Universidade Federal de Viçosa
Campus Rio Paranaíba

Aldo Henrique Dias Mendes
Matrícula 1610

Rafael Martins
Matrícula 1566

Michel Junio Ferreira Rosa
Matrícula 1540

Mineração de Dados

Classificação de Dados

Rio Paranaíba, Minas Gerais
06/2014

Sumário

1. Base de dados estação UFV-CRP

Base de dados gerada pela estação climática localizada no campus de Rio Paranaíba da Universidade Federal de Viçosa, está capturando dados desde 13/06/11 e a captura dos dados acontece de hora em hora. Até a data 04/06/2014 são aproximadamente 25000 objetos e 62 atributos + 1 classe que foi criada.

Entre os 62 atributos foram escolhidos alguns em específicos que são: Umidade, Velocidade do vento, Chuva, Radiação, Evapotranspiração. A escolha ocorreu através da referência bibliográfica e baseado em gráficos que mostraram que estes atributos são relevantes para determina o atributo escolhido como classe que foi a Temperatura.

Os objetos que possuem algum atributo faltante foram simplesmente retirado da base, uma vez que ao substituir o dado faltante por 0, significaria uma grande perda nos resultados finais.

Como este atributo está na forma de valores contínuos foi necessário transformá-lo em um atributo nominal, dessa forma foi utilizado um algoritmo para conversão representado da seguinte forma:

```
if(temperatura<18)
   classe="Frio";
if(temperatura>=18 && temperatura<25)
   classe="Normal";
if(temperatura>=25)
   classe="Quente";
```

Para classificação foi utilizado o *software* WEKA, e os algoritmos J48 que é a implementação da arvore de decisão e o algoritmo NaiveBayesSimple que é a implementação do Bayesiano Simples.

Do dia 04/06/2014 até o dia 15/06/2014, foram capturado dados para serem classificados nos algoritmos treinados com os dados anteriores, entre esses dados não possuem dados faltantes, o que ajuda na sua utilização.

1.1 Resultados

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances       13416               54.051  %
Incorrectly Classified Instances     11405               45.949  %
Kappa statistic                          0.3258
Mean absolute error                      0.3097
Root mean squared error                  0.5075
Relative absolute error                 76.0315 %
Root relative squared error            112.4475 %
Total Number of Instances            24821

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.923     0.433      0.48      0.923     0.632      0.774    Frio
                 0.26      0.134      0.672     0.26      0.375      0.61     Normal
                 0.694     0.113      0.58      0.694     0.632      0.905    Quente
Weighted Avg.    0.541     0.221      0.597     0.541     0.5        0.714

=== Confusion Matrix ===

    a    b    c   <-- classified as
 6932  552   29 |    a = Frio
 7166 3317 2260 |    b = Normal
  329 1069 3167 |    c = Quente
```

Bayesiano NaiveBayesSimple utilizando **Cross-validation = 10.**

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances       16553               66.6895 %
Incorrectly Classified Instances      8268               33.3105 %
Kappa statistic                          0.4622
Mean absolute error                      0.2679
Root mean squared error                  0.3828
Relative absolute error                 65.7656 %
Root relative squared error             84.8284 %
Total Number of Instances            24821

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.682     0.208      0.587     0.682     0.631      0.83     Frio
                 0.636     0.296      0.694     0.636     0.664      0.726    Normal
                 0.729     0.054      0.752     0.729     0.74       0.93     Quente
Weighted Avg.    0.667     0.225      0.672     0.667     0.668      0.795

=== Confusion Matrix ===

    a    b    c   <-- classified as
 5121 2367   25 |    a = Frio
 3566 8105 1072 |    b = Normal
   34 1204 3327 |    c = Quente
```

Árvore J48 utilizando **Cross-validation = 10.**

```
=== Evaluation on test set ===
=== Summary ===

Correctly Classified Instances         203               73.8182 %
Incorrectly Classified Instances        72               26.1818 %
Kappa statistic                          0.5474
Mean absolute error                      0.2653
Root mean squared error                  0.3696
Relative absolute error                 63.0988 %
Root relative squared error             79.3765 %
Total Number of Instances              275

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.738     0.119      0.88      0.738     0.803      0.917    Frio
                 0.762     0.276      0.616     0.762     0.681      0.78     Normal
                 0.64      0.036      0.64      0.64      0.64       0.951    Quente
Weighted Avg.    0.738     0.169      0.761     0.738     0.743      0.869

=== Confusion Matrix ===

   a   b   c   <-- classified as
 110  39   0 |   a = Frio
  15  77   9 |   b = Normal
   0   9  16 |   c = Quente
```

Árvore J48 utilizando **Novos objetos.**

```
=== Evaluation on test set ===
=== Summary ===

Correctly Classified Instances         208               75.6364 %
Incorrectly Classified Instances        67               24.3636 %
Kappa statistic                          0.5719
Mean absolute error                      0.1838
Root mean squared error                  0.3623
Relative absolute error                 43.7193 %
Root relative squared error             77.8084 %
Total Number of Instances              275

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.98      0.214      0.844     0.98      0.907      0.961    Frio
                 0.455     0.057      0.821     0.455     0.586      0.844    Normal
                 0.64      0.12       0.348     0.64      0.451      0.888    Quente
Weighted Avg.    0.756     0.148      0.791     0.756     0.748      0.911

=== Confusion Matrix ===

   a   b   c   <-- classified as
 146   3   0 |   a = Frio
  25  46  30 |   b = Normal
   2   7  16 |   c = Quente
```

Bayesiano NaiveBayesSimple utilizando **Novos objetos**.

