Aula Prática 7

Tema: Comparação de Classificadores

Murielly Oliveira Nascimento – 11921BSI222

Ferramentas Usadas

As análises foram feitas com o uso da ferramenta Weka. Os algoritmos usados para as análises foram K-vizinhos mais próximos com K=3 e K=6; Árvore de Decisão com o mínimo número de instâncias em cada folha igual a 2 e 4; e BayesNet com algoritmos de busca K2 e TAN, a explicação deste se encontra na próxima seção. Para a divisão da base de dados em treino e teste foi usada a estratégia Cross-Validation com K=10.

As medidas de avaliação usadas foram Precisão, Revocação e Medida F. Todas são baseadas na matriz de confusão:

		Classe Verdadeira	
		Positivo	Negativo
Classe Prevista	Positivo	Verdadeiro Positivo (VP)	Falso Positivo (FP)
	Negativo	Falso Negativo (FN)	Verdadeiro Negativo (VN)

Medida F = $\frac{2}{\frac{1}{prec}+\frac{1}{rev}}$

Precisão = $\frac{VP}{VP+FP}$

Revocação = $\frac{VP}{VP+FN}$

As Bases de Dados escolhidas para testes foram Breast Cancer Data Set e Space Shuttle Autolanding Domain, que podem ser encontradas nos seguintes links:

https://archive.ics.uci.edu/ml/datasets/breast+cancer

https://archive.ics.uci.edu/ml/datasets/Shuttle+Landing+Control

O Algoritmo BayesNet

> *A principal vantagem de raciocínio probabilístico sobre raciocínio lógico é o fato de que agentes podem tomar decisões racionais mesmo quando não existe informação suficiente para se provar que uma ação funcionará (*RUSSEL)

Grafos conseguem representar relações causais entre eventos. Uma Rede Bayesiana leva em consideração a dependência dos atributos entre si com o uso de grafos acíclicos. Assim cada nó são instâncias e as arestas dependências condicionais, se dois nós não estão conectados então eles são independentes.

Um possível método para construção das redes bayesianas é como segue:

Escolha *um conjunto de variáveis Xi que descrevam o domínio*

Escolha *uma ordem para as variáveis*

Enquanto *existir variáveis*

Escolha *uma variável Xi e adicione um nó a rede*

Determine *os nós Pais(Xi) dentre os nós que já estejam na rede e que tenham influência direta em Xi*

Defina *a tabela de probabilidades condicionais para Xi.*

De acordo com a professora Inês Dutra da UFRJ, Redes Bayesianas constituem uma forma natural para representação de informações condicionalmente independentes. É uma boa solução para problemas onde conclusões não podem ser obtidas apenas do domínio do problema.

> *Como resultado, vários algoritmos especializados em realizar a pesquisa em diferentes tipos de topologias de grafos acíclicos direcionados (DAG) foram desenvolvidos, sendo a maioria deles extensões (usando arcos de aumento) ou modificações de a topologia básica Naive Bayes. Esta abordagem geralmente obtém resultados mais satisfatórios. (ACID, 2005)*

Na ferramenta WEKA é possível escolher os algoritmos de busca a serem usados no classificador NetBayes. Foram usados o K2 e o TAN. O primeiro, de acordo com a professora Carolina Ruiz, busca heuristicamente a mais provável estrutura de rede de crenças dado uma base de dados. O segundo é um acrônimo para Tree Augmented Naive Bayes. É um método de aprendizado bayesiano semi-ingênuo. Ele relaxa a suposição de independência do atributo ingênuo de Bayes, empregando uma estrutura de árvore, na qual cada atributo depende apenas da classe e de um outro atributo. Uma árvore de abrangência ponderada máxima que maximiza a probabilidade dos dados de treinamento é usada para realizar a classificação (ZHENG, 2011)

Análise da Base Breast Cancer

Informações

A base Breast Cancer é composta por dados coletados da University Medical Centre, Instituto de Oncology, Ljubljana, Yugoslavia. Tem um total de 286 instâncias, sendo 201 da classe câncer de mama *recorrente* e 85 da *não recorrente.* Há a presença de valores ausentes. São 9 atributos do tipo nominal descritos da seguinte forma:

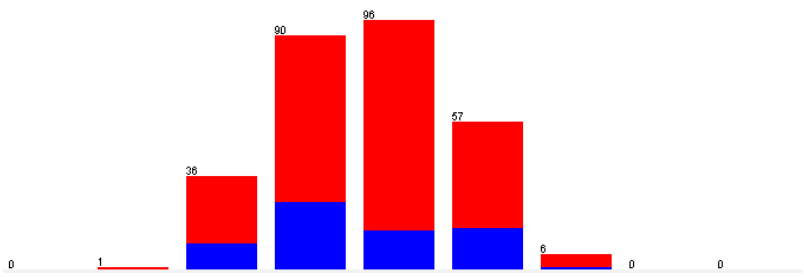
No.	Name
1	class
2	age
3	menopause
4	tumor-size
5	inv-nodes
6	node-caps
7	deg-malig
8	breast
9	breast-quad
10	irradiat

- **Idade**: idade do paciente no momento do diagnóstico que cobre os seguintes intervalos (10-19, 20-29, 30-39, 40-49, 50-59, 60-69, 70-79, 80-89, 90-99.)

Selected attribute
Name: age
Missing: 0 (0%)
Distinct: 6
Type: Nominal
Unique: 1 (0%)

No.	Label	Count	Weight
1	10-19	0	0
2	20-29	1	1
3	30-39	36	36
4	40-49	90	90
5	50-59	96	96
6	60-69	57	57
7	70-79	6	6
8	80-89	0	0
9	90-99	0	0

Class: irradiat (Nom)
Visualize All



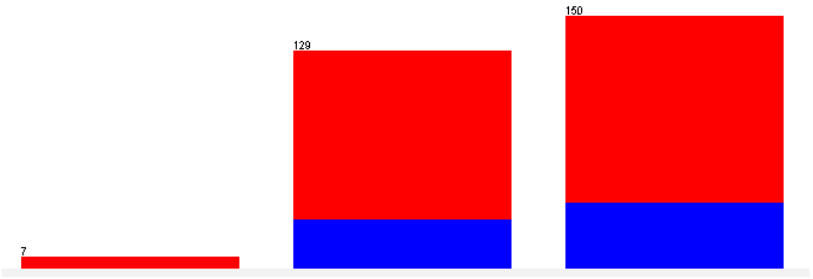
- **Menopausa**: se a paciente está na pré ou pós-menopausa no momento do diagnóstico, descrito da seguinte forma: lt40, ge40, premeno).

Selected attribute

Name: menopause Type: Nominal
Missing: 0 (0%) Distinct: 3 Unique: 0 (0%)

No.	Label	Count	Weight
1	lt40	7	7
2	ge40	129	129
3	premeno	150	150

Class: irradiat (Nom) Visualize All



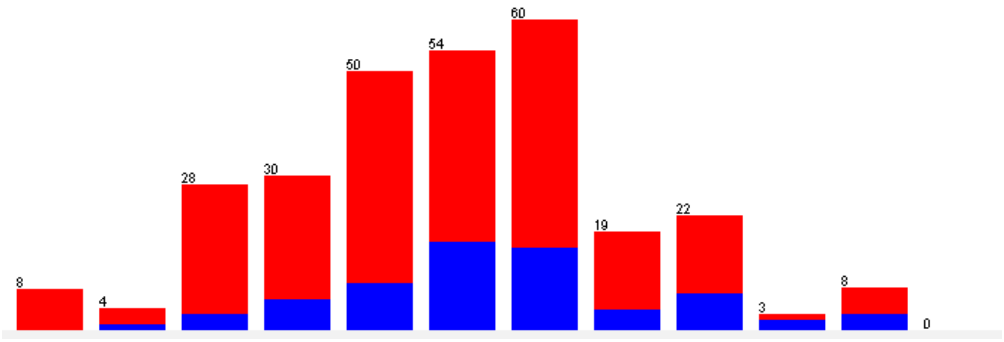
- **Tamanho do tumor**: o maior diâmetro (em mm) do tumor excisado, divido por intervalos (0-4, 5-9, 10-14, 15-19, 20-24, 25-29, 30-34, 35-39, 40-44, 45-49, 50-54, 55-59).

