 | 1 | 1 | 3 | 29 | 3 | 2 | 1 | 1 |
| 5 | 3 | 3 | 3 | 2 | 30 | 1 | 1 | 1 | 3 |
| 6 | 1 | 2 | 3 | 1 | 31 | 3 | 3 | 2 | 2 |
| 7 | 2 | 1 | 2 | 3 | 32 | 2 | 1 | 3 | 3 |
| 8 | 1 | 3 | 1 | 2 | 33 | 2 | 3 | 1 | 2 |
| 9 | 3 | 3 | 3 | 2 | 34 | 2 | 1 | 3 | 3 |
| 10 | 2 | 1 | 3 | 3 | 35 | 1 | 3 | 2 | 2 |
| 11 | 2 | 1 | 2 | 3 | 36 | 2 | 3 | 1 | 2 |
| 12 | 3 | 1 | 1 | 1 | 37 | 1 | 2 | 3 | 1 |
| 13 | 1 | 2 | 1 | 1 | 38 | 2 | 2 | 2 | 1 |
| 14 | 2 | 3 | 2 | 2 | 39 | 1 | 1 | 2 | 1 |
| 15 | 1 | 1 | 2 | 1 | 40 | 1 | 3 | 2 | 2 |
| 16 | 3 | 2 | 3 | 2 | 41 | 2 | 2 | 1 | 1 |
| 17 | 1 | 3 | 1 | 2 | 42 | 2 | 2 | 1 | 1 |
| 18 | 3 | 1 | 2 | 1 | 43 | 3 | 3 | 3 | 2 |
| 19 | 3 | 2 | 2 | 1 | 44 | 1 | 2 | 1 | 1 |
| 20 | 1 | 3 | 1 | 2 | 45 | 3 | 3 | 2 | 2 |
| 21 | 3 | 3 | 3 | 2 | 46 | 3 | 2 | 3 | 1 |
| 22 | 3 | 1 | 3 | 3 | 47 | 2 | 2 | 3 | 1 |
| 23 | 3 | 1 | 2 | 3 | 48 | 3 | 2 | 3 | 1 |
| 24 | 3 | 2 | 2 | 1 | 49 | 3 | 3 | 3 | 2 |
| 25 | 2 | 3 | 2 | 2 | 50 | 2 | 2 | 1 | 1 |

A tabela 2 é o resultado após a discretização de todos os atributos. Conforme cada valor discreto representando um inrtevalo, poderá o algoritmo está perdendo um pouco de informação, pois a discretização acaba generalizando a informação por agrupar os dados em intervalos representando-os de forma mais igual. Mas um fato a se levar em consideração é o número de faixas de valores, de uma variável discreta, pois se esse número for muito grande poderá acarretar ausência de generalização.

### 3.4.2 Processo (II) - Algoritmos Supervisionados

Ao chegar nessa etapa, Processo (II) da figura 8, já se tem uma base discretizada e clusters formados, tabela 2. É feita a execução do algorimo de aprendizado supervisionado e identificado os atributos de maior importância de cada cluster.

Uma vez com o cluster, serão percorridos todos os atributos, onde a cada iteração um atributo será a classe da vez. Nesse exemplo primeiramente o atributo **atr1** será classe, e os demais irão participar como entrada junto ao algoritmo, e verificar seu grau de importância entre eles. Depois o atributo **atr2** irá ser classe, e depois o **atr3**, fechando o ciclo de todos os atributos do cluster. Como visualizado na figura 11

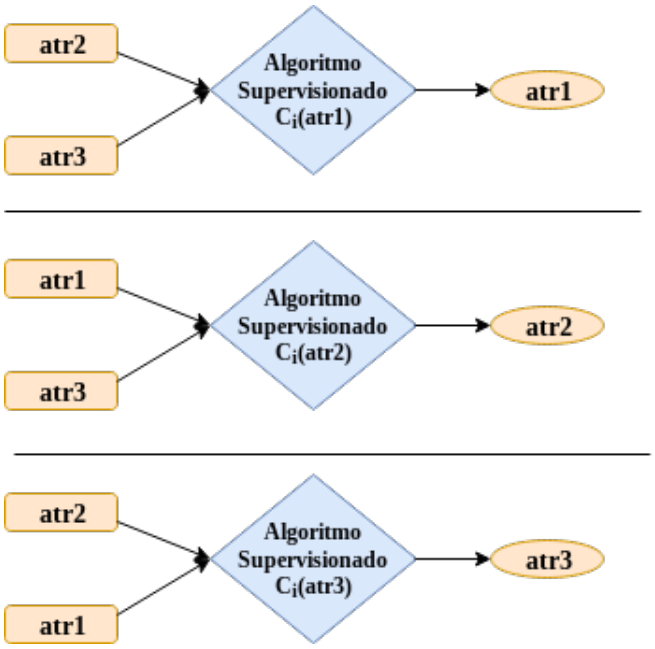
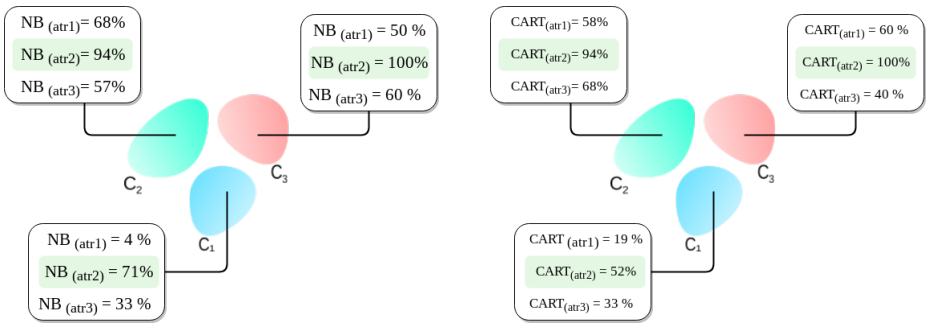



Figura 11 – Exemplo da técnica de correlação aplicada aos 3(*três*) atributos, cada um sendo classe em determinada iteração

A cada aplicação do algoritmo supervisionado é armazanado para cada cluster $c_i(atr)$ os valores de relevância dos atributos representada por uma porcentagem de acerto. O algoritmo será executado o mesmo número de vezes do número de atributos existente na base de dados, pois a cada iteração um atributo se torna um atributo classe, consequentemente é gerado seu valor de relevância em porcentagem. Quanto maior

sua porcentagem, mais bem correlacionado é o atributo em relação aos demais (figura 12).Portanto esse atributo poderá resumir as características do problema, podendo ser considerado atributo mais relevante, e escolhido como rótulo.




(a) Execução do Naive Bayes

(b) Execução do algorimo de Árvore de Decisão

Figura 12 – Resultado dos Algoritmos

Na figura 12a mostra o resultado da execução do Naive Bayes trabalhando com a base modelo (tabela 2) exibindo os resultados em porcentagem de acerto de cada atributo em relação aos demais. O mesmo acontece com a figura 12b onde é aplicado um algorimo de Árvore de Decisão - CART - exibindo o resultado de todas as taxas de acerto, em porcentagem, dos atributos de seus respectivos clusters.