2. Base de dados estação INMET Patrocínio

Base de dados gerada pela estação climática PATROCINIO-A523 localizada na cidade de Patrocínio e fornecida pelo Instituto Nacional de Meteorologia - INMET. A coleta dos dados é feita minuto a minuto e integralizada na base de dados a cada hora. É disponível para download somente os dados dos últimos 90 dias, desta forma, o intervalo dos dados coletados foi de 17/03/2014 até 15/06/2014, totalizando 2181 objetos. A base é composta por 17 atributos + 1 classe que foi criada.

Entre os 17 atributos, foram escolhidos alguns em específicos. São eles: Umidade, Pressão, Direção do vendo, Velocidade do vento e Radiação. A escolha ocorreu através de estudos bibliográficos e gráficos referentes aos atributos escolhidos, que mostram ser relevantes ao determinar o atributo escolhido como classe que foi a Temperatura.

O atributo Temperatura era representado através de dados contínuos, desta forma foi necessário transformá-lo em um atributo nominal, através da utilização de um algoritmo para conversão representado da seguinte forma:

```
If (temperatura < 18)
   classe = "Frio";

if (temperatura >= 18 && temperatura < 25)
   classe = "Normal";

if (temperatura >= 25)
   classe = "Quente";
```

Vale ressaltar, que todos os objetos foram normalizados para o intervalo de [0 a 1].

Para classificação foi utilizado o *software* WEKA, o algoritmo J48, que é a implementação da árvore de decisão e o algoritmo NaiveBayesSimple que é a implementação do Classificador Bayesiano Simples.

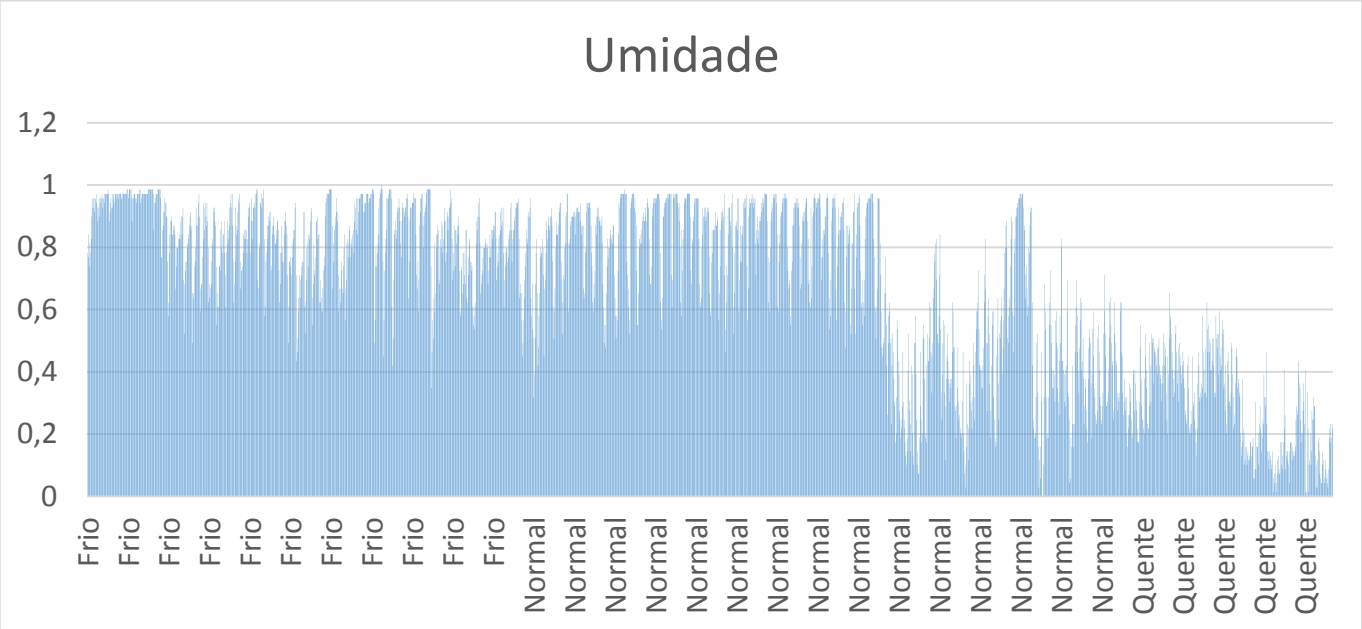
Após o treinamento dos algoritmos com a base de dados: Patrocínio, foram realizados testes de classificação com uma base de dados da cidade de Araxá, sendo que