Selected attribute

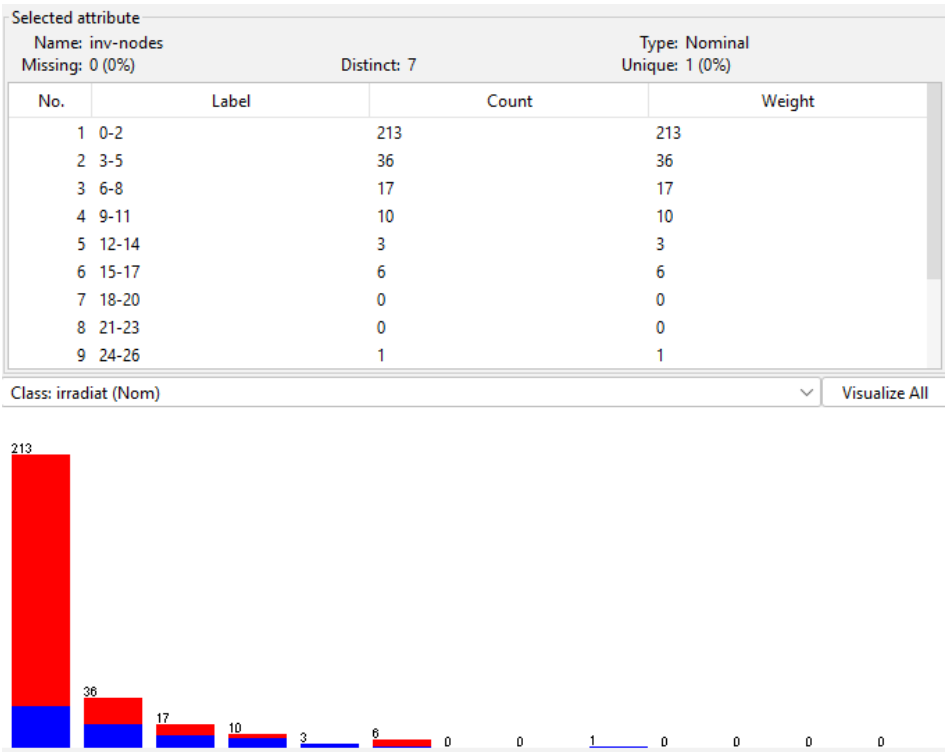
Name: tumor-size Type: Nominal
Missing: 0 (0%) Distinct: 11 Unique: 0 (0%)

No.	Label	Count	Weight
1	0-4	8	8
2	5-9	4	4
3	10-14	28	28
4	15-19	30	30
5	20-24	50	50
6	25-29	54	54
7	30-34	60	60
8	35-39	19	19
9	40-44	22	22

Class: irradiat (Nom) Visualize All



- **Inv-nodes**: o número (intervalo de 0 - 39) de linfonodos axilares que contêm câncer de mama metastático visível no exame histológico (0-2, 3-5, 6-8, 9-11, 12-14, 15-17, 18-20, 21-23, 24-26, 27-29, 30-32, 33-35, 36-39).



- **Caps de linfonodo**: se o câncer metástase para um linfonodo, embora fora do local original do tumor, ele pode permanecer “contido” pela cápsula do linfonodo. Porém, com o tempo, e com doença mais agressiva, o tumor pode substituir o linfonodo e então penetrar na cápsula, permitindo que invada os tecidos circundantes. (sim ou não)

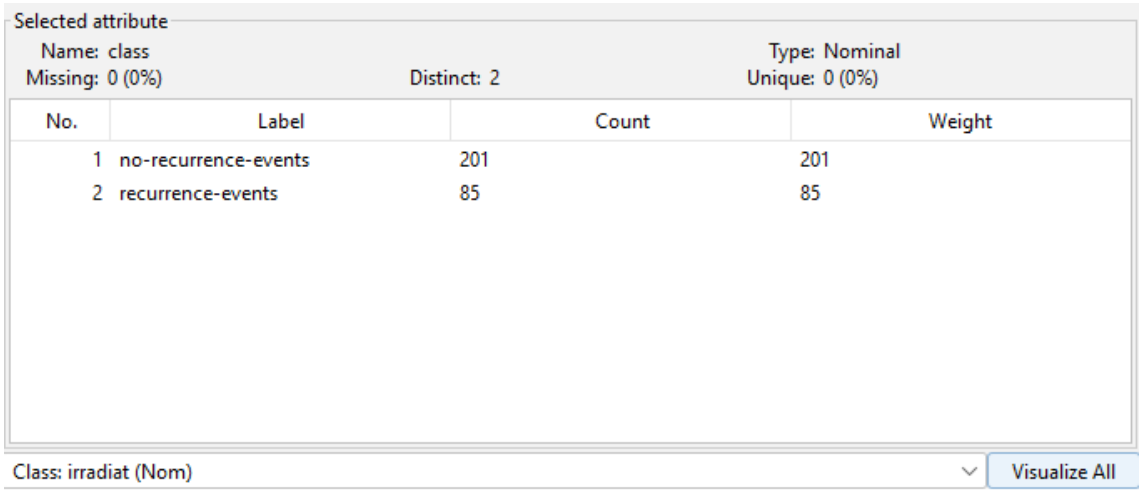
- **Grau de malignidade**: o grau histológico (intervalo 1-3) do tumor. Os tumores de grau 1 consistem predominantemente em células que, embora neoplásicas, retêm muitas de suas características usuais. Os tumores de grau 3 consistem predominantemente em células altamente anormais. (1, 2, 3.)

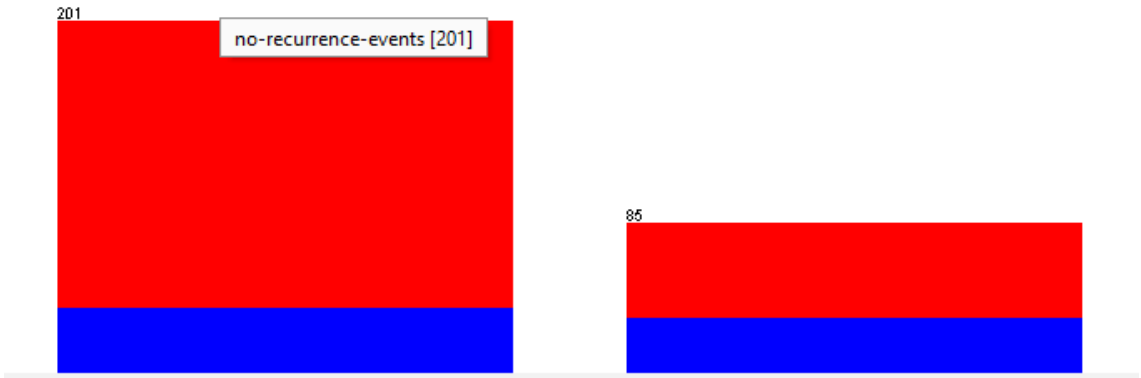
- **Mama**: o câncer de mama pode obviamente ocorrer em qualquer uma das mamas. (direita ou esquerda)

- **Quadrante da mama**: a mama pode ser dividida em quatro quadrantes, tendo como ponto central o mamilo (esquerda cima, esquerda baixo, direita cima, direita baixo, central).

- **Irradiação**: a radioterapia é um tratamento que utiliza raios-x de alta energia para destruir as células cancerígenas (sim ou não).

