
UNIVERSIDADE FEDERAL DE UBERLÂNDIA
FACULDADE DE COMPUTAÇÃO - FACOM

Elton Henrique Lunardi Gimenes
Nick Ishida
Carlos Daniel
Pablo Soares

Trabalho Prático de Mineração de Dados: Classificação

Relatório do Trabalho Prático

Elton Henrique Lunardi Gimenes

Trabalho Prático de Mineração de Dados: Classificação

Relatório apresentado à disciplina GSI556 - Mineração de Dados, como parte dos requisitos necessários à obtenção da nota correspondente à disciplina.

Professor(a): Carlos Cesar Mansur Tuma

Disciplina: Mineração de Dados

Turma: Turma S - 2021/2

Monte Carmelo

2022

Resumo

Relacionamento dos metodos ID3, Naive-Bayes e KNN de classificação para mineração de dados devidamente configurados para as bases “breast-cancer, glass,hypothyroid e ionosphere” com validação do 10-cross.

Lista de ilustrações

Figura 1 – Print da base visualizada no Weka	8
Figura 2 – Atributos do Câncer de Mama	8
Figura 3 – Localização do J48	9
Figura 4 – Configuração do J48(ID3)	10
Figura 5 – Resultado da Classificação J48	11
Figura 6 – Resultado da Classificação J48	12
Figura 7 – Arvore J48.....	12
Figura 8 – Diretório Naive-Bayes	13
Figura 9 – Configuração Naive-Bayes.....	13
Figura 10 – Resultado do Método Naive-Bayes	14
Figura 11 – Resultado do Método Naive-Bayes	15
Figura 12 – Diretório do Método IBK(KNN)	15
Figura 13 – Configuração IBK(KNN).....	16
Figura 14 – Resultados IBK(KNN)	17
Figura 15 – Resultados IBK(KNN)	17
Figura 16 – Resultado ID3.....	17
Figura 17 – Resultado Naive-Bayes	18
Figura 18 – Resultado KNN.....	18
Figura 19 – Atributos restantes.....	18
Figura 20 – Base glass inicial	19
Figura 21 – Localização do J48.....	20
Figura 22 – Configuração do J48(KNN)	21
Figura 23 – Resultados do J48.....	22
Figura 24 – Resultados do J48.....	23
Figura 25 – Resultados do J48.....	24
Figura 26 – Arvore J48.....	24
Figura 27 – Resultado Naive-Bayes	25
Figura 28 – Resultados KNN.....	26
Figura 29 – Resultados do ID3	27
Figura 30 – Resultados do ID3	28
Figura 31 – Resultados do ID3	29
Figura 32 – Resultado Naive-Bayes	29
Figura 33 – Resultado KNN.....	30
Figura 34 – Base Hypothyroid inicial	31
Figura 35 – Resultado ID3.....	32
Figura 36 – Resultado ID3.....	33
Figura 37 – Arvore ID3	33

Figura 38 – Resultado Naive-Bayes	34
Figura 39 – Resultado KNN.....	35
Figura 40 – Resultado KNN (K=3).....	36
Figura 41 – Resultado ID3.....	37
Figura 42 – Resultado ID3.....	38
Figura 43 – Resultado Naive-Bayes	38
Figura 44 – Resultado KNN.....	39
Figura 45 – Base Limpa	40

Sumário

1	Método de Classificação	6
1.1	ID3	6
1.2	Naive-Bayes	6
1.3	KNN	6
2	Método de Avaliação	7
2.1	Validação 10-cross	7
3	Metodologia e Aplicação	8
3.1	Cancer de Mama	8
3.1.1	Método ID3	9
3.1.2	Método Naive-Bayes.....	12
3.1.3	Método KNN	15
3.1.4	Comparação dos Resultados	17
3.2	Glass.....	18
3.2.1	Método ID3.....	19
3.2.2	Método Naive-Bayes.....	24
3.2.3	Método KNN	25
3.2.4	Comparação dos Resultados	26
3.3	Hypothyroid	30
3.3.1	Método ID3.....	31
3.3.2	Método Naive-Bayes.....	34
3.3.3	Método KNN.....	34
3.3.4	Comparação dos Resultados	37
3.4	Ionosphaera	30
3.4.1	Método ID3.....	31
3.4.2	Método Naive-Bayes.....	34
3.4.3	Método KNN.....	34
3.4.4	Comparação dos Resultados	37

1 Método de Classificação

1.1 ID3

O método de mineração de dados id3 é uma abordagem poderosa para identificar as relações entre atributos e entidades visuais. Ele pode: (1) filtrar dados irrelevantes ou redundantes, executando-os através de uma função de filtro que descarta alguns dos atributos a fim de focar em atributos particulares no conjunto de treinamento; (2) prever resultados esperados para novos casos que não estejam no conjunto de treinamento, pois esta é uma técnica de mineração de dados utilizada para alcançar o aprendizado supervisionado.

1.2 Naive-Bayes

O método Naive Bayes é um exemplo de métodos estatísticos não paramétricos. Ele envolve a modelagem de uma variável contínua com várias variáveis discretas. O método Naive Bayes não é um conceito novo; foi introduzido pela primeira vez em 1959 por Thomas Bayes (1702-1761) para prever a probabilidade de que um volume de vários eventos ocorrerá dado o conhecimento de suas probabilidades individuais. Onde há mais de dois resultados discretos de interesse, vários modelos de classificação binária separados poderiam ser usados como uma alternativa à previsão de Naive Bayes.