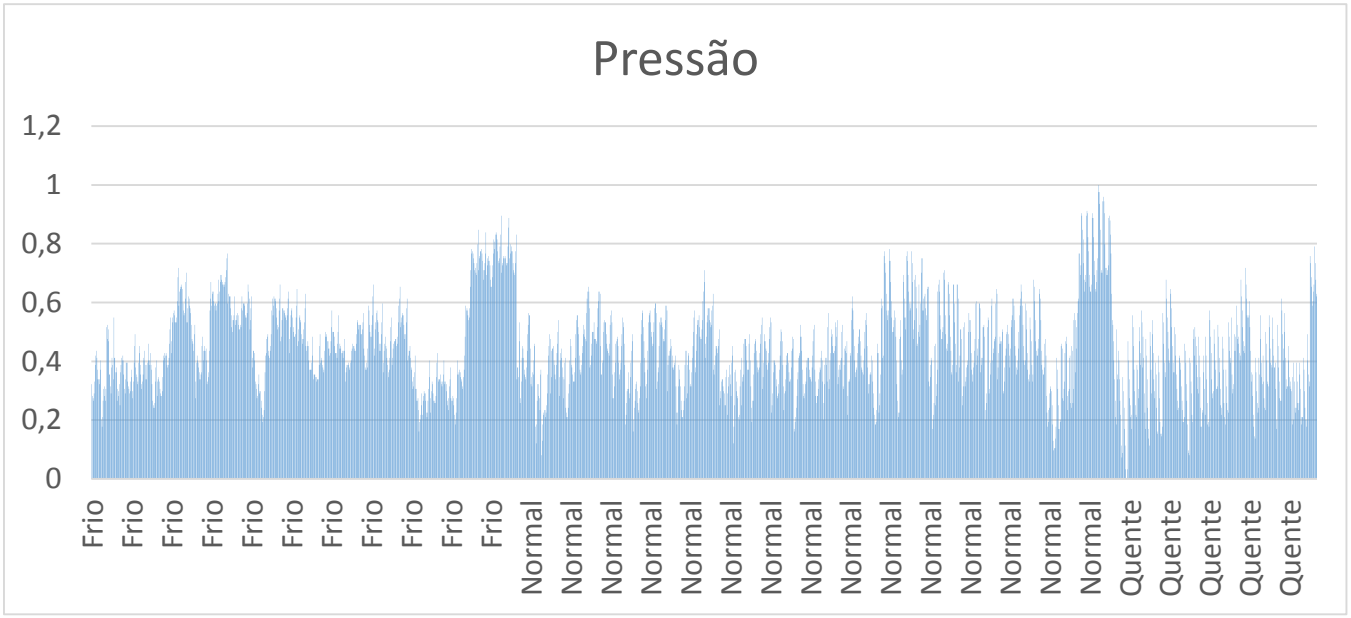
a mesma, possui o mesmo número de objetos e foi pré-processada com as mesmas especificações da base de treinamento.

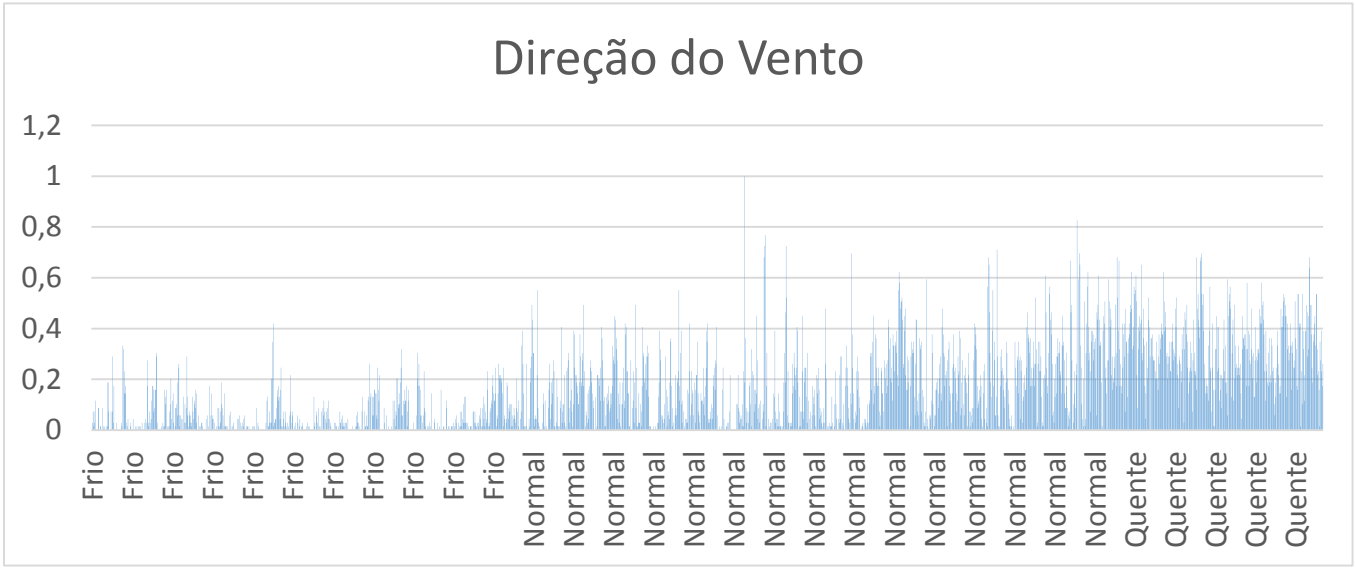
Os resultados podem ser avaliados na seção abaixo.