Uma forma de eliminar uma possível ambiguidade entre os clusters foi adicionar na implementação uma variável $V$. Essa variável é utilizada para seleção dos atributos rótulos de um clusters, caso aconteça de os rótulos se repetirem em clusters diferentes. Logo, todos os atributos que tiverem até uma diferença $V$ em relação ao atributo de maior taxa de acerto, expresso em porcentagem, serão escolhidos como rótulo. Isto posto, se o atributo de maior taxa de acerto possuir 90%, e o $V = 10\%$ então todos outros atributos que tiverem valores a partir de 80% são selecionados como rótulo do cluster.

O valor da variável $V$ é subjetivo e irá ser arbitrado de acordo com os resultados em cada aplicação do algoritmo em cima de um conjunto de dados. Nesse exemplo os atributos importantes com $V = 12$ utilizando a figura 12a, teriam os rótulos, por clusters, $r_{c_i} : r_{c_1} = \{atr2\}$, $r_{c_2} = \{atr2\}$, $r_{c_3} = \{atr2\}$.

### 3.4.3 Processo (III) - Rotulação

Nesse processo de rotulação serão calculados os intervalos dos atributos que estão na figura 8 com o atributos de maior relevância (maior porcentagem) selecionado na etapa anterior. Para compor o rótulo $r_{c_i}$ do cluster $c_i$ é calculado a faixa do atributo que tiver maior frequência. É possível verificar neste exemplo, da Base Modelo, o resultado da figura 12a, onde o rótulo $r_{c_1}$ é o **atr2=]67.66, 88.36]**, porque o valor da faixa de maior

frequência do cluster $c_1$ em relação ao atributo **atr2** é a faixa 2 (figura 10c), que representa o limite inferior ]67.66s e o limite superior, 88.36].

Uma vez terminado o processo (III) de rotulação, o fluxo $b$ da figura 8, só será executato caso seja necessário testar com outro algoritmo.

O resultado da rotulação dos Algoritmos Naive Bayes 12a e de Árvore de Decisão 12b - CART - aplicado na BD Modelo são:

- $r_{c_1} = (atr2, ]67.66, 88.36])$;
- $r_{c_2} = (atr2, ]88.36, 108.36])$ ;
- $r_{c_3} = (atr2, [39.67, 67.66])$;

O algoritmo 2 exibe a rotina em forma de pseudocódigo para melhor entendimento.

**Algorithm 2:** Rotina de Rotulação

```
1  Carrega_valores_auxiliares(V, R, TipoDiscretização);
2  Carrega_BD;
3  Discretiza_BD;
4  Separa_em_clusters_de_acordo_com_classificação_BD;
5  while existir clusters do
6      while existir atributos do
7          atributo_classe=seleciona_nova_classe(atributos) ;
8          Aplica_algoritmo_supervisionado(atributo_classe, atributos_naoClasse);
9          Calcula_matriz_de_porcentagem_de_acertos;
10     Carrega_atributos_importantes_considerando_V;
11     Associa_valores_aos_intervalos;
12 Exibe_rótulos_todos_clusters;
```

# 4 Resultados

Foram realizados testes com algumas bases de dados da UCI Machine Learning[1], um repositório de dados a serviço da comunidade de aprendizado de máquina. Criado por estudantes de pós-graduação na UC Irvine em 1987 e até hoje é utilizado não só por estudantes mas também por educadores e pesquisadores como fonte primária de aprendizado de máquina.

As bases de dados foram escolhidas não só por critérios comparativos de outro trabalho que também já as utilizaram servindo de referência para os resultados, como também, um cuidado de só escolher bases que estão classificadas, uma vez que esta pesquisa trabalhará com os clusters já formados e não na criação de grupos.

A divisão deste capítulo de resultados iniciará por uma explanação da implementação do trabalho, e logo após, cada base de dados utilizada é destacada em uma seção. A cada seção referente a uma base de dados houve uma divisão de subseções de acordo com os algoritmos utilizados: Naive Bayes e CART.

## 4.1 Implementação

Para conseguir gerar os resultados aqui escritos foram feitas implementações utilizando a ferramenta MATLAB[2,3], onde foi possível utilizar suas funções de aprendizado de máquina já implementadas na Statistics and Machine Learning Toolbox, uma toolbox com implementações preparadas para aprendizado de máquina. Por aprensentar linguagem técnica e funcões já prontas direcionada para aprendizado de máquina essa ferramenta foi escolhida para colocar em prática essa pesquisa.

Ao longo da pesquisa foram realizados vários testes, porém, nesses testes houveram alterações de algumas variáveis, e métodos de discretização, sempre tentando melhorar os resultados. Essas alterações dependendo da base de dados utilizadas são: variável "V", quantidade de faixas "R" e métodos de discretização "EWD,EFD".

Como dito na subseção 3.4.2 a variável $V$ existe para evitar a ambiguidade dos rótulos, ou seja, quando rótulos apresentarem os mesmos resultados: atributo e faixa de valor. Além de evitar a ambiguidade dos rótulos a variável $V$ pode ser utilizada também para selecionar mais de um atributo para ser o rótulo do cluster.

Essa situação é necessária após uma análise da tabela de correlação dos atributos

[1] http://archive.ics.uci.edu/ml/
[2] http://www.mathworks.com/products/matlab/
[3] versão: R2016a(9.0.0.341360); 64-bit (glnxa64)

(seção 3.3), a exemplo da tabela 4. E existindo valores muito próximos em relação a outros, e se necessário, utilizar esses atributos também como rótulos, a variável $V$ pode ser configurada com um valor que possa abranger esses atributos que possuem valor de porcentagem próximos ao do atributo mais relevante. O valor de $V$ é subjetivo e sua adição é condicionada a análise da aplicação do algoritmo na bases de dados.

A composição da tabela de correlação de atributos é o resultado da aplicação do algoritmo enquanto o atributo era a classe da vez, coforme figura 11, seção 3.4.2. O atributo de maior valor junto com os atributos da diferença de $V$ com o mais relevante, são escolhidos para ser rótulos. Na linha (cluster) 1 o maior valor é o atributo perimetro. Então o valor de perimetro subtraído de $V = 3$ vai resultar no limete inferior. A partir daí o(s) atributo(s) que possuí(rem) um valor que está entre este resultado, limite inferior, até o de maior porcentagem, irá compor o rótulo.

Na implementação onde é feita a composição da tabela de correlação de atributos é percorrido todos os atributos e a cada iteração um atributo será a saída, ou seja, o atributo classe conforme figura 11, seção 3.4.2. A cada iteração um valor da correlação do atributo classe com os demais atributos é gerado compondo a tabela de correlação.

Após a tabela estar montada o atributo rótulo é selecionado a partir do maior valor em relação aos outros atributos, e caso seja necessário também é selecionado como rótulo os atributos que possuam o valor entre a diferença de $V$ com o mais relevante. Na linha (cluster) 1 o maior valor é o atributo perimetro. Então o valor de perimetro subtraído de $V = 3$ vai resultar no limete inferior. A partir daí o(s) atributo(s) que possuí(rem) um valor que está entre este resultado, limite inferior, até o de maior porcentagem, irá compor o rótulo.