1.3 KNN

O KNN é um método para resolver problemas de clusterização. Ele assume que a preferência dos usuários pode ser representada por um conjunto de pares de chaves de valor para cada instância, onde a chave representa o item que os usuários selecionam frequentemente e o valor representa os outros itens que os usuários gostam menos por exemplo

2 Método de Avaliação

2.1 Validação 10-cross

A validação 10-cross é um método para avaliar a qualidade de um modelo de previsão. Os dados de treinamento são divididos em duas amostras, cada uma das quais é usada para treinar um conjunto de testes separado, e depois são feitas previsões sobre o conjunto de dados combinados. O desempenho do algoritmo é comparado com uma resposta correta conhecida (tal como um conjunto de dados da verdade do terreno) para avaliar a capacidade de generalização.

3 Metodologia e Aplicação

3.1 Cancer de Mama

Nosso trabalho começa como a base de dados “breast-cancer.arff” que relaciona 286 instancias e 10 atributos oriundos e característicos do cancer de mama, atributos e base essa listada a seguir:

Figura 1 – Print da base visualizada no Weka

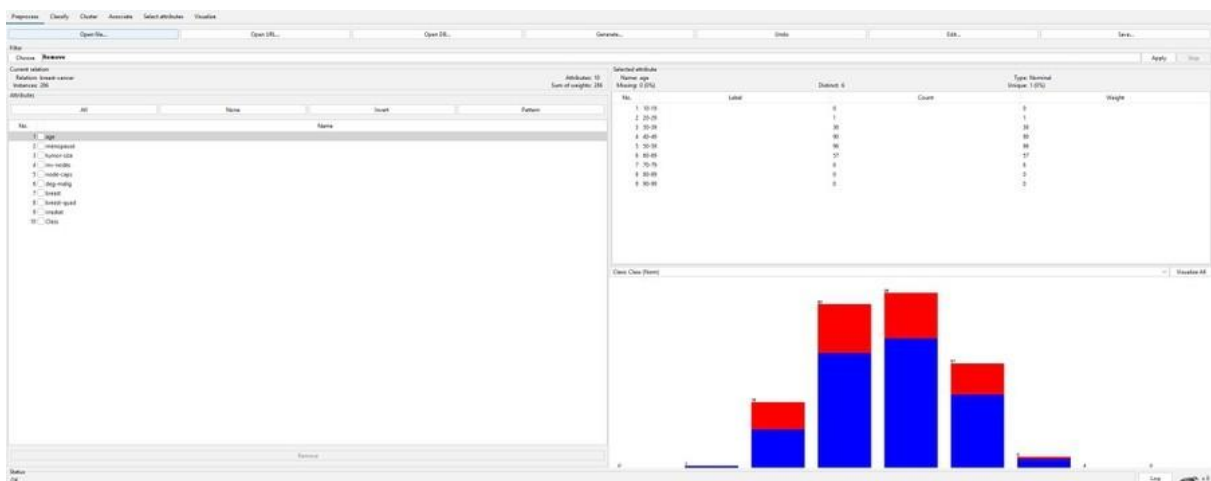


Figura 2 – Atributos do Câncer de Mama

Current relation:
Relation: breast-cancer
Instances: 286

Attributes:

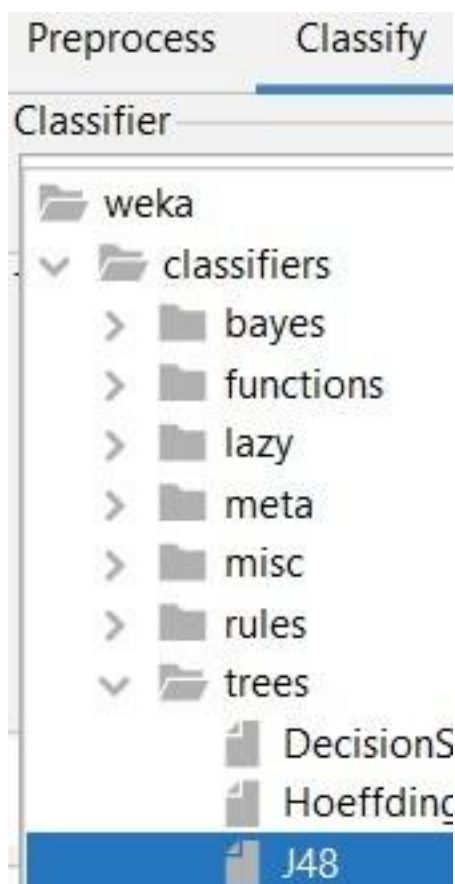
All

No.	Attribute
1	age
2	menopause
3	tumor-size
4	inv-nodes
5	node-caps
6	deg-malign
7	breast
8	breast-quadrant
9	inradiat
10	Class