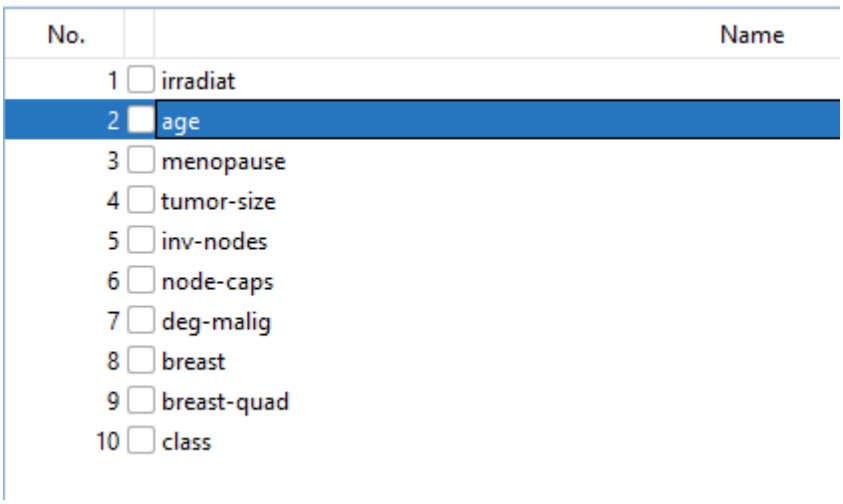
- **Classe**: A classificação é dada por recorrente e não recorrente.





Pré-processamento

O atributo classe é o primeiro na base de dados. Para a ferramenta Weka reconhecê-lo como a *classe* usei o filtro Reorder para colocá-la como o último atributo. Os dados da base são categóricos e precisam ser convertidos em numéricos para análises. O filtro usado para a conversão foi OrdinalToNumeric. Por fim os valores ausentes foram substituídos pelas médias dos respectivos atributos usando o filtro ReplaceMissingValues.



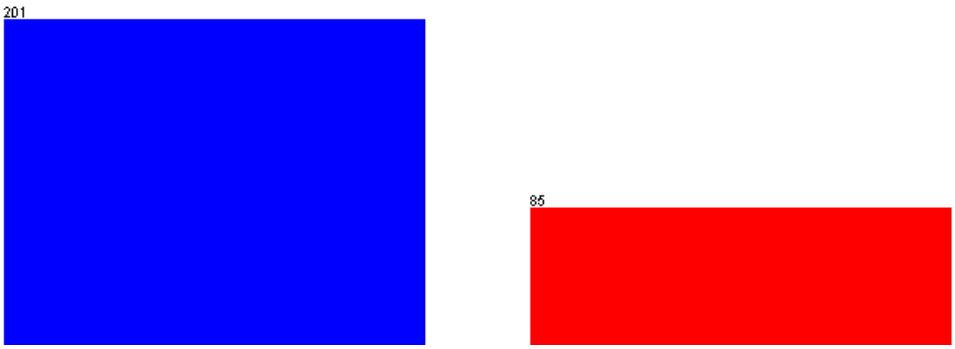
- Classe

Selected attribute

Name: class Type: Nominal
Missing: 0 (0%) Distinct: 2 Unique: 0 (0%)

No.	Label	Count	Weight
1	no-recurrence-events	201	201
2	recurrence-events	85	85

Class: class (Nom) Visualize All



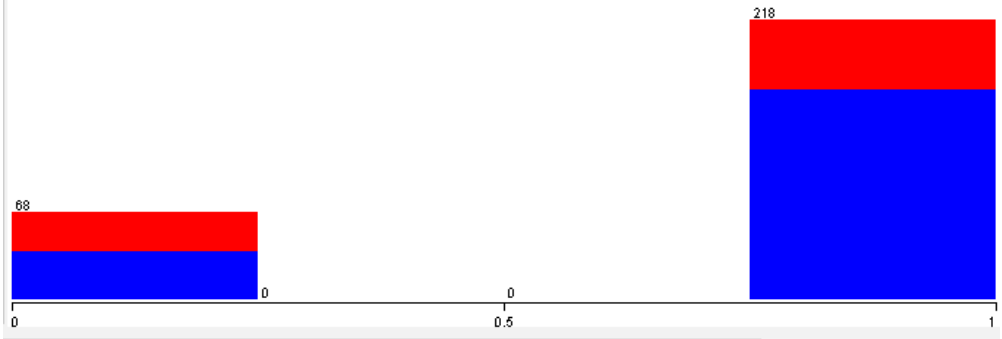
- Irradiação

Selected attribute

Name: irradiat Type: Numeric
Missing: 0 (0%) Distinct: 2 Unique: 0 (0%)

Statistic	Value
Minimum	0
Maximum	1
Mean	0.762
StdDev	0.426

Class: class (Nom) Visualize All



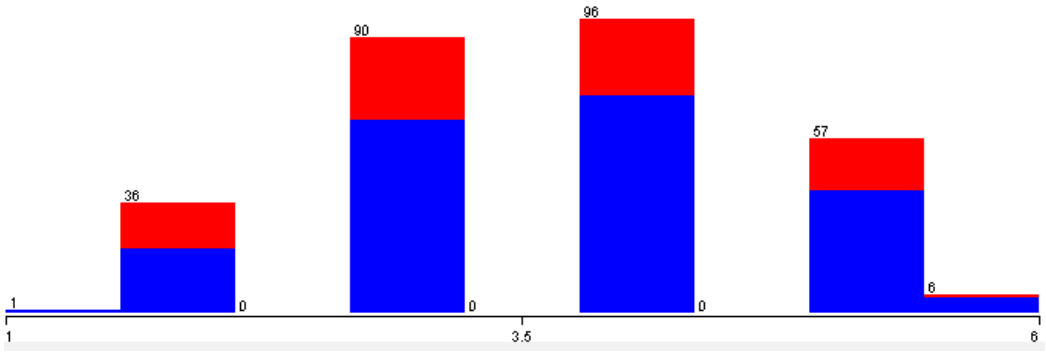
- Idade

Selected attribute

Name: age
Missing: 0 (0%)
Distinct: 6
Type: Numeric
Unique: 1 (0%)