A cada base de dados descritas nas seções, são configuradas as variáveis e método de discretização utilizados e implementado dois algoritmos de aprendizado supervisionado com paradigmas diferentes para fazer rotulação. Cada algoritmo terá como resultado um rótulo por cluster de dados servindo de amostra para testar a acurácia.

Os algoritmos utilizados foram o Naive Bayes, subseção 2.1.1.2, com paradigma estatístico. E também o algoritmo Classification e Regression Trees - CART, subseção 2.1.1.1, com paradigma simbólico.

## 4.2 Seeds - Identificação de Tipos de Semente

Essa base pertence a UCI Machine Learning, composta por sete atributos definindo suas características e mais um atributo classe responsável por identificar os tipos de sementes. Em seus atributos seus valores são todos contínuos e não existem valores em branco, possuindo um total de 210 registros classificados em três categorias:

- 70 elementos do tipo Kama;
- 70 elementos do tipo Rosa;
- 70 elementos do tipo Canadian.

Para classificar as sementes, como Kama, Rosa e Canadian foi utilizada uma técnica de raio X, que é relativamente mais barata que outras técnicas de imagem, como microscopia ou tecnologia a laser. O material foi colhido de campos expermentais, explorados no Instituo de Agrofísica da Academia Polonês de Ciências em Lublin.

Como já mencionado neste capítulo, seção 4.1, antes de executar o algoritmo algumas configuração são necessárias. A primeira configuração é o método de discretização do tipo EFD, a segunda é a divisão dos valores dos atributos em faixas, $R = 3$ para todos os atributos, e também o valor de variação $V = 0\%$, não obstante, esta variável $V$ só assumirar valor maior que zero após análise dos resultados caso haja ambiguidade.

Na tabela 3 e tabela 6 são apresentados os resultados de rotulação dos algoritmo Naive Bayes e CART respectivamente. Essas tabelas são compostas por colunas informando os número dos **Clusters**, **Rótulos** integrando **Atributo** e sua **Faixa** de valor. Junto também a coluna **Relevância** exibindo a resposta do algoritmo de rotulação em porcentagem da correlação do atributo em relação aos outros atributos do cluster, retirado da tabela 4 e da tabela 7 respectivamente. E por último a coluna **Elem Fora da Faixa** que mostra a quantidade de elementos que não estão dentro da faixa designada pelo do rótulo encontrado.

Essa última coluna, **Elem Fora da Faixa**, tem a função de exibir, em números, a quantidade de valores que não estão participando da porcentagem da coluna de **Relevância**. Através de expermentos percebeu-se o mérito de apresentar em números a quantidade de elementos que não estão sendo representados pelo rótulo gerando mais realidade as informações, ao invés de exibir em porcentagem.

### 4.2.1 Naive Bayes

Tabela 3 – Resultado da rotulação com o algoritmo Naive Bayes

| | Rótulos | | | |
|---|---|---|---|---|
| Cluster | Atributos | Faixa | Relevância(%) | Elem fora da Faixa |
| 1 | area | ] 12.78 ∼ 16.14 ] | 92% | 14 |
| 2 | area | ] 16.14 ∼ 21.18 ] | 95% | 6 |
| 3 | perimetro | [ 12.41 ∼ 13.73 ] | 95% | 5 |

Analisando a coluna Rótulos da tabela 3, nota-se que o atributo **area** aparece tanto no cluster 1 como também no cluster 2. A técnica envolve não só o atributo mais relevante, como também, a faixa que os valores mais se repetem dentro do atributo. Nesse

caso pode-se observar que o atributo se repete entre os clusters, mas no cluster 1 a faixa de valores difere do cluster 2, sendo considerados rótulos distintos.

Caso os resultados gerados na tabela 4 expusessem clusters com rótulos ambíguos, poderia ser utilizado a variação de $V$. Quando houver ambiguidade dos rótulos a seleção dos atributos, que compõem os rótulos, acontecerá da diferença da variável $V$ em relação ao atributo de maior relevência do cluster. Caso essa variável tenha o valor alterado, os rótulos dos clusters poderão sofrer mudanças, pois poderia aumentar ou diminuir o número de atributos dos rótulos, dependendo do valor inserido em $V$. Através da tabela 4 é possível analisar todos os valores de relevância gerados para os atributos e analisar qual valor pode-se inserir em $V$ para montar o rótulo.

Para exemplificar a utilização da variável $V$ pode-se utilizar como exemplo os dados do cluster 2 da tabela 4 e adotando $V = 3\%$. Neste exemplo não só o atributo de maior relevância, **area** com 95.7% seria escolhido como rótulo, mas também o atributo **lkernel** com valor 92.8%, pois a diferença entre o valor de **area** com $V$ resultaria em 92.7%. Através dessa diferença todos os atributos que estivessem na faixa de 92.7% a 95.7% seriam selecionados como atributos do rótulo.

Tabela 4 – Resultado da Correlação dos atributos pelo Naive Bayes; Legenda dos Atributos: (A)area, (B)perimetro, (C)compacteness, (D)Lkernel, (E)Wkernel, (F)asymetry, (G)lkgroove

| | | Atributos | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | A | B | C | D | E | F | G |
| Clusters | 1 | 92.8 | 87.1 | 50.0 | 75.7 | 85.7 | 60.0 | 65.7 |
| | 2 | 95.7 | 91.4 | 47.1 | 92.8 | 90.0 | 28.5 | 85.7 |
| | 3 | 91.4 | 95.7 | 71.4 | 85.7 | 91.4 | 64.2 | 58.5 |

A tabela 4 é formada por clusters representado pelas linhas, e atributos representado por colunas. Essa tabela é fruto da implementação do Naive Bayes na base de dados **Seeds**, e foi gerada para auxiliar a retirada dos atributo(s) rótulo(s). Uma análise pode ser feita através desses dados e ajudar a definir um valor para a variável $V$ caso necessário. Percebe-se que algumas características são mais bem correlacionadas que outras, através de seus valores mais altos. Isso indica o grau de relacionamento entre os atributos após a aplicação do algoritmo.

Para provar empiricamente os resultados, na tabela 5 é exposto o resultado de 4(*quatro*) execuções do Algoritmo Naive Bayes, e pode-se constatar que mesmo havendo algumas alterações em seus valores nos atributos em cada execução, a correlação entre os atributos não oferece muita alteração. Como exemplo, o atributo **area** nos clusters 1 e 2, possuem o melhor grau de correlacionamento em seus grupos, mesmo nas quatro execuções, como mostrado na tabela 5.