3.1.1 Método ID3

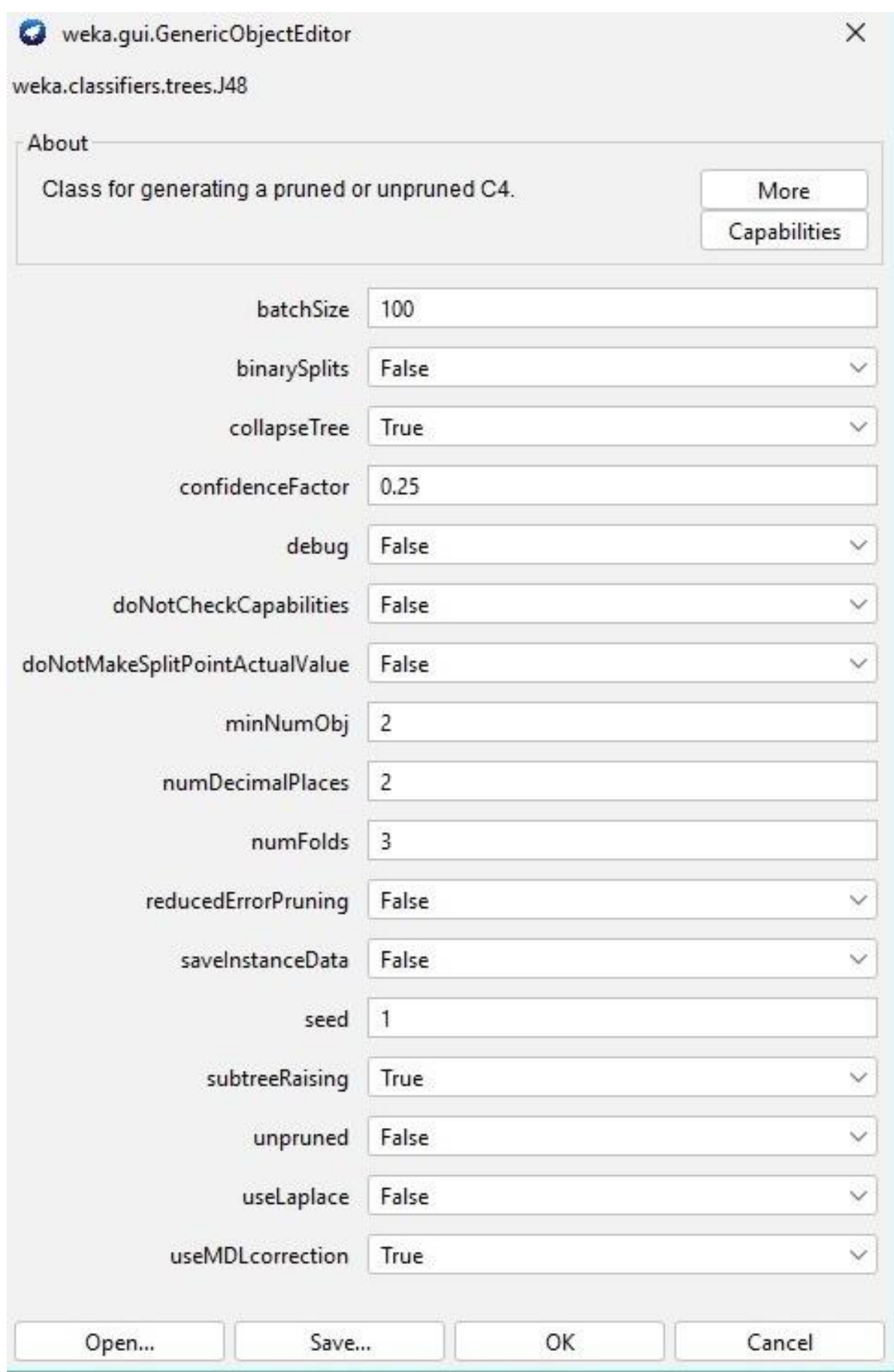
A aplicação do método ID3 será realizado dentro da aplicação Weka em sua versão estável 3.8, dentro do sistema operacional Windows 10; a inicialização da classificação começa na aba Classify do Weka, escolhemos a opção J48, que corresponde ao método ID3, dentro da pasta “trees” pela opção “Choose” do programa.

Figura 3 – Localização do J48



Com o método selecionado, cabe-nos agora a configuração de acordo com a nossa necessidade.

Figura 4 – Configuração do J48(ID3)



weka.gui.GenericObjectEditor

weka.classifiers.trees.J48

About

Class for generating a pruned or unpruned C4.

More

Capabilities

batchSize 100

binarySplits False

collapseTree True

confidenceFactor 0.25

debug False

doNotCheckCapabilities False

doNotMakeSplitPointActualValue False

minNumObj 2

numDecimalPlaces 2

numFolds 3

reducedErrorPruning False

saveInstanceData False

seed 1

subtreeRaising True

unpruned False

useLaplace False

useMDLcorrection True

Open... Save... OK Cancel

Realizado, as devidas configurações e seleção do métodos, inicia-se a classificação, e com ela obtemos os seguintes resultados provenientes do Weka.