Statistic	Value
Minimum	1
Maximum	6
Mean	3.664
StdDev	1.012

Class: class (Nom) Visualize All



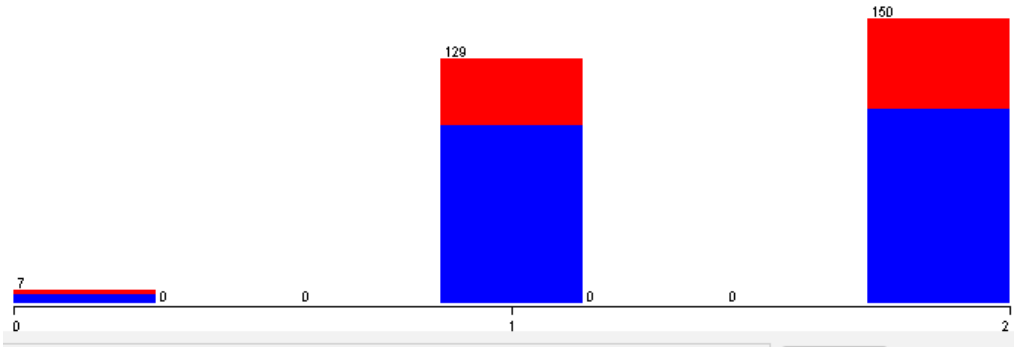
- Menopausa

Selected attribute

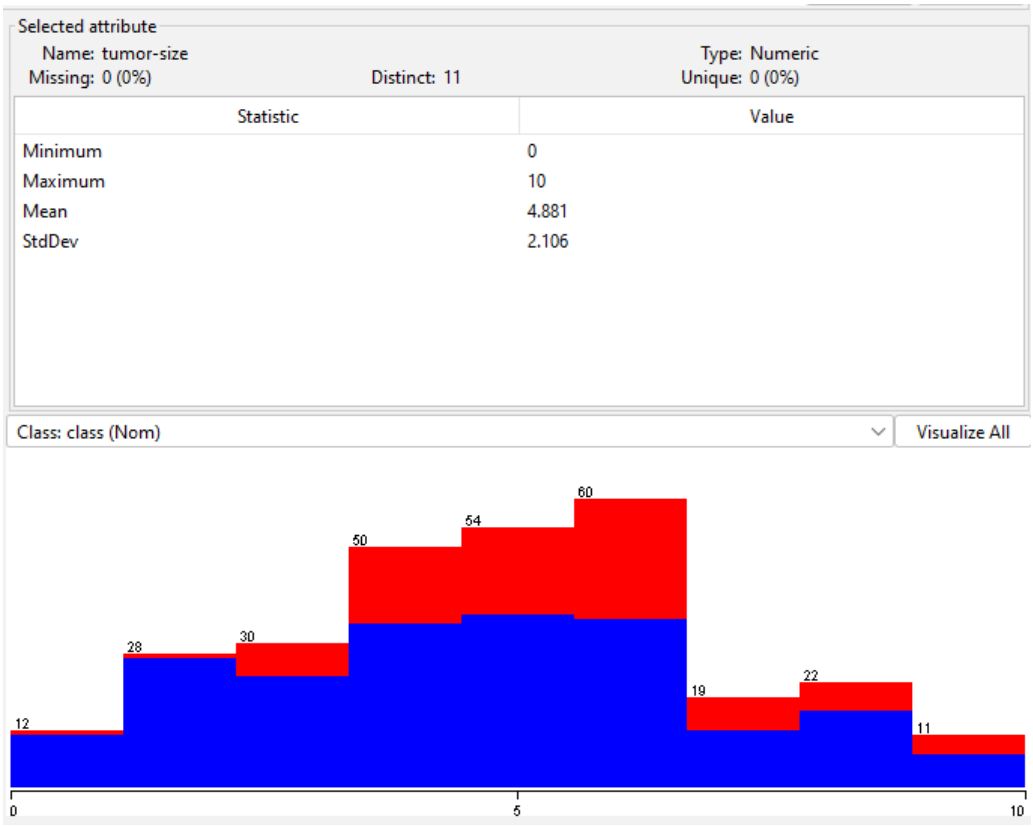
Name: menopause
Missing: 0 (0%)
Distinct: 3
Type: Numeric
Unique: 0 (0%)

Statistic	Value
Minimum	0
Maximum	2
Mean	1.5
StdDev	0.548

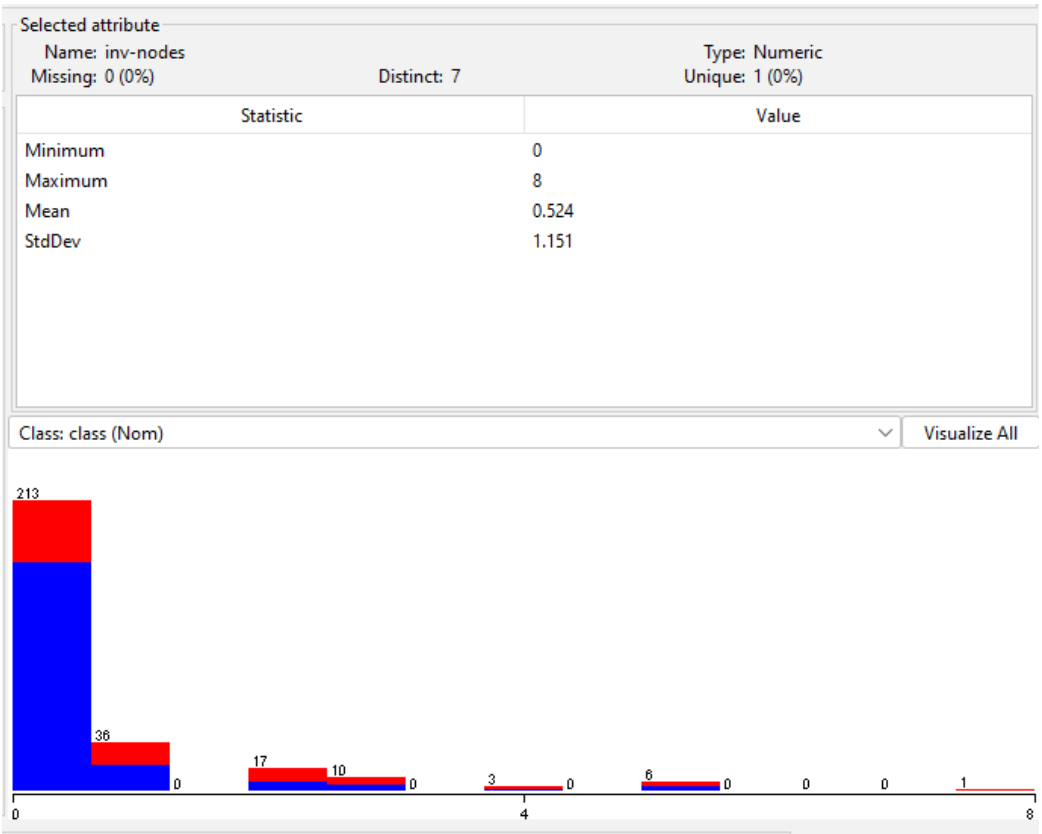
Class: class (Nom) Visualize All



- Tamanho do Tumor



- Inv-Nodes



- Caps de linfonodo

- Grau de malignidade

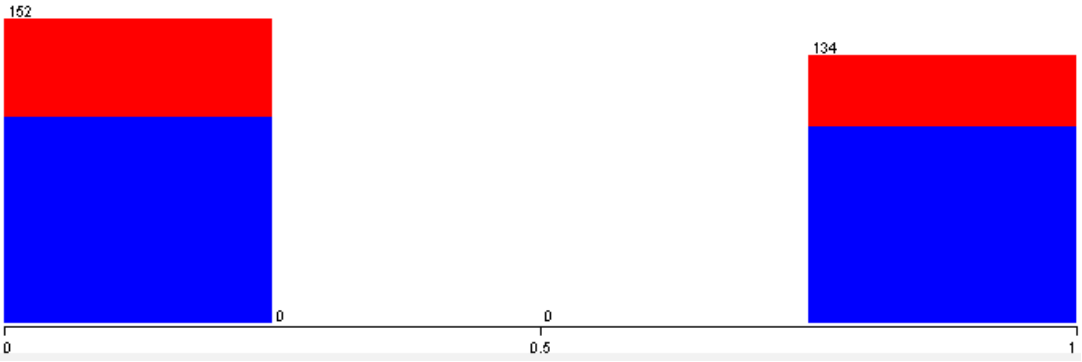
- Mama

Selected attribute

Name: breast | Type: Numeric
Missing: 0 (0%) | Distinct: 2 | Unique: 0 (0%)

Statistic	Value
Minimum	0
Maximum	1
Mean	0.469
StdDev	0.5

Class: class (Nom) | Visualize All



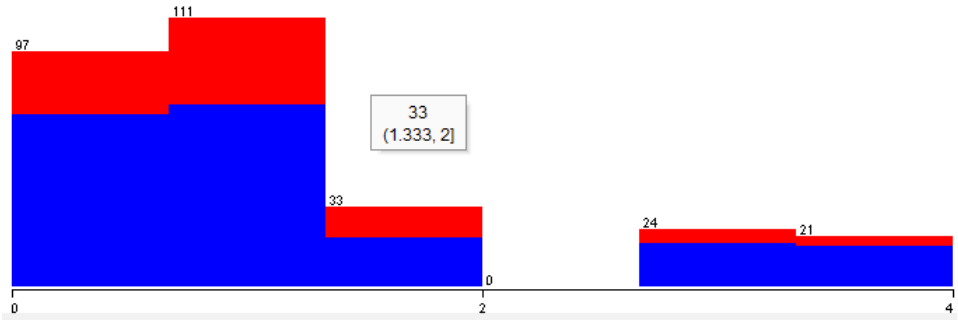
- Quadrante da mama

Selected attribute

Name: breast-quad | Type: Numeric
Missing: 0 (0%) | Distinct: 6 | Unique: 1 (0%)

Statistic	Value
Minimum	0
Maximum	4
Mean	1.165
StdDev	1.195

Class: class (Nom) | Visualize All



Avaliação

Para todos os casos estou considerando a classe Recorrentes como positiva, por ter o menor número de instâncias e a classe Não Recorrente como negativa. A ferramenta Weka faz o cálculo de considerando cada uma das classes, porém para as análises feitas nesse trabalho usei a lógica acima.