Segue abaixo o resultado do algoritmo Naive Bayes na base de dados **Seeds** com

Tabela 5 – Resultado de 4(*quatro*) execuções do algoritmo Naive Bayes; Legenda dos Atributos: (A)area, (B)perimetro, (C)compacteness, (D)Lkernel, (E)Wkernel, (F)asymetry, (G)lkgroove

| 1a. Execução | | Atributos | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | A | B | C | D | E | F | G |
| Clusters | 1 | 92.8 | 87.1 | 48.5 | 77.1 | 82.8 | 57.1 | 65.7 |
| | 2 | 94.2 | 90.0 | 45.7 | 92.8 | 90.0 | 38.5 | 87.1 |
| | 3 | 91.4 | 95.7 | 72.8 | 85.7 | 91.4 | 64.2 | 60.0 |

| 2a. Execução | | Atributos | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | A | B | C | D | E | F | G |
| Clusters | 1 | 92.8 | 87.1 | 47.1 | 77.1 | 87.1 | 60.0 | 65.7 |
| | 2 | 94.2 | 90.0 | 47.1 | 92.8 | 91.4 | 32.8 | 87.1 |
| | 3 | 91.4 | 95.7 | 72.8 | 85.7 | 92.8 | 64.2 | 60.0 |

| 3a. Execução | | Atributos | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | A | B | C | D | E | F | G |
| Clusters | 1 | 94.2 | 85.7 | 48.5 | 77.1 | 82.8 | 61.4 | 65.7 |
| | 2 | 92.8 | 90.0 | 50.0 | 92.8 | 90.0 | 32.8 | 87.1 |
| | 3 | 91.4 | 95.7 | 72.8 | 85.7 | 92.8 | 64.2 | 60.0 |

| 4a. Execução | | Atributos | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | A | B | C | D | E | F | G |
| Clusters | 1 | 91.4 | 88.5 | 54.2 | 75.7 | 85.7 | 62.8 | 61.4 |
| | 2 | 95.7 | 90.0 | 50.0 | 92.8 | 90.0 | 38.5 | 85.7 |
| | 3 | 91.4 | 95.7 | 72.8 | 85.7 | 94.2 | 64.2 | 57.1 |

seus rótulos:

- $r_{c_1} = \{(area, ]12.78 \sim 16.14])\}$
- $r_{c_2} = \{(area, ]16.14 \sim 21.18])\}$
- $r_{c_3} = \{(perimetro, [12.41 \sim 13.73])\}$

### 4.2.2 CART

Já na tabela 6, tem-se o resultado da aplicação do algoritmo supervisionado na rotulação. Ele é um algoritmo de classificação de árvore de decisão utilizado pela toolbox do MATLAB. O intuito é testar a base de dados em diferentes paradigmas.

Tabela 6 – Resultado da aplicação do algoritmo CART

| | Rótulos | | | |
|---|---|---|---|---|
| Cluster | Atributos | Faixa | Relevância(%) | Elem fora da Faixa |
| 1 | perimetro | [ 13.73 ∼ 15.18 ] | 94% | 14 |
| 2 | area | ] 16.14 ∼ 21.18 ] | 98% | 6 |
| | perimetro | ] 15.18 ∼ 17.25 ] | 98% | 7 |
| 3 | wkernel | [ 2.63 ∼ 3.049 ] | 97% | 9 |

Foram realizadas vários teste, onde alguns desses testes estão na tabela 8. Essas operações foram execuções do algoritmo CART na base, para provar que a técnica de correlação de atributos, seção 3.3, é funcional para este algoritmo. O mesmo comportamento entre execuções pode ser visto no algoritmo de paradigma estatístico, subseção 4.2.1, realizado nessa pesquisa. O comportamento de ambos foram bem semelhantes, como tambéms seus valores não se alteram muito a cada iteração.

O resultado da rotulaçao utilizando o algoritmo CART na base de dados **Seeds** tem como rótulos:

Tabela 7 – Resultado da Correlação dos atributos pelo CART; Legenda dos Atributos: (A)area, (B)perimetro, (C)compacteness, (D)Lkernel, (E)Wkernel, (F)asymetry, (G)lkgroove

| | | Atributos | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | A | B | C | D | E | F | G |
| Clusters | 1 | 91.4 | 94.2 | 58.5 | 80.0 | 81.4 | 61.4 | 61.4 |
| | 2 | 98.5 | 98.5 | 51.4 | 90.0 | 88.5 | 42.8 | 88.5 |
| | 3 | 92.7 | 95.7 | 80.0 | 88.5 | 97.1 | 58.5 | 78.5 |

Tabela 8 – Resultado de 4(*quatro*) iterações do algoritmo CART; Legenda dos Atributos: (A)area, (B)perimetro, (C)compacteness, (D)Lkernel, (E)Wkernel, (F)asymetry, (G)lkgroove

| 1a. Execução | | Atributos | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | A | B | C | D | E | F | G |
| Clusters | 1 | 91.4 | 94.2 | 58.5 | 80.0 | 74.2 | 55.7 | 60.0 |
| | 2 | 98.5 | 98.5 | 50.0 | 90.0 | 88.5 | 41.4 | 90.0 |
| | 3 | 92.8 | 95.7 | 80.0 | 88.5 | 97.1 | 55.7 | 77.1 |

| 2a. Execução | | Atributos | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | A | B | C | D | E | F | G |
| Clusters | 1 | 91.4 | 94.2 | 62.8 | 78.5 | 81.4 | 61.4 | 57.1 |
| | 2 | 98.5 | 98.5 | 54.2 | 90.0 | 88.5 | 40.0 | 90.0 |
| | 3 | 92.8 | 95.7 | 80.0 | 88.5 | 97.1 | 60.0 | 77.1 |

| 3a. Execução | | Atributos | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | A | B | C | D | E | F | G |
| Clusters | 1 | 93.8 | 93.6 | 61.8 | 83.2 | 89.2 | 53.2 | 71.0 |
| | 2 | 98.2 | 98.3 | 61.9 | 93.0 | 90.5 | 25.2 | 90.1 |
| | 3 | 95.5 | 96.3 | 82.4 | 90.9 | 97.7 | 59.3 | 77.0 |

| 4a. Execução | | Atributos | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | A | B | C | D | E | F | G |
| Clusters | 1 | 92.8 | 94.2 | 60.0 | 80.0 | 84.2 | 64.2 | 60.0 |
| | 2 | 98.5 | 98.5 | 47.1 | 91.4 | 90.0 | 42.8 | 88.5 |
| | 3 | 91.4 | 95.7 | 80.0 | 88.5 | 97.1 | 55.7 | 77.1 |

- $r_{c_1} = \{(perimetro, ]13.73 \sim 15.18])\}$
- $r_{c_2} = \{(area, ]16.14 \sim 21.18]), (perimetro, ]15.18 \sim 17.25])\}$
- $r_{c_3} = \{(wkernet, [2.63 \sim 3.049])\}$

## 4.3 Iris - Identificação de Tipos de Plantas

A base de dados **Iris**, também pertencente a UCI Machine Learning, é muito conhecida em outras pesquisas[4] como também na literatura em reconhecimentos de padrões por utilizar classes de plantas bem definidas. Contêm 3 classes de 50 instâncias cada, totalizando 150 registros de amostra de plantas. O atributo classe classifica o tipo de planta em 3 tipos:

- 50 elementos da classe Iris-setosa ;
- 50 elementos da classe Iris-versicolour;
- 50 elementos da classe Iris-virginica.