Figura 5 – Resultado da Classificação J48

```

Class
Test mode: 10-fold cross-validation

=== Classifier model (full training set) ===

J48 pruned tree
-----

node-caps = yes
| deg-malig = 1: recurrence-events (1.01/0.4)
| deg-malig = 2: no-recurrence-events (26.2/8.0)
| deg-malig = 3: recurrence-events (30.4/7.4)
node-caps = no: no-recurrence-events (228.39/53.4)

Number of Leaves : 4

Size of the tree : 6

Time taken to build model: 0.01 seconds

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

Correctly Classified Instances      216          75.5245 %
Incorrectly Classified Instances    70           24.4755 %
Kappa statistic                    0.2826
Mean absolute error                 0.3676
Root mean squared error             0.4324
Relative absolute error             87.8635 %
Root relative squared error         94.6093 %
Total Number of Instances          286

```

Figura 6 – Resultado da Classificação J48

```

=== Detailed Accuracy By Class ===

```

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0,960	0,729	0,757	0,960	0,846	0,339	0,584	0,736	no-recurrence-events
	0,271	0,040	0,742	0,271	0,397	0,339	0,584	0,436	recurrence-events
Weighted Avg.	0,755	0,524	0,752	0,755	0,713	0,339	0,584	0,647	

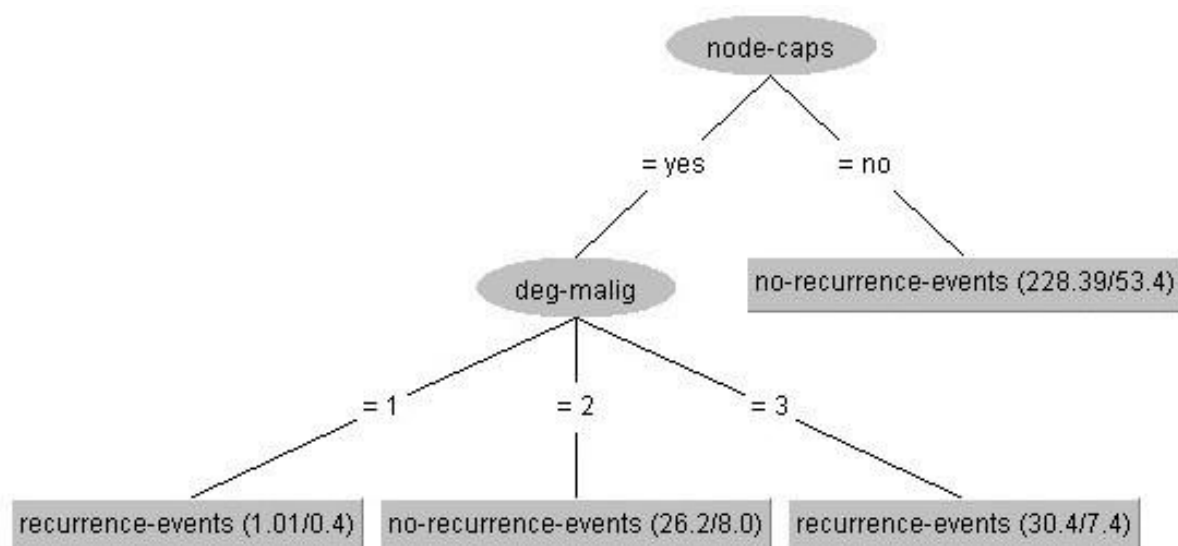
```

=== Confusion Matrix ===
  a  b  <-- classified as
193  8 |  a = no-recurrence-events
 62 23 |  b = recurrence-events

```

Após a geração dos resultados, podemos visualizar com o botão direito, a árvore de decisões que foi formada.