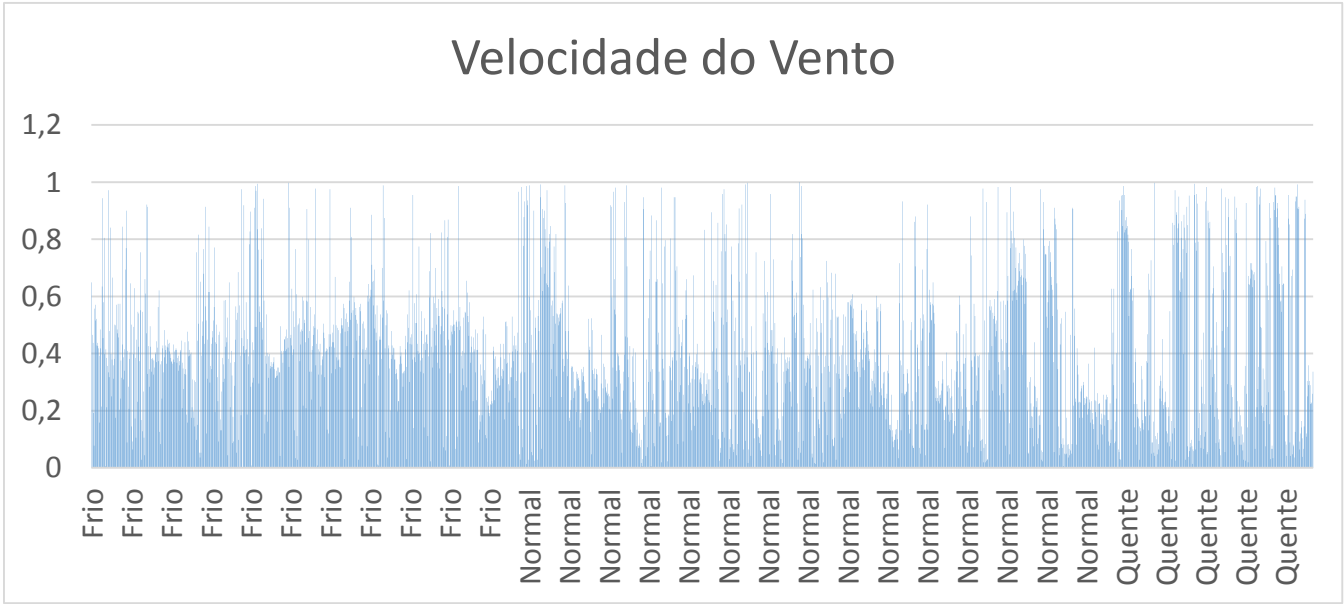
2.1 Resultados

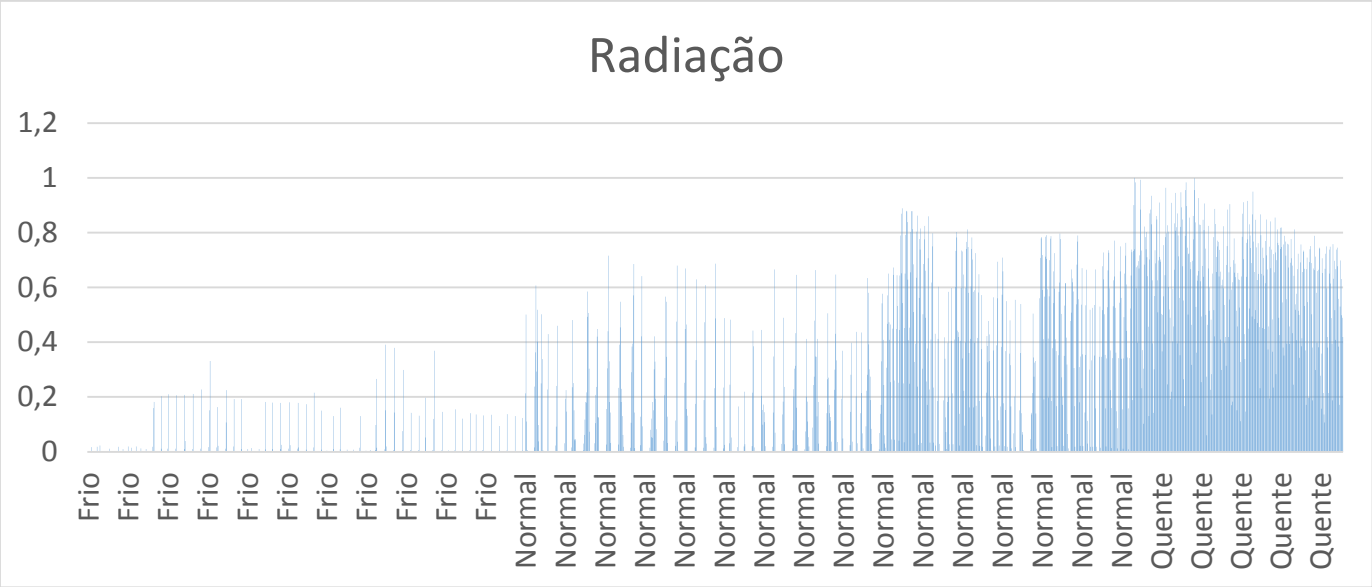
Gráficos gerados através dos objetos da base, para auxílio da interpretação dos dados




Pressão
1,2
1
0,8
0,6
0,4
0,2
0
Frio
Normal
Quente


Direção do Vento
1,2
1
0,8
0,6
0,4
0,2
0
Frio
Normal
Quente


Velocidade do Vento
1,2
1
0,8
0,6
0,4
0,2
0
Frio
Normal
Quente



Análises obtidas a partir do WEKA

Bayesiano NaiveBayesSimple utilizando **Cross-validation = 10.**

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        1413               64.8165 %
Incorrectly Classified Instances       767               35.1835 %
Kappa statistic                          0.4642
Mean absolute error                      0.2436
Root mean squared error                  0.4418
Relative absolute error                 59.3391 %
Root relative squared error             97.5158 %
Total Number of Instances             2180

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.91      0.333      0.592     0.91      0.717      0.845    Frio
                 0.398     0.115      0.765     0.398     0.523      0.711    Normal
                 0.833     0.09       0.65      0.833     0.73       0.954    Quente
Weighted Avg.    0.648     0.186      0.686     0.648     0.625      0.798

=== Confusion Matrix ===

   a   b   c   <-- classified as
 688  68   0 |   a = Frio
 474 421 164 |   b = Normal
   0  61 304 |   c = Quente
```

Árvore J48 utilizando **Cross-validation = 10.**

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        1577               72.3394 %
Incorrectly Classified Instances       603               27.6606 %
Kappa statistic                          0.5553
Mean absolute error                      0.2278
Root mean squared error                  0.3626
Relative absolute error                 55.4792 %
Root relative squared error             80.0316 %
Total Number of Instances             2180

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.728     0.145      0.728     0.728     0.728      0.884    Frio
                 0.695     0.25       0.724     0.695     0.709      0.777    Normal
                 0.797     0.064      0.713     0.797     0.753      0.938    Quente
Weighted Avg.    0.723     0.182      0.724     0.723     0.723      0.841

=== Confusion Matrix ===

   a   b   c   <-- classified as
 550 206   0 |   a = Frio
 206 736 117 |   b = Normal
   0  74 291 |   c = Quente
```

Bayesiano NaiveBayesSimple utilizando **Novos objetos**.