Os atributos correspondentes são comprimento da sepala - SL, largura da sepala - SW, comprimento da pétala - PL e largura da pétala - PW. Através dessas características

[4] (LOPES, 2014; **??**; FILHO, 2015) e outros

há uma classificação para dizer qual tipo de planta.

Foi aplicado na configuração de execução do algoritmo o método de discretização, tipo EFD[5], a divisão de três faixas de valores R = 3 para todos os atributos, e inserido o valor de variação $V = 0\%$. Mais uma vez, o valor $V$ existe para evitar ambiguidade dos rótulos, podendo ser utilizado pelo pesquisador quando necessário após análise dos valores de correlação dos atributos nos grupos, tabela 10.

Seguindo a análise, semelhante da base de dados anterior, serão realizados testes utilizando dois algoritmos[6], e cada resultado será exibido em tabelas. Portando as colunas são formadas por **Clusters**, **Rótulos**, **Relevância** e **Elem fora da Faixa** representando os valores que não estão dentro da faixa escolhida como rótulo. Também foi posto nas tabelas 12 e 10 os resultados das correlações entre os atributos de cada grupo, servindo de informação para decisão do valor de $V$, caso fosse necessário. E também apresentado os resultados de outras iterações de cada algoritmo, para mostrar o comportamento dos atributos entre eles no grupo.

### 4.3.1 Naive Bayes

Através da tabela 9 os resultados da rotulação são exibidos após a aplicação do algoritmo. Com essa base de dados nota-se que no cluster 1 houve um acerto de 100% da rotulação. O cluster 2 e cluster 3 obtiveram rótulos distintos, cada um com com grau de relevância acima de 80% em relação aos outros atributos de cada grupo.

Tabela 9 – Resultado da aplicação do algoritmo Naive Bayes

| | Rótulos | | | |
|---|---|---|---|---|
| Cluster | Atributos | Faixa | Relevância(%) | Elem fora da Faixa |
| 1 | petallength | [ 1.0 ∼ 3.7 ] | 100% | 0 |
| | petalwidth | [ 0.1 ∼ 1.0 ] | 100% | 0 |
| 2 | petallength | ] 3.7 ∼ 5.1 ] | 84% | 7 |
| 3 | petalwidth | ] 1.7 ∼ 2.5 ] | 90% | 5 |

Os valore na coluna de relevância não podem ser analisados isoladamente. Para isso a tabela 10 possui os valores de todos os atributos no momento que ele são classes. Os valores são em porcentagem para melhor análise do grau de relacionamento entre os outros atributos.

Na tabela 10 foram inseridas quatro resultados de execuções do algoritmo. Foi escolhida na tabela 10 a 1a. execução para montar a tabela de rótulos, tabela 9. A partir dessas execuções o pesquisador poderá arbitrá sobre o valor de $V$ para melhor adaptá-lo a base. Das várias execuções expostas na tabela 10, percebe-se que não há muita diferença

[5] seção 2.2.2
[6] sessões 4.2.1,4.2.2

Tabela 10 – Resultado de 4(*quatro*) execuções do algoritmo Naive Bayes; Legenda dos Atributos: (SL)sepallength,(SW)sepalwidth,(PL)petallength,(PW)petalwidth

| 1a. Execução | | Atributos | | | |
|---|---|---|---|---|---|
| | | SL | SW | PL | PW |
| Clusters | 1 | 80 | 68 | **100** | **100** |
| | 2 | 72 | 76 | **84** | **82** |
| | 3 | 76 | 74 | 68 | **90** |

| 2a. Execução | | Atributos | | | |
|---|---|---|---|---|---|
| | | SL | SW | PL | PW |
| Clusters | 1 | 80 | 68 | 100 | 100 |
| | 2 | 72 | 76 | 88 | 84 |
| | 3 | 70 | 74 | 70 | 90 |

| 3a. Execução | | Atributos | | | |
|---|---|---|---|---|---|
| | | SL | SW | PL | PW |
| Clusters | 1 | 80 | 68 | 100 | 100 |
| | 2 | 72 | 74 | 84 | 84 |
| | 3 | 74 | 74 | 68 | 90 |

| 4a. Execução | | Atributos | | | |
|---|---|---|---|---|---|
| | | SL | SW | PL | PW |
| Clusters | 1 | 80 | 68 | 100 | 100 |
| | 2 | 72 | 74 | 86 | 82 |
| | 3 | 70 | 74 | 70 | 92 |

entre os valores de cada execução. Isso mostra um padrão de valores de acordo com a base. No caso da 1a. execução os valores escolhidos como rótulo estão destacados em cada cluster.

Se a tabela escolhida fosse a da 3a. execução, os valores de rótulos seriam modificados, em virtude dos valores mais altos serem iguais, fazendo que o rótulo assumisse dois atributos: PL e PW. Em análise do cluster 2 perecebe-se que os valores de PL e PW nas quatro execuções são bem próximos e até idênticos na terceira execução, como já dito anteriormente, então caso fosse necessário inserir um valor de variação $V$, um valor aceitável seria $V = 3$. Desta maneira manteria os rótulos dos clusters 1 e 3 sem alteração, e um novo atributo seria incluído no cluster 2, assumindo o novo rótulo com dois atributos: PL e PW.

Os rótulos com o algoritmo Naive Bayes na base de dados **Iris** são dados abaixo:

- $r_{c_1} = \{(petallength, [1.0 \sim 3.7]), (petalwidth, [0.1 \sim 1.0])\}$
- $r_{c_2} = \{(petallength, ]3.7 \sim 5.1])\}$
- $r_{c_3} = \{(petalwidth, ]1.7 \sim 2.5])\}$

### 4.3.2 CART

A aplicação do algoritmo CART na base de dados **Iris** gerou a tabela 13 como resultado, e ao examinar pode-se observar uma semelhança com a subseção anterior 4.3.1 onde foi aplicado o Naive Bayes.

Ao observar a tabela 13 percebe-se que o resultado de rotulação no cluster 1 e 3 são idênticos ao do algoritmo apresentado anteriormente, mas no cluster 2 o rótulo é diferenciado pelo atributo petalwidth que atinge valores mais altos em todas as execuções, como mostra a tabela 10.