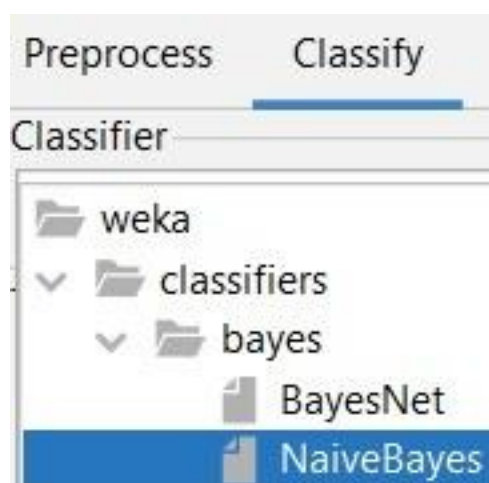
Figura 7 – Árvore J48



3.1.2 Método Naive-Bayes

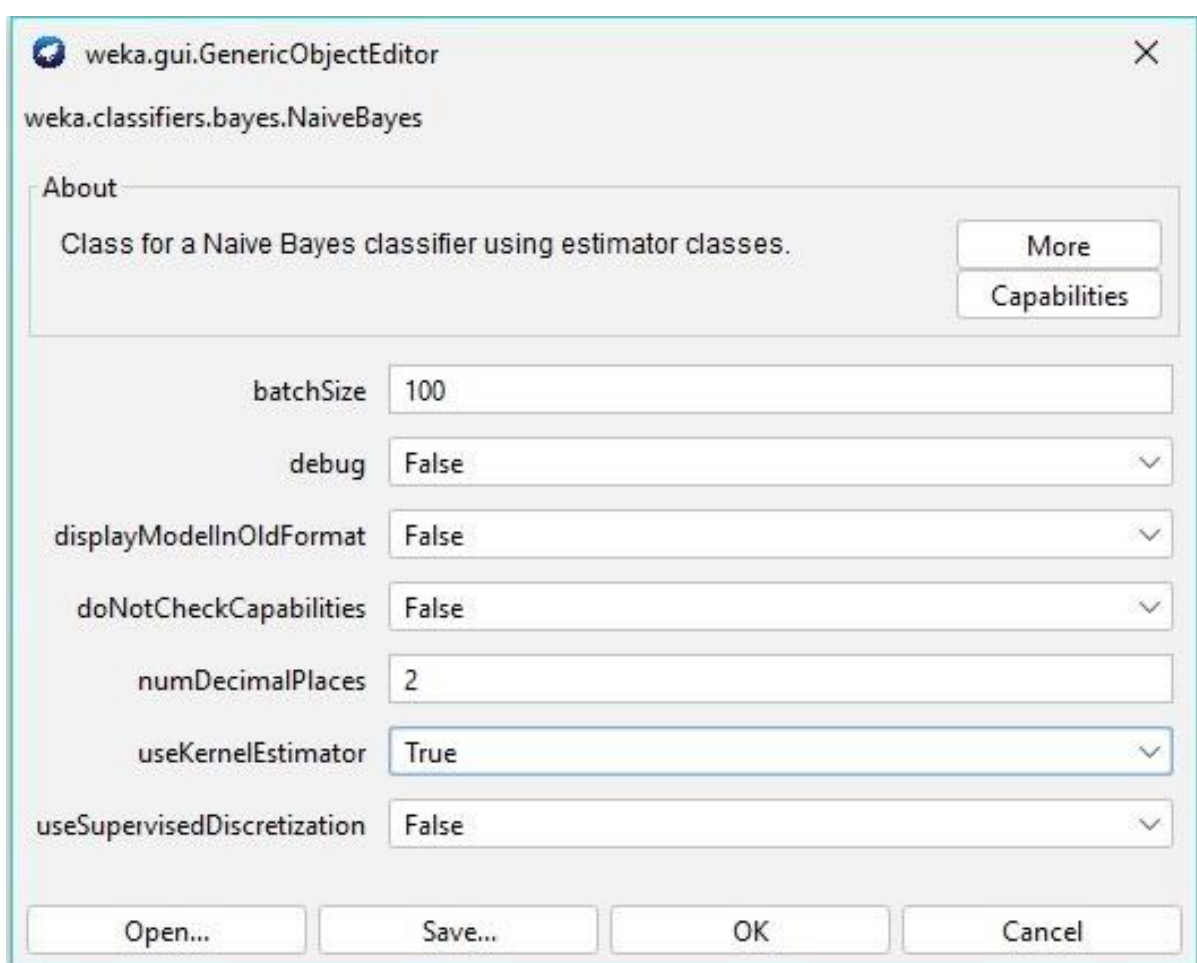
Para o método Naive-Bayes acessamos de maneira similar, contudo, dentro da pasta “bayes” que se encontra dentro dos “classifiers”.

Figura 8 – Diretório Naive-Bayes



Para acessarmos o menu e fazer as configurações, se dá da mesma forma do método anterior, dessa forma, a fazemos.

Figura 9 – Configuração Naive-Bayes



Feito, as devidas configurações acima, podemos obter os seguintes resultados.

Figura 10 – Resultado do Método Naive-Bayes

```

Test mode:      10-fold cross-validation

=== Classifier model (full training set) ===

Naive Bayes Classifier

              Class
Attribute      no-recurrence-events  recurrence-events
              (0.7)                  (0.3)
=====
node-caps
  yes          26.0                  32.0
  no          172.0                  52.0
  [total]     198.0                  84.0

deg-malig
  1           60.0                  13.0
  2          103.0                  29.0
  3           41.0                  46.0
  [total]     204.0                  88.0

Time taken to build model: 0 seconds

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

Correctly Classified Instances      215          75.1748 %
Incorrectly Classified Instances    71           24.8252 %
Kappa statistic                    0.2754
Mean absolute error                 0.352
Root mean squared error             0.4275
Relative absolute error             84.134 %
Root relative squared error         93.5325 %
Total Number of Instances          286

```


Figura 11 – Resultado do Método Naive-Bayes

```

=== Detailed Accuracy By Class ===

```

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0,955	0,729	0,756	0,955	0,844	0,327	0,646	0,763	no-recurrence-events
	0,271	0,045	0,719	0,271	0,393	0,327	0,646	0,479	recurrence-events
Weighted Avg.	0,752	0,526	0,745	0,752	0,710	0,327	0,646	0,678	

```

=== Confusion Matrix ===

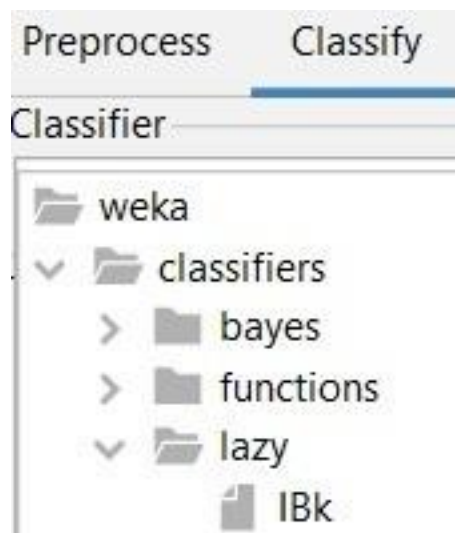
```

a	b	<-- classified as
192	9	a = no-recurrence-events
62	23	b = recurrence-events

3.1.3 Método KNN

Para a realização do KNN, precisamos procurar pelo seu método correspondente, o IBK do Weka, este, encontramos agora na pasta “lazy” dos classificadores.