A matriz de confusão do Algoritmo K-vizinhos mais próximos é como segue:

- Com K = 3

```
=== Confusion Matrix ===

   a   b   <-- classified as
 164  37 |   a = no-recurrence-events
  58  27 |   b = recurrence-events
```

- Com K = 6

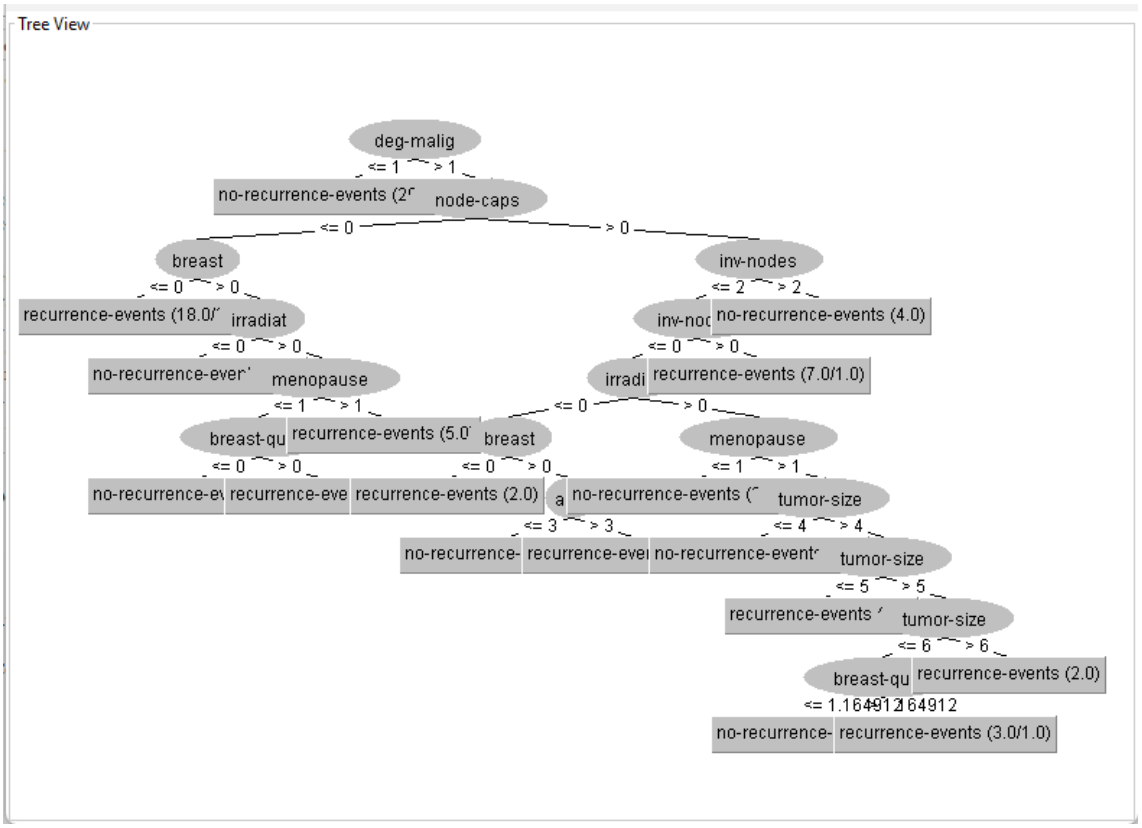
```
=== Confusion Matrix ===

   a   b   <-- classified as
 187  14 |   a = no-recurrence-events
  67  18 |   b = recurrence-events
```

	K = 3	K = 6
Precisão	0,422	0,563
Revocação	0,318	0,212
Medida F	0,362	0,308

A matriz de confusão do Algoritmo Árvore de Decisão e a árvore gerada são como seguem:

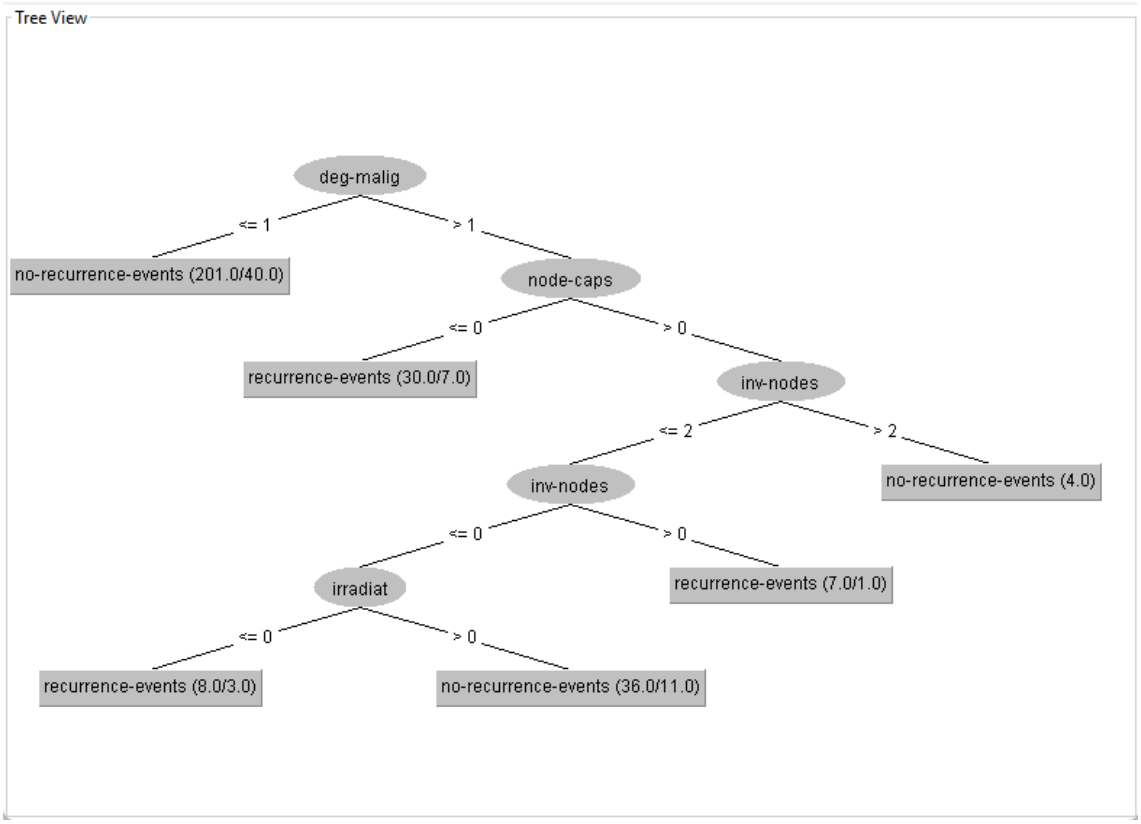
- Com número de instâncias por folha igual a 2

```
=== Confusion Matrix ===

   a   b   <-- classified as
 181  20 |   a = no-recurrence-events
  64  21 |   b = recurrence-events
```



- Com número de instâncias por folha igual a 4

```
=== Confusion Matrix ===

   a   b   <-- classified as
 178  23 |   a = no-recurrence-events
  62  23 |   b = recurrence-events
```