Segue abaixo os rótulos na base de dados **Iris** aplicado no algoritmo CART:

Tabela 11 – Resultado da aplicação do algoritmo CART

| Cluster | Rótulos | | Relevância(%) | Elem fora da Faixa |
|---|---|---|---|---|
| | Atributos | Faixa | | |
| 1 | petallength | [ 1.0 ∼ 3.7 ] | 100% | 0 |
| | petalwidth | [ 0.1 ∼ 1.0 ] | 100% | 0 |
| 2 | petalwidth | ] 1.0 ∼ 1.7 ] | 90% | 8 |
| 3 | petalwidth | ] 1.7 ∼ 2.5 ] | 90% | 5 |

Tabela 12 – Resultado de 4(*quatro*) iterações do algoritmo CART; Legenda dos Atributos: (SL)sepallength,(SW)sepalwidth,(PL)petallength,(PW)petalwidth

| 1a. Execução | | Atributos | | | |
|---|---|---|---|---|---|
| | | SL | SW | PL | PW |
| Clusters | 1 | 80 | 68 | **100** | **100** |
| | 2 | 74 | 76 | **88** | **90** |
| | 3 | 68 | 68 | 74 | **90** |

| 2a. Execução | | Atributos | | | |
|---|---|---|---|---|---|
| | | SL | SW | PL | PW |
| Clusters | 1 | 80 | 68 | 100 | 100 |
| | 2 | 74 | 76 | 88 | 90 |
| | 3 | 70 | 70 | 74 | 90 |

| 3a. Execução | | Atributos | | | |
|---|---|---|---|---|---|
| | | SL | SW | PL | PW |
| Clusters | 1 | 80 | 68 | 100 | 100 |
| | 2 | 74 | 76 | 86 | 90 |
| | 3 | 70 | 66 | 78 | 90 |

| 4a. Execução | | Atributos | | | |
|---|---|---|---|---|---|
| | | SL | SW | PL | PW |
| Clusters | 1 | 80 | 68 | 100 | 100 |
| | 2 | 72 | 74 | 86 | 90 |
| | 3 | 68 | 66 | 78 | 90 |

- $r_{c_1} = \{(petallength, [1.0 \sim 3.7]), (petalwidth, [0.1 \sim 1.0])\}$
- $r_{c_2} = \{(petallength, ]3.7 \sim 5.1]), (petalwidth, ]1.0 \sim 1.7])\}$
- $r_{c_3} = \{(petalwidth, ]1.7 \sim 2.5])\}$

## 4.4 Glass - Identificação de Tipos de Vidros

Essa base ficou conhecida por Vina Spiehler, Ph.D. da DABFT Diagnostic Products Corporation, onde conduzio pesquisas e teste de comparação em seu sistema baseado em regras determinando, se o tipo de vidro era temperado ou não. Institutos de investigação criminológica motivaram os estudos de classificação de tipos de vidros, porque uma classificação corretamente identificada em uma cena de crime pode ser utilizada como prova.

Possui um total de 214 instancias, caracterizados por 9 atributos (RI, Na, Mg, Al, Si, K, Ca, Ba e Fe), sendo que o atributo **RI** indica o índice de refração, e quanto aos demais atributos são valores correspondentes a porcentagem do óxido.

Os tipos de vidro (atributo classe) foram divididos em 7 grupos distintos:

- 1 janelas de construção - vidro temperado: 70 registros
- 2 janelas de construção - vidro não-temperado: 76 registros

- 3 janelas de veículos - vidro temperado: 17 registros
- 4 janelas de veículos - vidro não-temperado: 0 registro
- 5 recipientes: 13 registros
- 6 louças de mesa: 9 registros
- 7 lâmpadas: 29 registros

Para execução dos algoritmos terão que ser definidos em quantas em quantas faixas ($R$) serão divididos os valores dos atributos, qual é o método de discretização e o valor de variação $V$ caso haja ambiguidade. Nos teste desenvolvidos nesta pesquisa os valore de referência foram, $R = 0$, o método de discretização EFD e o valor de $= 0$.

### 4.4.1 Naive Bayes

Tabela 13 – Resultado da aplicação do algoritmo CART

| | Rótulos | | | |
|---|---|---|---|---|
| Cluster | Atributos | Faixa | Relevância(%) | Elem fora da Faixa |
| 1 | Mg | [ 2.245 ∼ 4.490 ] | 100% | 0 |
| | K | [ 0.0 ∼ 1.5525 ] | 100% | 0 |
| | Ba | [ 0.0 ∼ 0.7875 ] | 100% | 0 |
| 2 | K | ] 0.0 ∼ 1.5525 ] | 100% | 0 |
| 3 | Mg | ] 2.245 ∼ 4.490 ] | 100% | 0 |
| | K | ] 0.0 ∼ 1.5525 ] | 100% | 0 |
| | Ca | ] 8.12 ∼ 10.81 ] | 100% | 0 |
| | Ba | [ 0.0 ∼ 0.7875 ] | 100% | 0 |
| 4 | Al | [ 1.0925 ∼ 1.895 ] | 92% | 4 |
| | K | [ 0.0 ∼ 1.5525 ] | 92% | 3 |
| | Ba | [ 0.0 ∼ 0.7875 ] | 92% | 1 |
| 5 | Na | [14.055 ∼ 17.38 ] | 100% | 2 |
| | K | [ 0.0 ∼ 1.5525 ] | 100% | 0 |
| | Ba | [ 0.0 ∼ 0.7875 ] | 100% | 0 |
| | Fe | [ 0.0 ∼ 0.1275 ] | 100% | 0 |

### 4.4.2 Naive Bayes

# 5 Conclusões, Trabalhos Futuros e Cronograma

Aqui serão abordados as conclusões dessa proposta de mestrado referentes aos resultados do capítulo 4, bem como uma sessão de Trabalhos Futuros e Cronograma. Em Trabalhos Futuros a pretenção é expor melhoriras e expansões de tudo que fora realizado nesta pesquisa, e mostrar que existe uma continuidade para todo esse estudo aqui elaborado. Já no Cronograma, será criado uma tabela divida em meses definindo os passos a serem seguidos até a conclusão da dissertação.

## 5.1 Conclusão

O que se pretende fazer aqui nesta sessão é comentar as conclusões dos resultados realizados neste trabalho, capítulo 4, onde nesse capítulo mostrou a aplicação dos algoritmos supervisionados em algumas bases de dados. E com todo conteúdo produzdio relatar se o problema proposto pelo trabalho foi solucionado, ou não.

Uma vez conhecido o problema, foi realizado a execução de dois algoritmos supervisionados, servindo de amostra para provar que era possível fazer rotulação de dados com estes algoritmos supervisionados, tema deste trabalho. E já identificando alguns trabalhos que já haviam feito rotulação, (LOPES; MACHADO; RABELO, ), utilizando algoritmos supervisionados, o intuito era executar outros algoritmos com paradigmas diferentes aos que foram realizados em pesquisas anteriores.

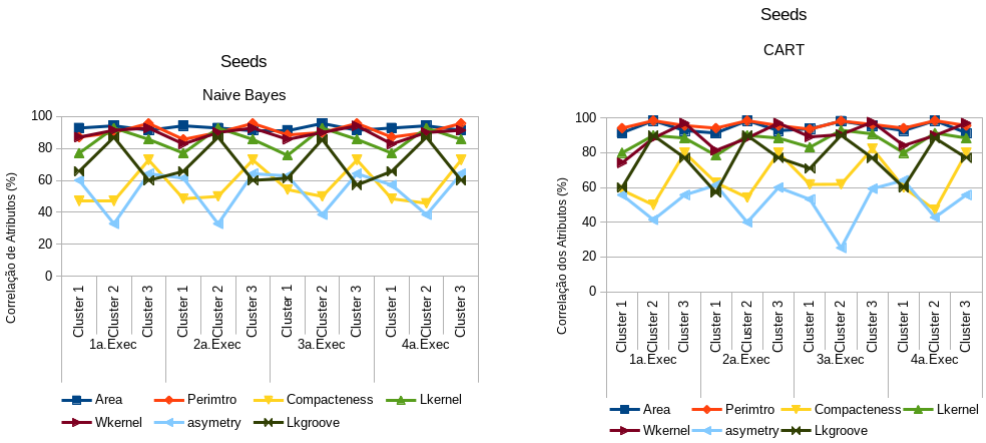
Dos dois algoritmos apresentados, um pertencente ao paradigma simbólico e o outro estatístico. Ademais cada execução desses algoritmos resultaram respostas satisfatórias no âmbito da rotulação de dados. Embora cada um tenham suas peculiaridades.