Figura 12 – Diretório do Método IBK(KNN)



Após a seleção do método correspondente, seguimos para as configurações com KNN = 3.

Figura 13 – Configuração IBK(KNN)

The screenshot shows the 'weka.gui.GenericObjectEditor' window for the 'weka.classifiers.lazy.IBk' classifier. The 'About' section describes it as a 'K-nearest neighbours classifier.' with buttons for 'More' and 'Capabilities'. The configuration parameters are as follows:

Parameter	Value
KNN	3
batchSize	100
crossValidate	True
debug	False
distanceWeighting	No distance weighting
doNotCheckCapabilities	False
meanSquared	False
nearestNeighbourSearchAlgorithm	Choose LinearNNSearch -A "weka.core.Euclidean"
numDecimalPlaces	2
windowSize	0

At the bottom, there are buttons for 'Open...', 'Save...', 'OK', and 'Cancel'.

Realizadas as configurações, obtivemos os seguintes resultados.

Figura 14 – Resultados IBK(KNN)

```

Test mode:      10-fold cross-validation

=== Classifier model (full training set) ===

IB1 instance-based classifier
using 2 nearest neighbour(s) for classification

Time taken to build model: 0 seconds

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

Correctly Classified Instances      199           69.5804 %
Incorrectly Classified Instances    87           30.4196 %
Kappa statistic                     0.1405
Mean absolute error                 0.3516
Root mean squared error            0.4997
Relative absolute error             84.0345 %
Root relative squared error        109.3175 %
Total Number of Instances          286

```

Figura 15 – Resultados IBK(KNN)

```

=== Detailed Accuracy By Class ===

      TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area  PRC Area  Class
      0,896   0,776   0,732     0,896   0,805     0,157   0,617   0,777   no-recurrence-events
      0,224   0,104   0,475     0,224   0,304     0,157   0,617   0,425   recurrence-events
Weighted Avg.   0,696   0,577   0,655     0,696   0,656     0,157   0,617   0,672

=== Confusion Matrix ===

  a  b  <-- classified as
180 21 |  a = no-recurrence-events
 66 19 |  b = recurrence-events

```

3.1.4 Comparação dos Resultados

Listamos agora os 3 resultados obtidos para análise e discussão.

Figura 16 – Resultado ID3

```

=== Detailed Accuracy By Class ===

      TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area  PRC Area  Class
      0,960   0,729   0,757     0,960   0,846     0,339   0,584   0,736   no-recurrence-events
      0,271   0,040   0,742     0,271   0,397     0,339   0,584   0,436   recurrence-events
Weighted Avg.   0,755   0,524   0,752     0,755   0,713     0,339   0,584   0,647

=== Confusion Matrix ===

  a  b  <-- classified as
193  8 |  a = no-recurrence-events
 62 23 |  b = recurrence-events

```

Figura 17 – Resultado Naive-Bayes

```

=== Detailed Accuracy By Class ===

          TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area  PRC Area  Class
          0,955    0,729    0,756     0,955    0,844     0,327    0,646    0,763    no-recurrence-events
          0,271    0,045    0,719     0,271    0,393     0,327    0,646    0,479    recurrence-events
Weighted Avg.    0,752    0,526    0,745     0,752    0,710     0,327    0,646    0,678

=== Confusion Matrix ===

  a  b  <-- classified as
192  9 |  a = no-recurrence-events
 62 23 |  b = recurrence-events

```

Figura 18 – Resultado KNN

```

=== Detailed Accuracy By Class ===

          TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area  PRC Area  Class
          0,896    0,776    0,732     0,896    0,805     0,157    0,617    0,777    no-recurrence-events
          0,224    0,104    0,475     0,224    0,304     0,157    0,617    0,425    recurrence-events
Weighted Avg.    0,696    0,577    0,655     0,696    0,656     0,157    0,617    0,672

=== Confusion Matrix ===

  a  b  <-- classified as
180 21 |  a = no-recurrence-events
 66 19 |  b = recurrence-events

```

Após a classificação dos dados, obtivemos a seguinte base de dados com apenas 3 atributos, após a “limpeza”

Figura 19 – Atributos restantes.



Após a limpeza o Naive-Bayes teve uma melhora de 71% para 75% e o KNN teve uma queda de 73% para 69%.

3.2 Glass

Começamos a nova base denominada “glass.arff” com 214 instancias e 10 atributos de materiais da tabela periódica e um atributo “Type”.