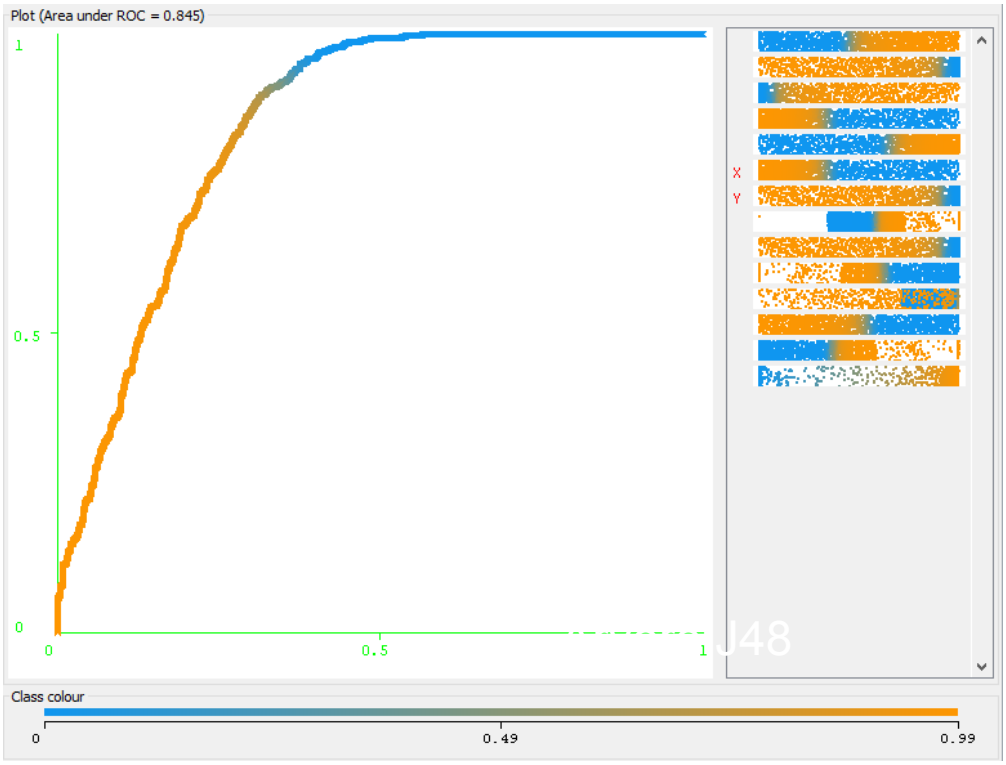
```
=== Evaluation on test set ===
=== Summary ===

Correctly Classified Instances        1134               57.7687 %
Incorrectly Classified Instances       829               42.2313 %
Kappa statistic                          0.2536
Mean absolute error                      0.2887
Root mean squared error                  0.4906
Relative absolute error                 72.9358 %
Root relative squared error            112.3895 %
Total Number of Instances             1963

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.262     0.151      0.318     0.262     0.287      0.655    Frio
                 0.615     0.478      0.688     0.615     0.649      0.576    Normal
                 0.854     0.15       0.514     0.854     0.641      0.916    Quente
Weighted Avg.    0.578     0.357      0.582     0.578     0.571      0.646

=== Confusion Matrix ===

   a   b   c   <-- classified as
 109 301   6 |   a = Frio
 234 762 243 |   b = Normal
   0  45 263 |   c = Quente
```

Árvore J48 utilizando **Novos objetos.**

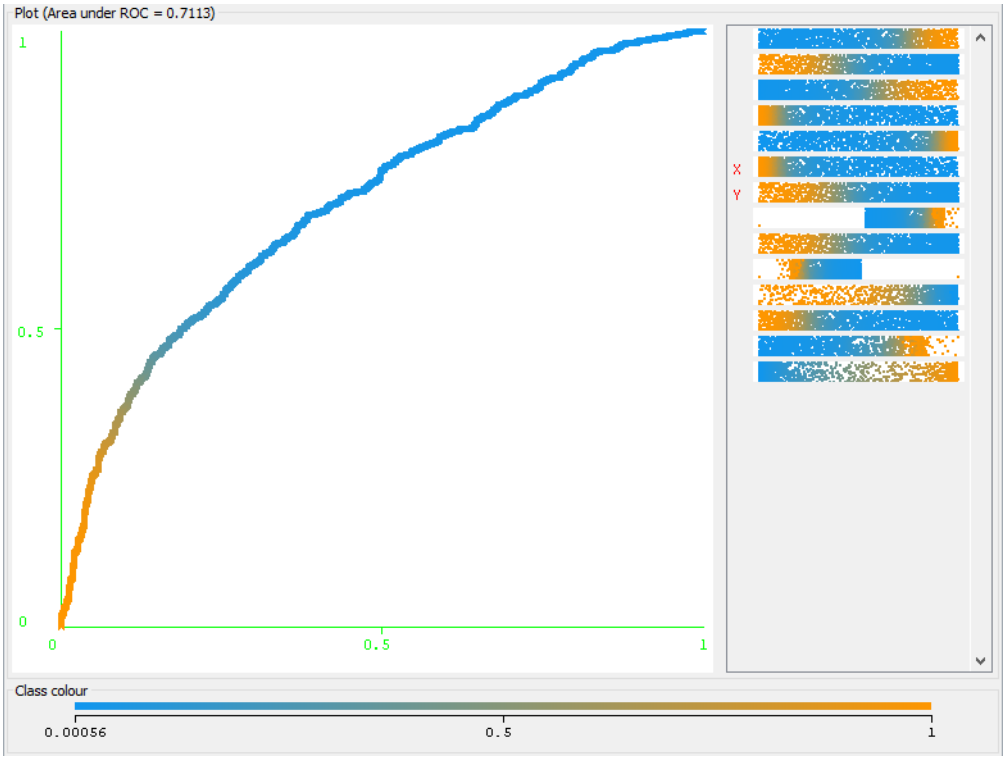
```
=== Evaluation on test set ===
=== Summary ===

Correctly Classified Instances        1160               59.0932 %
Incorrectly Classified Instances       803               40.9068 %
Kappa statistic                          0.2683
Mean absolute error                      0.2823
Root mean squared error                  0.4308
Relative absolute error                 71.3154 %
Root relative squared error             98.7042 %
Total Number of Instances             1963

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.139     0.078      0.326     0.139     0.195      0.704    Frio
                 0.646     0.489      0.694     0.646     0.669      0.612    Normal
                 0.977     0.199      0.478     0.977     0.642      0.902    Quente
Weighted Avg.    0.591     0.356      0.582     0.591     0.564      0.677

=== Confusion Matrix ===

   a   b   c   <-- classified as
  58 347  11 |   a = Frio
 120 801 318 |   b = Normal
   0   7 301 |   c = Quente
```