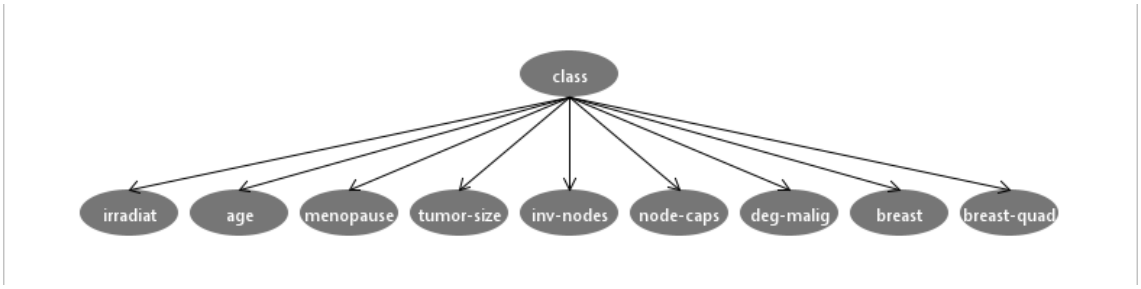


	N = 2	N = 4
Precisão	0,512	0,500
Revocação	0,247	0,271
Medida F	0,333	0,351

A matriz de confusão e o grafo gerado pelo algoritmo BayesNet são como seguem:

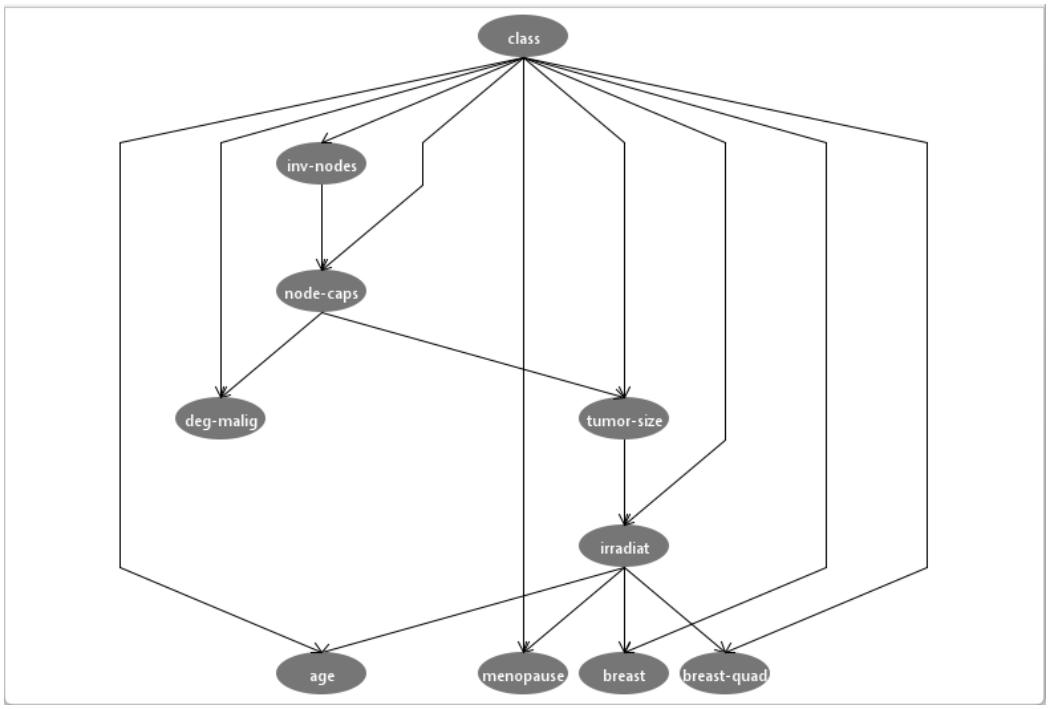
- Com algoritmo de busca K2

```
=== Confusion Matrix ===

   a   b   <-- classified as
 175  26 |   a = no-recurrence-events
  54  31 |   b = recurrence-events
```



- Com algoritmo de busca TAN

```
=== Confusion Matrix ===

   a   b   <-- classified as
 189  12 |   a = no-recurrence-events
  58  27 |   b = recurrence-events
```

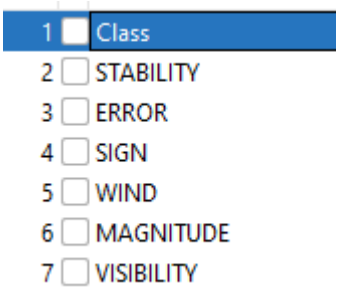


	K2	TAN
Precisão	0,544	0,692
Revocação	0,365	0,318
Medida F	0,437	0,435

Análise da Base de Dados Shuttle Landing Control

Informações

A Base de Dados Shuttle Landing Control foi criada artificialmente por Roger Burke e sua equipe na NASA para avaliar sobre quais condições o pouso automático seria melhor do que o manual. Trata-se de uma base de dados pequena com 15 instâncias, valores ausentes e 6 atributos do tipo categórico descritos a seguir:

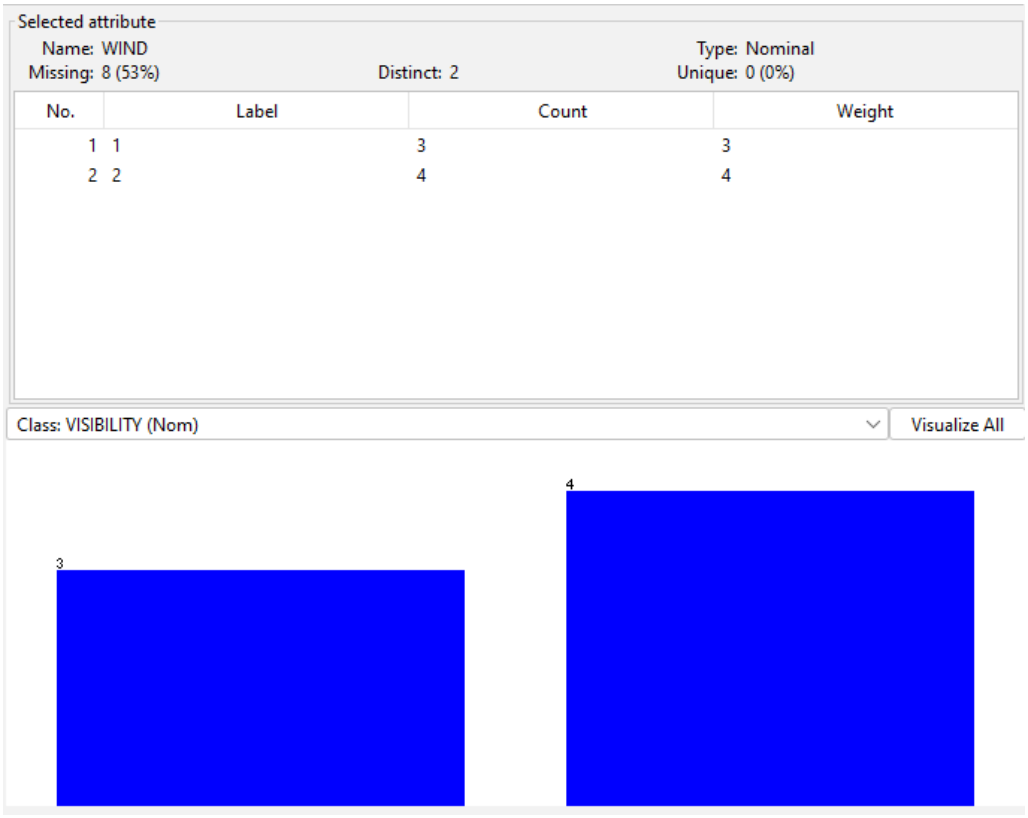


- **Estabilidade**: estável e muito estável (1 e 2)