Figura 20 – Base glass inicial

Current relation

Relation: Glass
Instances: 214

Attributes: 10
Sum of weights: 214

Attributes

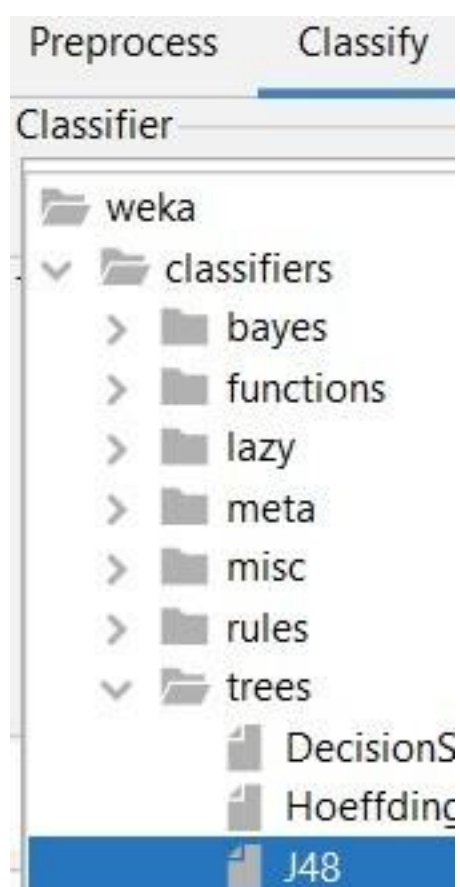
All None Invert Pattern

No.		Name
1	<input checked="" type="checkbox"/>	RI
2	<input type="checkbox"/>	Na
3	<input type="checkbox"/>	Mg
4	<input type="checkbox"/>	Al
5	<input type="checkbox"/>	Si
6	<input type="checkbox"/>	K
7	<input type="checkbox"/>	Ca
8	<input type="checkbox"/>	Ba
9	<input type="checkbox"/>	Fe
10	<input type="checkbox"/>	Type

3.2.1 Método ID3

Da mesma forma da primeira base, fazemos igualmente nesta.

Figura 21 – Localização do J48



E da mesma forma com as mesmas configurações.

Figura 22 – Configuração do J48(KNN)

The screenshot shows the 'weka.gui.GenericObjectEditor' window for the 'weka.classifiers.lazy.IBk' classifier. The 'About' section describes it as a 'K-nearest neighbours classifier.' with buttons for 'More' and 'Capabilities'. The configuration parameters are as follows:

Parameter	Value
KNN	3
batchSize	100
crossValidate	True
debug	False
distanceWeighting	No distance weighting
doNotCheckCapabilities	False
meanSquared	False
nearestNeighbourSearchAlgorithm	Choose LinearNNSearch -A "weka.core.Euclidean"
numDecimalPlaces	2
windowSize	0

At the bottom, there are buttons for 'Open...', 'Save...', 'OK', and 'Cancel'.

Com as devidas configurações realizadas, obtemos os seguintes resultados.

Figura 23 – Resultados do J48

```

Test mode: 10-fold cross-validation

=== Classifier model (full training set) ===

J48 pruned tree
-----

Ba <= 0.27
|  Mg <= 2.41
|  |  K <= 0.03
|  |  |  Na <= 13.75: build wind non-float (3.0)
|  |  |  Na > 13.75: tableware (9.0)
|  |  |  K > 0.03
|  |  |  Na <= 13.49
|  |  |  |  RI <= 1.5241: containers (13.0/1.0)
|  |  |  |  RI > 1.5241: build wind non-float (3.0)
|  |  |  Na > 13.49: build wind non-float (7.0/1.0)
|  Mg > 2.41
|  |  Al <= 1.41
|  |  |  RI <= 1.51707
|  |  |  |  RI <= 1.51596: build wind float (3.0)
|  |  |  |  RI > 1.51596
|  |  |  |  |  Fe <= 0.12
|  |  |  |  |  |  Mg <= 3.54: vehic wind float (5.0)
|  |  |  |  |  |  Mg > 3.54
|  |  |  |  |  |  |  RI <= 1.51667: build wind non-float (2.0)
|  |  |  |  |  |  |  RI > 1.51667: vehic wind float (2.0)
|  |  |  |  |  |  Fe > 0.12: build wind non-float (2.0)
|  |  |  RI > 1.51707
|  |  |  |  K <= 0.23
|  |  |  |  |  Mg <= 3.34: build wind non-float (2.0)
|  |  |  |  |  Mg > 3.34
|  |  |  |  |  |  Si <= 72.64
|  |  |  |  |  |  |  Na <= 14.01: build wind float (14.0)
|  |  |  |  |  |  |  Na > 14.01
|  |  |  |  |  |  |  |  RI <= 1.52211
|  |  |  |  |  |  |  |  |  Na <= 14.32: vehic wind float (3.0)
|  |  |  |  |  |  |  |  |  Na > 14.32: build wind float (2.0)
|  |  |  |  |  |  |  |  RI > 1.52211: build wind float (3.0)
|  |  |  |  |  |  |  Si > 72.64: vehic wind float (3.0)
|  |  |  K > 0.23
|  |  |  |  Mg <= 3.75
|  |  |  |  |  Fe <= 0.14
|  |  |  |  |  |  RI <= 1.52043: build wind float (36.0)
|  |  |  |  |  |  RI > 1.52043: build wind non-float (2.0/1.0)
|  |  |  |  Fe > 0.14