Curvas ROC - Bayesiano NaiveBayesSimple

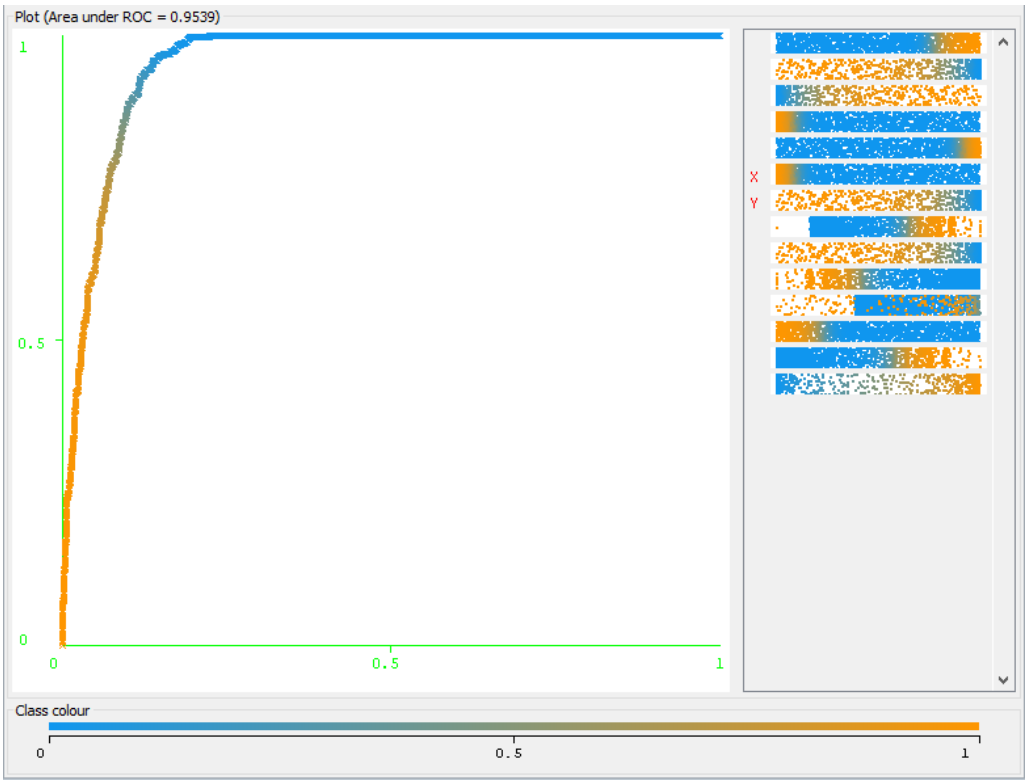
Classe: Frio



Classe: Normal

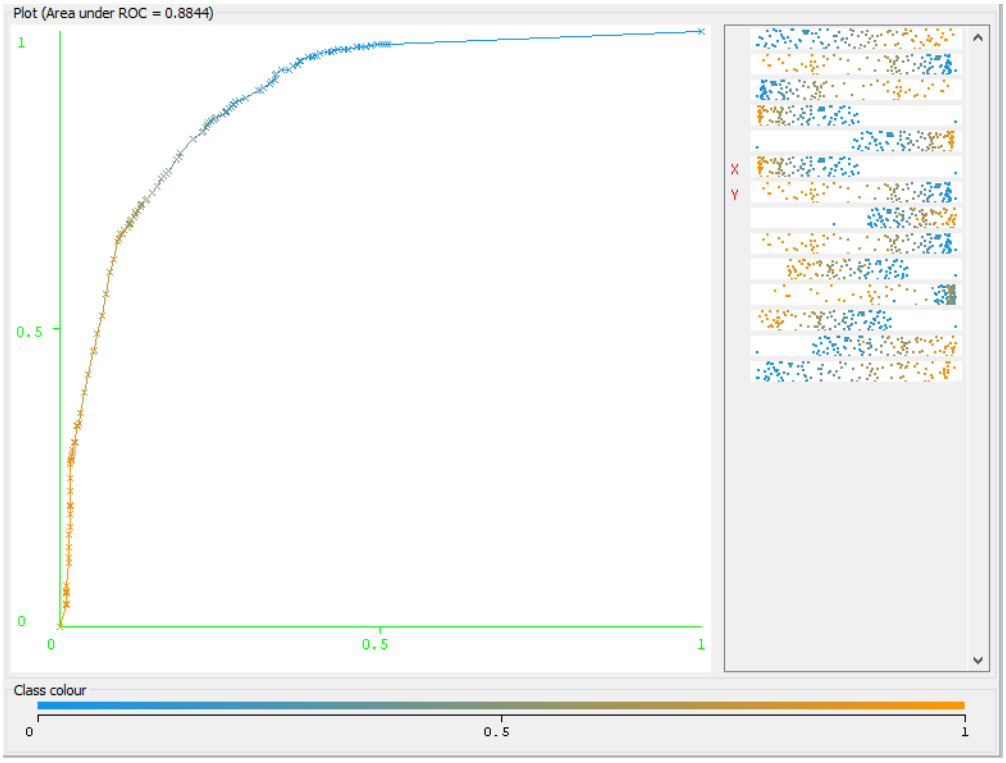


Classe: Quente

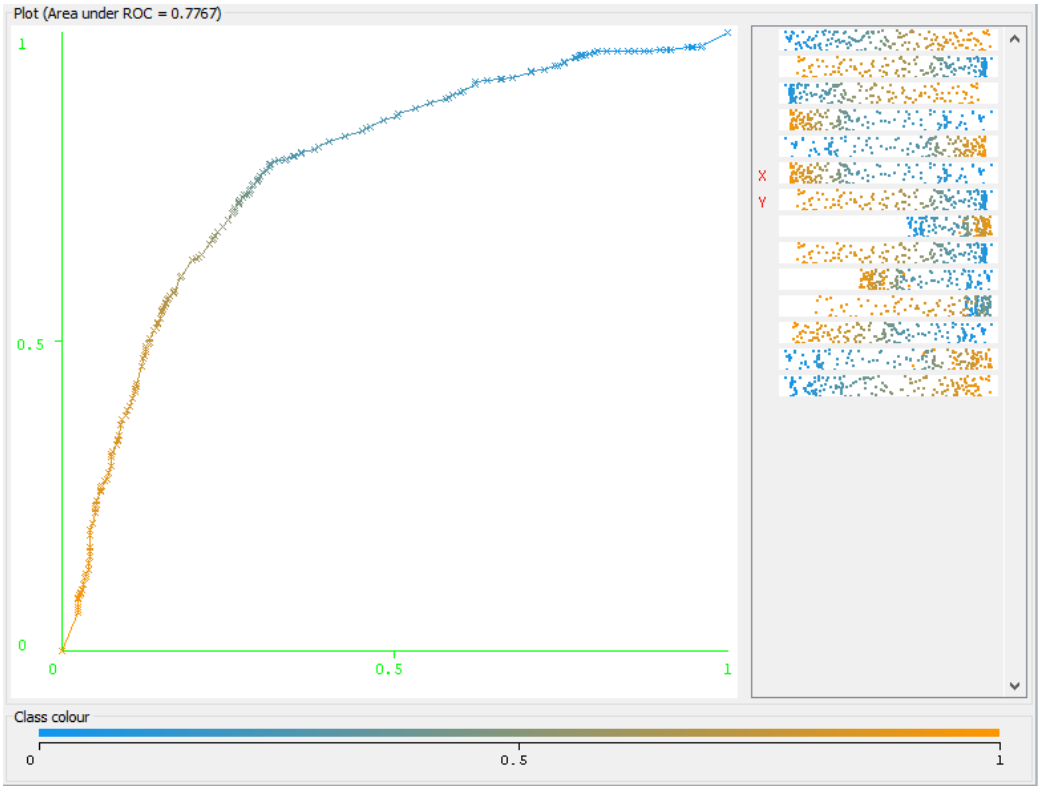


Curvas ROC - Árvore J48

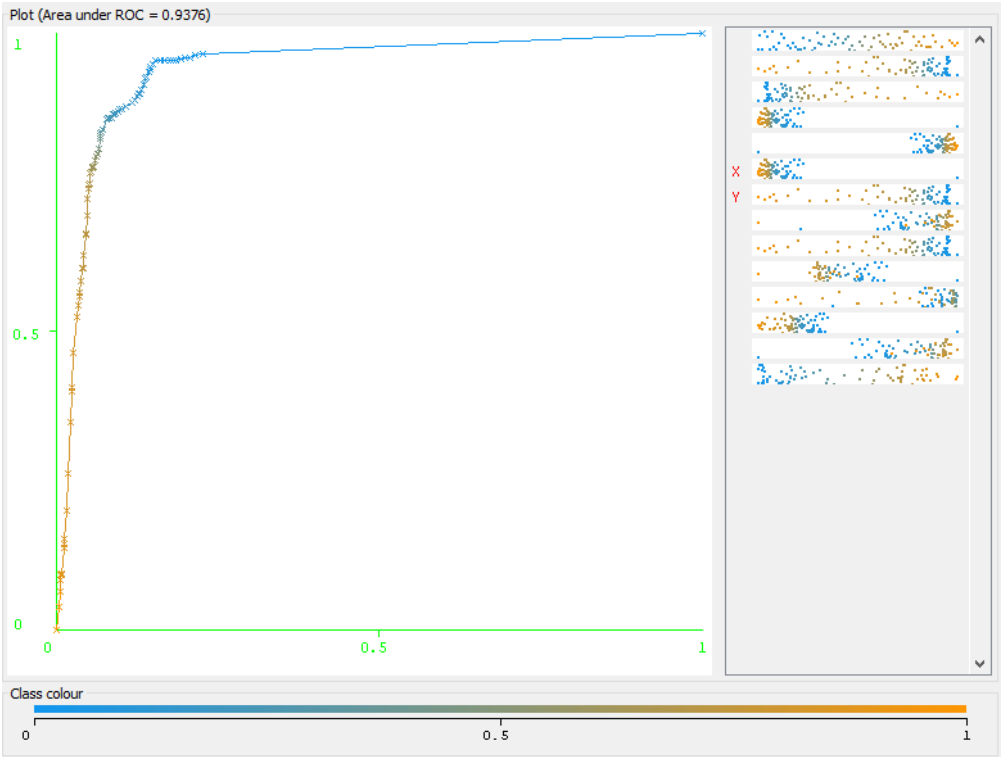
Classe: Frio



Classe: Normal



Classe: Quente



É possível observar com a base de dados testada, que a partir da classe Quente, o classificador bayesiano obteve um melhor desempenho em relação a árvore de decisão, porém as outras duas classes: Frio e Normal, indica uma melhor performance ao executar sobre a árvore de decisão.