No modelo de resolução proposto foi inicialmente utilizado Naive Bayes na base de dados Seeds[1]. Logo o resultado mostrou-se bem confiável pois o método consegue mostrar através da tabela 5, os valores de correlação entre os atributos, também exibido na coluna **Relevância** da tabela 3. Dessa forma fica fácil identificar quais os atributos podem ser os rótulos dos clusters. Embora essa decisão possa ser modificada de acordo com o valor da variável $V$. Variável essa criada para melhor escolher os atributos do rótulo, e podendo assumir valores diferentes dependendo do comportamento da base de dados.

Continuando com a base Seeds, após a escolha do atributo que fará parte do rótulo, o segundo passo é a escolha da faixa de valores do atributo. Essa segunda etapa

[1] Sessão 4.2.1

é dependente totalmente da discretização[2] e independente da primeira etapa. O método é capaz de gerar a faixa de maior repetição de valores de qualquer atributo, mas aqui neste trabalho o que importa é a faixa do atributo rótulo. Para ter mais confiabilidade no rótulo o método escolhe a faixa de valores que mais se repetem. No caso desse algoritmo o resultado na tabela 3 consegue provar uma boa eficiência, pois em cada 70 elementos do cluster 1, somente 14, ficaram de fora dessa faixa. No cluster 2, somando os dois atributos rótulos tem-se 12 elementos que não estão dentro da representatividade do rótulo. Outro valor pequeno em relação aos 70 elementos. E no cluster 3, somente 5 elementos não estão dentro da faixa considerada rótulo.




(a) Naive Bayes (b) CART

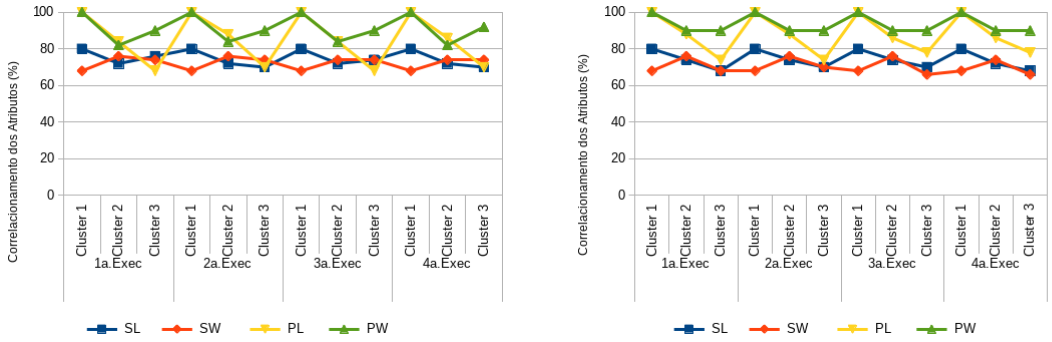
Figura 13 – Gráfico de Execuções dos algoritmos supervisionados na base de dados SEEDS.

No caso do algoritmo CART, os resultados foram diferentes dos apresentados pelo Naive Bayes, mas nem por isso foram insatisfatórios. Contudo uma breve análise sobre as execuções das tabelas 5 e 8 podem ser observadas nos gráficos da figura 13. Como já comentado anteriormente o comportamento dos valores do correlacionamento dos atributos ao longo das execuções mostra-se equilibrada, figura 13b. O gráfico do CART tem um movimento semelhante ao do aplicado do Naive Bayes 13a, embora a variável **asymetry** saia um pouco do padrão, mas como seus valores são baixos, nada alterou nos rótulos , contudo o valor de **perimetro** ficou bastante encostado ao valor da **area**, fazendo o rótulo **perimetro** aparecer nos grupos 1 e 2. E também só não foi escolhido pelo grupo 3 , pois a variável **Wkernel** estava com valor mais alto. E no gráfico percebe-se que **Wkernel** mantém valores altos em todas as execuções do grupo 3.

De acordo com o exposto acima pode-se dizer nesta análise, que o Naive Bayes acabou tendo resultados um pouco melhores, pois no que diz respeito ao número de elementos fora da faixa definida pelo rótulo, o CART, acabou por ter mais elementos fora da faixa de rótulo, comparando com os resultados do Naive Bayes. Quer dizer, o rótulo

[2] sessão 2.2

deixa de representar mais elementos usando o CART ao invés do Naive Bayes, ou em outras palavras, o Naive Bayes representou mais elementos que o CART.




(a) Naive Bayes (b) CART

Figura 14 – Gráfico de Execuções dos algoritmos supervisionados na base de dados IRIS.

Ja na base de dados Iris, os dois algoritmos supervisionados testados apresentaram os mesmos rótulos. Mantendo as mesmas configurações, $R = 3$, $V = 3\%$ e EFD na discretização. Nos gráficos da figura 14 pode-se acompanhar como os valores dos atributos se comportam em seus clusters nas 4 iterações.

Nesta base Iris os algoritmos tem resultados nos gráficos, bastantes semelhantes, e logo se percebe que essa base contém características que possuem mais coerência com a classe em relação ao da base SEEDS, pois nenhum atributo possui valor abaixo da linha 65(%) de relacionamento entre eles. Embora no gráfico as linhas referentes aos comportamentos dos atributos não sejam totalmentes iguais em cada algoritmo executado, não chegou a um valor que diferenciasse para modificar o resultado dos rótulos como resposta.

Na rotulação encontrada pelo dois algoritmos nos resultados da tabela 9 e 13 o rótulo escolhido no cluster 1 teve dois atributos, **petallength** e **petalwidth** e cada um deles definiram uma faixa onde foi possível abranger 100% de elementos dentro das faixas escolhidas por cada um dos atributos. Já no cluster 2 os mesmos atributos são escolhidos mas com faixas de valores diferentes. Embora não tivesse fechado os mesmos valores do cluster 1, obteve um total, de 7 elementos do **petallength** e 8 do **petalwidth**, totalizando 15 elementos que não estão dentro da faixa delimitada pelos rótulos. O cluster 2 possui um total de 50 elementos e 15 deles não são representados pelo rótulo do cluster. E no cluster 3 o atributo escolhido para compor o rótulo é mais uma vez o **petalwidth**. Logo se percebe a importância do atributo nos clusters, mas em nenhum deles a faixa é a mesma. Isso define bem o rótulo do cluster 3, pois o rótulo representa 45 elementos dentro do cluster, possuindo somente 5 elementos fora dessa faixa representada pelo atributo.