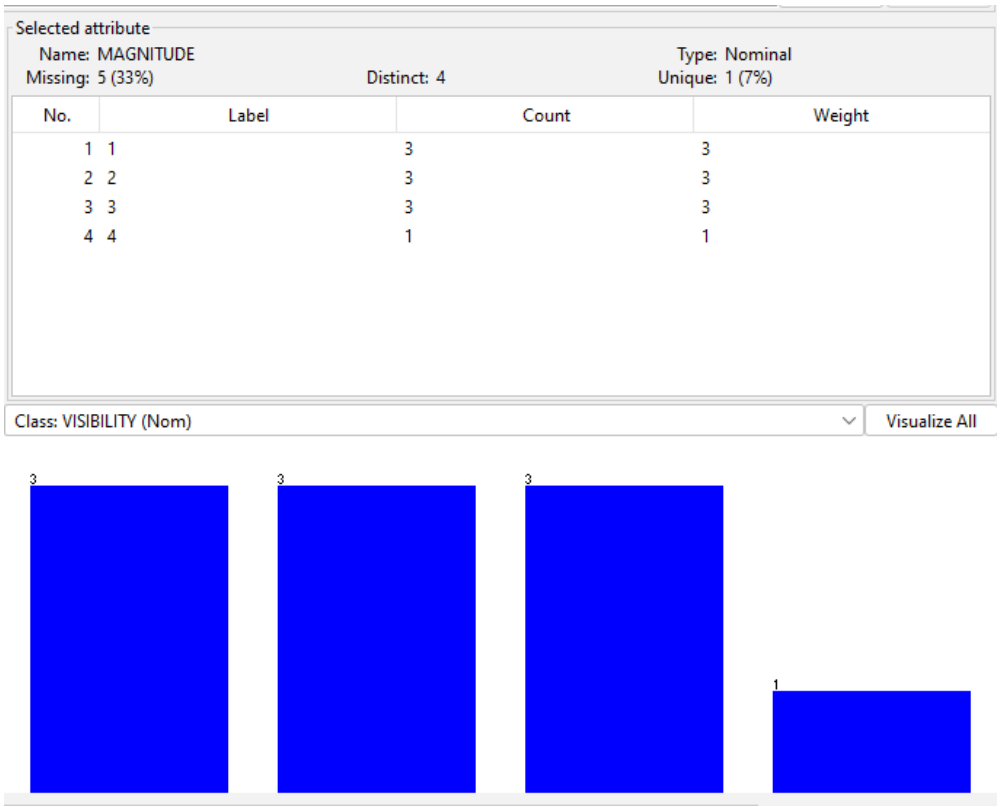
- **Erro**: tamanho do erro de pequeno a grande (1 a 4)

- **Sinal**: positivo ou negativo (1 e 2)

- **Vento**: Cabeça ou cauda (1 e 2)

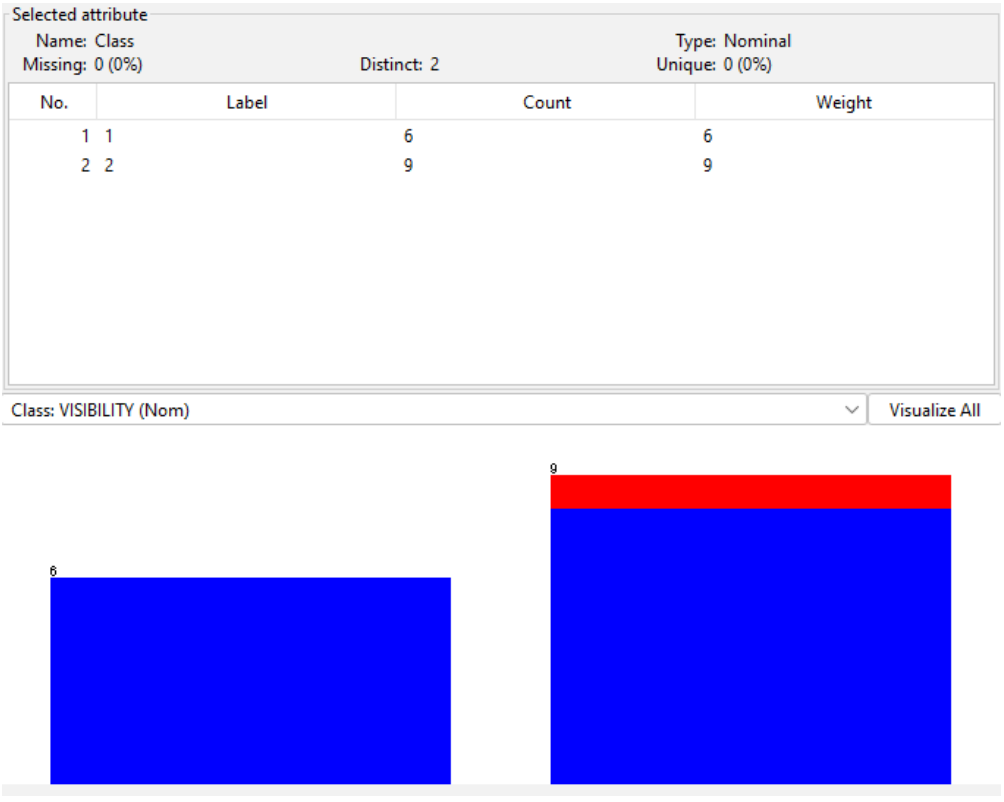


- **Magnitude**: Força do vento (1), leve (2), média (3), forte(4)



- **Visibilidade**: Sim ou Não

- **Classe**: dividida em manual (1) e automático (2).



Pré-processamento

Assim como na base Breast-Cancer é preciso reordenar os atributos, transformá-los em numéricos e tratar os valores ausentes. Novamente os filtros Reorder, OrdinalToNumeric e ReplaceMissingValues foram usados.

No.	Name
1	VISIBILITY
2	STABILITY
3	ERROR
4	SIGN
5	WIND
6	MAGNITUDE
7	Class

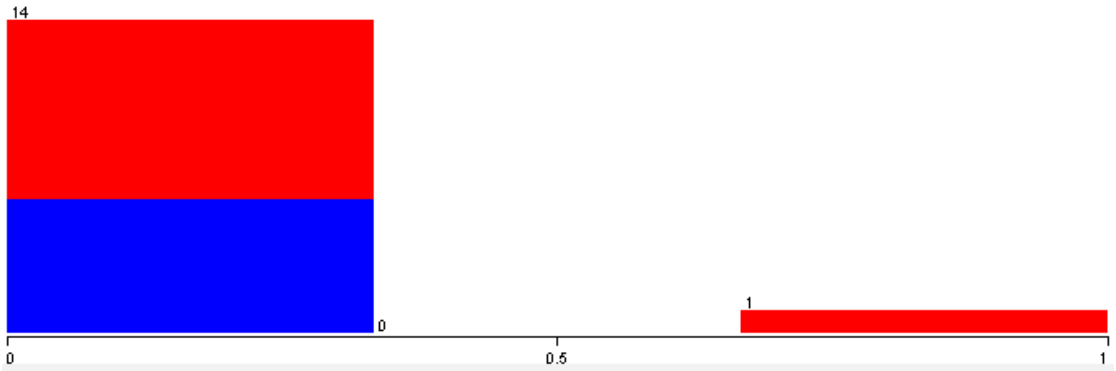
- **Visibilidade**

Selected attribute

Name: VISIBILITY Type: Numeric
Missing: 0 (0%) Distinct: 2 Unique: 1 (7%)

Statistic	Value
Minimum	0
Maximum	1
Mean	0.067
StdDev	0.258

Class: Class (Nom) Visualize All



- **Estabilidade**

- **Erro**

- **Sinal**


Selected attribute
Name: SIGN
Missing: 0 (0%)
Distinct: 3
Type: Numeric
Unique: 1 (7%)
Statistic
Value
Minimum
0
Maximum
1
Mean
0.143
StdDev
0.247
Class: Class (Nom)
Visualize All
14
0
1
0
0.5
1

- **Vento**


Selected attribute
Name: WIND
Missing: 0 (0%)
Distinct: 3
Type: Numeric
Unique: 0 (0%)
Statistic
Value
Minimum
0
Maximum
1
Mean
0.571
StdDev
0.35
Class: Class (Nom)
Visualize All
12
3
0
0.5
1

- **Magnitude**

- **Classe**

Avaliação

Nesse caso, estou considerando a classe Manual, 6 instâncias, como positivo e Automático, 9 instâncias, como negativo.