3. Conjunto de dados de lentes

Base de dados que tem como finalidade verificar se o cliente pode ou não usar lentes de contato, se sim qual lente de contato especifica o cliente pode usar, exemplo lentes rígidas ou gelatinosas. Cada objeto é completa e correta. Sendo que a base é livre de ruídos. E se encontra para download em http://archive.ics.uci.edu/ml/datasets/Lenses. A base de dados tem exatamente 24 objetos, foi escolhida pelo número inferior, assim testamos o classificador bayesiano se é bom também para classificar poucos objetos.

A base possui 4 atributos mais 1 classe no qual o objeto será classificado. O primeiro atributo chamado de idade das vistas tem como classificação três opções: novo, Pre-presbiopia e Presbiopia. Onde novo é uma visão nova, Presbiopia é uma visão velha e a Pre-presbiopia é o meio termo. O Segundo atributo, a prescrição são problemas que o cliente já possui, como por exemplo: Mópe e Hipermetrope. O terceiro atributo, é o astigmatismo, se a pessoa tem ou não essa falha na visão. E por último o atributo lagrima, se o paciente tem a lagrima reduzida ou é normal.

3.1 Pre-Processamento

O único pre-processamento feito a trasformação dos dados da base que estavam com números discretos para atributos nomais, seguindo a documentação da base

de dados. Como os atributos nominais estavam completos, ou seja, objetos faltantes e não há a necessidade de normalizar os dados. Foi jogado a base no Weka e abaixo segue os resultados obtidos.

3.2 Resultados

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          20               83.3333 %
Incorrectly Classified Instances         4               16.6667 %
Kappa statistic                          0.71
Mean absolute error                      0.15
Root mean squared error                  0.3249
Relative absolute error                 39.7059 %
Root relative squared error             74.3898 %
Total Number of Instances               24

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.8       0.111      0.923     0.8       0.857      0.811    nao
                 1         0.053      0.833     1         0.909      0.947    gelatinosa
                 0.75      0.1        0.6       0.75      0.667      0.813    rigido
Weighted Avg.    0.833     0.097      0.851     0.833     0.836      0.84

=== Confusion Matrix ===

  a  b  c   <-- classified as
 12  1  2 |  a = nao
  0  5  0 |  b = gelatinosa
  1  0  3 |  c = rigido
```

Árvore J48 utilizando **Cross-validation = 10.**

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          17               70.8333 %
Incorrectly Classified Instances         7               29.1667 %
Kappa statistic                          0.4381
Mean absolute error                      0.2545
Root mean squared error                  0.3326
Relative absolute error                 67.3578 %
Root relative squared error             76.1544 %
Total Number of Instances               24

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.8       0.444      0.75      0.8       0.774      0.83     nao
                 0.8       0.053      0.8       0.8       0.8        0.947    gelatinosa
                 0.25      0.1        0.333     0.25      0.286      0.925    rigido
Weighted Avg.    0.708     0.305      0.691     0.708     0.698      0.87

=== Confusion Matrix ===

  a  b  c   <-- classified as
 12  1  2 |  a = nao
  1  4  0 |  b = gelatinosa
  3  0  1 |  c = rigido
```

Bayesiano NaiveBayesSimple utilizando **Cross-validation = 10.**

Utilizando todos os atributos o bayesiano obteve uma taxa de acerto de 71% abaixo da árvore que obteve 83%.

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          17               70.8333 %
Incorrectly Classified Instances         7               29.1667 %
Kappa statistic                          0.5
Mean absolute error                      0.2348
Root mean squared error                  0.3571
Relative absolute error                 62.1658 %
Root relative squared error             81.7569 %
Total Number of Instances               24

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.8       0          1         0.8       0.889      0.8      nao
                 1         0.368      0.417     1         0.588      0.716    gelatinosa
                 0         0          0         0         0          0.7      rigido
Weighted Avg.    0.708     0.077      0.712     0.708     0.678      0.766

=== Confusion Matrix ===

  a  b  c   <-- classified as
 12  3  0 |  a = nao
  0  5  0 |  b = gelatinosa
  0  4  0 |  c = rigido
```

Árvore J48 utilizando **Novos objetos.**

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          17               70.8333 %
Incorrectly Classified Instances         7               29.1667 %
Kappa statistic                          0.4381
Mean absolute error                      0.2545
Root mean squared error                  0.3326
Relative absolute error                 67.3578 %
Root relative squared error             76.1544 %
Total Number of Instances               24

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.8       0.444      0.75      0.8       0.774      0.83     nao
                 0.8       0.053      0.8       0.8       0.8        0.947    gelatinosa
                 0.25      0.1        0.333     0.25      0.286      0.925    rigido
Weighted Avg.    0.708     0.305      0.691     0.708     0.698      0.87

=== Confusion Matrix ===

  a  b  c   <-- classified as
 12  1  2 |  a = nao
  1  4  0 |  b = gelatinosa
  3  0  1 |  c = rigido
```

Bayesiano NaiveBayesSimple utilizando **Novos objetos**.

Quando pedimos para o WEKA selecionar os atributos e classificar os dois classificados chegaram a uma taxa de acerto de 71%, só que a árvore de decisão passou a não conseguir classificar mais as lentes rígidas, assim perdendo muito o seu desempenho.