A repetição do atributo **petalwidth** em todos os rótulos, acaba mostrando o grau de relevância desse atributo na base de dados. Para um especialista é interessante saber

que esse atributo possue um grande referencial na base de dados. Nos rótulos esse atributo assumiu faixas diferentes conseguindo assim o método da um significado ao cluster.

Com uma breve análise já se constata que os resultados nos dois algoritmos supervisionados foram bem satisfatórios nas bases utilizadas, e provando que é possível a rotulação de dados. Além de conseguir representar bem os clusters através dos rótulos encontrados. E uma obsevação técnica dos algoritmos utilizados é que o CART se mostrou bem mais rápido em relação ao Naive Bayes para gerar os resultados.

## 5.2 Trabalhos Futuros

A pesquisa ainda precisa de mais divulgação na esfera acadêmica, e para isso a plublicação de um artigo sobre os resultados apresentados aqui é uma consolidação dessa proposta de mestrado já voltada para a dissertação propriamente dita.

Fazer testes com mais bases de dados e com isso traçar uma estratégia caso seja necessário a utilização da idéia de rotulação por algum fim. Caso algum órgão/setor/empresa precise utilizar a rotulação em seu meio, seria interessante o analista de dados, saber com quais algoritmos supervisionados ele obteria melhores resultados. Embora se saiba nesse estudo que quaisquer algoritmos supervisionados são capazes de realizar a rotulação de dados, também foi provado que em algumas bases um algoritmo se sobressai a outro. Por consequencia disso ter um maior número de base com características diferentes ajudaria em uma tomada de decisão.

Outro ponto importante é inserir nos teste mais algoritmos, que pertençam a paradigmas diferentes dos que ainda não foram utilizados. * de acordo com autor PEARSON, onde os paradigmas são divididos em : ..... pode-se perceber que existe uma proximidade na afirmação de rotulação para quaisquer algoritmo supervisionado, não sendo ainda possível afirmar esse tema, por falta de testes em alguns algoritmos de paradigmas ainda não testados, mas já deixando par atrabalhos futuros

## 5.3 Cronograma

Tabela 14 – Cronograma de atividades

| Atividades | Meses | | | | | |
|---|---|---|---|---|---|---|
| | Março | Abril | Maio | Junho | Julho | Agosto |
| Testes com Novas Bases de Dados | | | | | | |
| Modificar Números de Faixa (R) | | | | | | |
| Testar com outros Métodos de Discretização | | | | | | |
| Testar com outros Algoritmos com Paradigmas Diferentes | | | | | | |
| Preparar Artigo | | | | | | |
| Escrita da Dissertação | | | | | | |

# Referências


BARBER, D. *Bayesian Reasoning and Machine Learning*. [s.n.], 2011. ISSN 9780521518147. ISBN 9780511804779. Disponível em: <http://ebooks.cambridge.org/ref/id/CBO9780511804779>. Citado 2 vezes nas páginas 6 e 10.

CATLETT, J. *On changing continuous attributes into ordered discrete attributes*. Springer, Berlin, Heidelberg: Lecture Notes in Computer Science (Lecture Notes in Artificial Intelligence), 1991. 164–178 p. Citado na página 19.

DOUGHERTY, J.; KOHAVI, R.; SAHAMI, M. Supervised and Unsupervised Discretization of Continuous Features. *Machine Learning Proceedings 1995*, v. 0, p. 194–202, 1995. ISSN 0717-6163. Disponível em: <http://linkinghub.elsevier.com/retrieve/pii/B9781558603776500323>. Citado na página 10.

FILHO, V. P. R. *Rotulacao de grupos utilizando conjuntos fuzzy*. Tese (Doutorado) — Universidade Federal do Piauí, 2015. Citado 3 vezes nas páginas 15, 14 e 29.

HWANG, G. J.; LI, F. A Dynamic Method for Discretization of Continuous Attributes. *Lecture Notes in Computer Science - Intelligent Data Engineering and Automated Learning - IDEAL 2002: Third International Conference*, v. 2412/2002, p. 506, 2002. ISSN 16113349. Disponível em: <http://www.springerlink.com/content/4n05b2n6x0cx4tlk>. Citado na página 10.

KOTSIANTIS, S.; KANELLOPOULOS, D. Discretization Techniques : A recent survey. *GESTS International Transactions on Computer Science and Engineering*, v. 32, n. 1, p. 47–58, 2006. Citado na página 10.

KUMAR, A.; ANDU, T.; THANAMANI, A. S. Multidimensional Clustering Methods of Data Mining for Industrial Applications. *International Journal of Engineering Science Invention*, v. 2, n. 7, p. 1–8, 2013. Citado na página 1.

LIMA, B. V. A. Método Semissupervisionado de Rotulação e Classificação Utilizando Agrupamento por Sementes e Classificadores. 2015. Citado 2 vezes nas páginas 15 e 13.

LOPES, L. A. Dissertação (Mestrado em Ciências da Computação), *Rotulação Automática de Grupos com Aprendizagem de Máquina Supervisionada*. 2014. 73 p. Citado 10 vezes nas páginas 15, 1, 2, 11, 12, 13, 16, 18, 19 e 29.

LOPES, L. A.; MACHADO, V. P.; RABELO, R. D. A. L. Automatic Labeling of Groupings through Supervised Machine Learning. Citado na página 34.

LUCCA, G. et al. Uma implementação do algoritmo Naïve Bayes para classificação de texto. *Centro de Ciências Computacionais - Universidade Federal do Rio Grande (FURG) Rio Grande - RS - Brasil*, p. 1–4, 2013. Citado na página 9.

MADUREIRA, D. F. *Analise de sentimento para textos curtos*. Tese (Doutorado) — Fundacao Getulio Vargas, Rio de Janeiro, 2017. Citado na página 9.

MCCALLUM, A.; NIGAM, K. A Comparison of Event Models for Naive Bayes Text Classification. 1997. Citado na página 9.

MITCHELL, T. M. *Machine learning*. [S.l.]: McGraw-Hill Science/Engineering/Math, 1997. 432 p. ISSN 10450823. ISBN 9781577354260. Citado na página 5.

MONTGOMERY, K. *Big Data Now*. 1. ed. [S.l.]: O'Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472, 2013. v. 53. 1689–1699 p. ISSN 1098-6596. ISBN 9788578110796. Citado na página 1.

RAIMUNDO, L. R.; MATTOS, M. C. D.; WALESKA, P. O Algoritmo de Classificação CART em uma Ferramenta de Data Mining. 2008. Citado na página 8.

RUSSEL, S.; NORVIG, P. *Inteligência Artificial*. 3ª. ed. Rio de Janeiro: [s.n.], 2013. ISBN 9780136042594. Citado 2 vezes nas páginas 5 e 6.

WU, X. et al. *Top 10 algorithms in data mining*. [S.l.: s.n.], 2008. v. 14. 1–37 p. ISSN 02191377. ISBN 1011500701. Citado na página 9.