A matriz de confusão do Algoritmo K-vizinhos mais próximos é como segue:

- Com K = 3

```
=== Confusion Matrix ===

 a b   <-- classified as
 0 6 | a = 1
 0 9 | b = 2
```

- Com K = 6

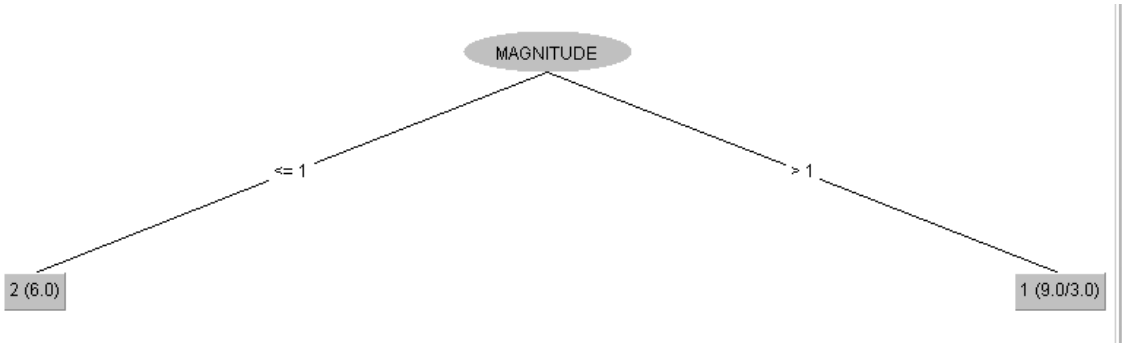
```
=== Confusion Matrix ===

 a b   <-- classified as
 0 6 | a = 1
 0 9 | b = 2
```

	K = 3	K = 6
Precisão	-	-
Revocação	0	0
Medida F	-	-

A matriz de confusão e a árvore gerada com o Algoritmo Árvore de Decisão são como seguem:

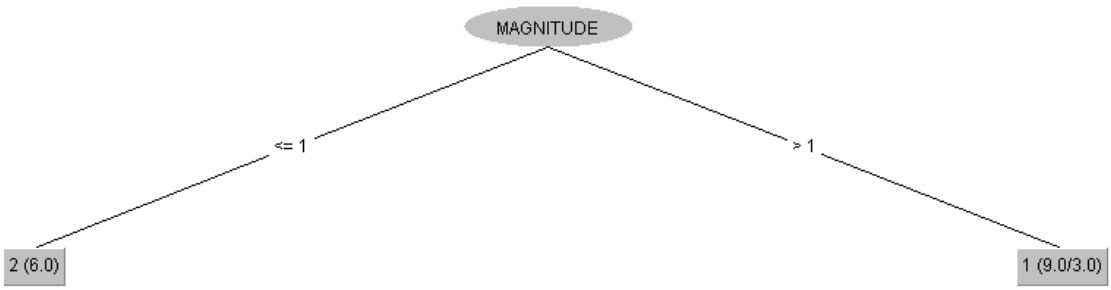
- Com número de instâncias por folha igual a 2.

```
=== Confusion Matrix ===

 a b   <-- classified as
 2 4 | a = 1
 3 6 | b = 2
```



- Com número de instâncias por folha igual a 4

```
=== Confusion Matrix ===

 a b   <-- classified as
 6 0 | a = 1
 3 6 | b = 2
```

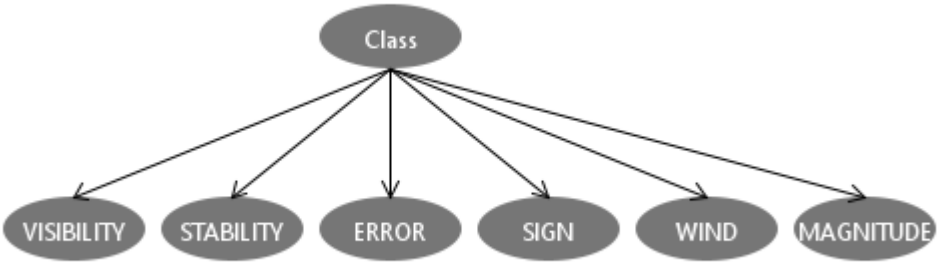


	N = 2	N = 4
Precisão	0,400	0,667
Revocação	0,333	1
Medida F	0,364	0,800

A matriz de confusão e o grafo gerados pelo algoritmo BayesNet são como seguem:

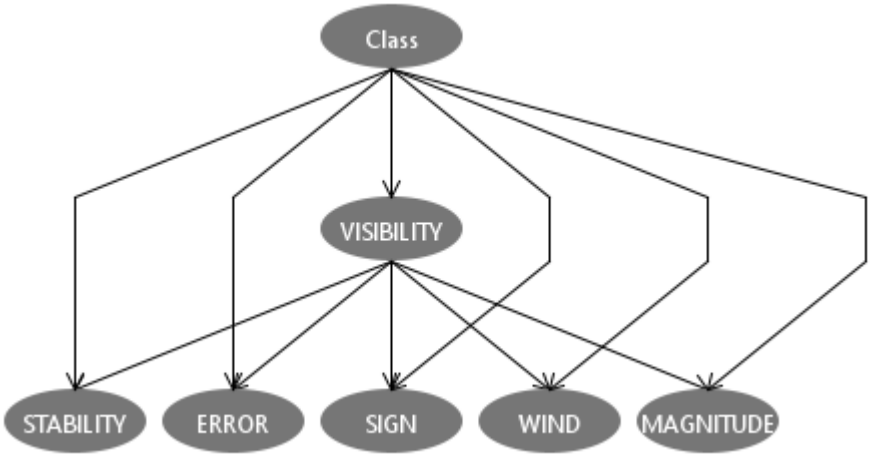
- Com algoritmo de busca K2

```
=== Confusion Matrix ===

 a b   <-- classified as
 1 5 | a = 1
 3 6 | b = 2
```



- Com o algoritmo de busca TAN

```
=== Confusion Matrix ===

 a b   <-- classified as
 1 5 | a = 1
 3 6 | b = 2
```



	K2	TAN
Precisão	0,250	0,250
Revocação	0,167	0,167
Medida F	0,200	0,200

Comparação dos Algoritmos

Neste trabalho foram usadas duas bases de dados de tamanhos diferentes. A primeira Breast Cancer tinha 286 instâncias e a Shuttle Landing Control 15 instâncias. Para avaliar qual algoritmo teve o melhor desempenho analisamos os resultados que cada um teve nas duas bases de dados. Como duas

configurações dos algoritmos foram usadas estou considerando essas para as análises finais: K=6 para o Algoritmo K-Vizinhos mais próximos, N=4 para a Árvore de Decisão, e o algoritmo de busca TAN para NetBayes.

- Precisão

	Breast-Cancer	Shuttle Landing Control
K vizinhos mais próximos	0,563	-
Árvore de Decisão	0,500	0,667
NetBayes	0,692	0,250

- Revocação

	Breast-Cancer	Shuttle Landing Control
K vizinhos mais próximos	0,212	0
Árvore de Decisão	0,271	1
NetBayes	0,318	0,167

- Medida F

	Breast-Cancer	Shuttle Landing Control
K vizinhos mais próximos	0,308	-
Árvore de Decisão	0,351	0,800
NetBayes	0,435	0,200

O Algoritmo com maior precisão, revocação e Medida F, na base de dados Breast Cancer é o NetBayes e na Shuttle Landing Control é a Árvore de Decisão. Podemos inferir que o algoritmo NetBayes possui uma performance melhor em bases de dados grandes, enquanto a Árvore de Decisão tem uma performance melhor em bases de dados pequenas.

Referências Bibliográficas

Russel, J. Stuart & Norvig, Peter. “Artifical Intelligence: A modern Approach”. Prentice Hall.

ACID, S. “Learning Bayesian Network Classifiers: Searching in a Space of Partially Directed Acyclic Graphs”. Springer Science 2005

RUIZ, C. “Illustration of the K2 Algorithm for Learning Bayes Net Structures”. Department of Computer Science, WPI.

Zheng, F., Webb, G.I. (2011). Tree Augmented Naive Bayes. In: Sammut, C., Webb, G.I. (eds) Encyclopedia of Machine Learning. Springer, Boston, MA. https://doi.org/10.1007/978-0-387-30164-8_850