```


Figura 24 – Resultados do J48

```

| | | | | | | RI <= 1.52043: build wind float (36.0)
| | | | | | | RI > 1.52043: build wind non-float (2.0/1.0)
| | | | | | | Fe > 0.14
| | | | | | | Al <= 1.17: build wind non-float (5.0)
| | | | | | | Al > 1.17: build wind float (6.0/1.0)
| | | | | | | Mg > 3.75: build wind non-float (10.0)
| | | | | | | Al > 1.41
| | | | | | | Si <= 72.49
| | | | | | | Ca <= 8.28: build wind non-float (6.0)
| | | | | | | Ca > 8.28: vehic wind float (5.0/1.0)
| | | | | | | Si > 72.49
| | | | | | | RI <= 1.51732
| | | | | | | Fe <= 0.22: build wind non-float (30.0/1.0)
| | | | | | | Fe > 0.22
| | | | | | | RI <= 1.51629: build wind float (2.0)
| | | | | | | RI > 1.51629: build wind non-float (2.0)
| | | | | | | RI > 1.51732
| | | | | | | RI <= 1.51789: build wind float (3.0)
| | | | | | | RI > 1.51789: build wind non-float (2.0)
Ba > 0.27
| | | | | | | Si <= 70.16: build wind non-float (2.0/1.0)
| | | | | | | Si > 70.16: headlamps (27.0/1.0)

Number of Leaves : 30

Size of the tree : 59

Time taken to build model: 0.01 seconds

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

Correctly Classified Instances 143 66.8224 %
Incorrectly Classified Instances 71 33.1776 %
Kappa statistic 0.55
Mean absolute error 0.1026
Root mean squared error 0.2897
Relative absolute error 48.4507 %
Root relative squared error 89.2727 %
Total Number of Instances 214

=== Detailed Accuracy By Class ===

TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class
0,714 0,174 0,667 0,714 0,690 0,532 0,806 0,667 build wind float
0,618 0,181 0,653 0,618 0,635 0,443 0,768 0,606 build wind non-float
0,353 0,046 0,400 0,353 0,375 0,325 0,766 0,251 vehic wind float

```


Figura 27 – Resultado Naive-Bayes

```

Time taken to build model: 0 seconds

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

Correctly Classified Instances      108      50.4673 %
Incorrectly Classified Instances    106      49.5327 %
Kappa statistic                    0.3276
Mean absolute error                0.1445
Root mean squared error            0.329
Relative absolute error            68.2172 %
Root relative squared error        101.3623 %
Total Number of Instances          214

=== Detailed Accuracy By Class ===

```

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0,886	0,500	0,463	0,886	0,608	0,374	0,779	0,605	build wind float
	0,118	0,072	0,474	0,118	0,189	0,077	0,772	0,545	build wind non-float
	0,059	0,041	0,111	0,059	0,077	0,025	0,699	0,137	vehic wind float
	?	0,000	?	?	?	?	?	?	vehic wind non-float
	0,231	0,030	0,333	0,231	0,273	0,239	0,876	0,392	containers
	0,889	0,029	0,571	0,889	0,696	0,698	0,990	0,731	tableware
	0,862	0,022	0,862	0,862	0,862	0,840	0,948	0,826	headlamps
Weighted Avg.	0,505	0,198	0,490	0,505	0,435	0,310	0,808	0,569	

```

=== Confusion Matrix ===

 a b c d e f g <-- classified as
62 1 4 0 0 2 1 | a = build wind float
56 9 4 0 3 3 1 | b = build wind non-float
15 0 1 0 0 1 0 | c = vehic wind float
 0 0 0 0 0 0 0 | d = vehic wind non-float
 0 9 0 0 3 0 1 | e = containers
 0 0 0 0 0 8 1 | f = tableware
 1 0 0 0 3 0 25 | g = headlamps

```

3.2.3 Método KNN

Igualmente à base anterior, é feito a seleção e configuração, obtendo os seguintes resultados.

Figura 28 – Resultados KNN

```

Time taken to build model: 0 seconds

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

Correctly Classified Instances      149          69.6262 %
Incorrectly Classified Instances    65          30.3738 %
Kappa statistic                    0.5829
Mean absolute error                0.098
Root mean squared error            0.2707
Relative absolute error            46.265 %
Root relative squared error        83.4126 %
Total Number of Instances          214

=== Detailed Accuracy By Class ===

      TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area  PRC Area  Class
      0,800    0,208    0,651     0,800    0,718     0,566    0,847     0,681    build wind float
      0,671    0,130    0,739     0,671    0,703     0,554    0,822     0,731    build wind non-float
      0,235    0,036    0,364     0,235    0,286     0,245    0,730     0,200    vehic wind float
      ?        0,000    ?         ?         ?         ?         ?         ?         vehic wind non-float
      0,615    0,015    0,727     0,615    0,667     0,650    0,954     0,613    containers
      0,778    0,020    0,636     0,778    0,700     0,689    0,953     0,603    tableware
      0,793    0,016    0,885     0,793    0,836     0,814    0,868     0,805    headlamps
Weighted Avg.  0,696    0,121    0,695     0,696    0,691     0,580    0,842     0,670

=== Confusion Matrix ===

 a b c d e f g  <-- classified as
56 9 5 0 0 0 0 | a = build wind float
19 51 2 0 1 2 1 | b = build wind non-float
 9 4 4 0 0 0 0 | c = vehic wind float
 0 0 0 0 0 0 0 | d = vehic wind non-float
 0 3 0 0 8 0 2 | e = containers
 0 1 0 0 1 7 0 | f = tableware
 2 1 0 0 1 2 23 | g = headlamps

```

3.2.4 Comparação dos Resultados

Listados os 3 resultados anteriores para análise.

Figura 29 – Resultados do ID3

```

Test mode:      10-fold cross-validation

=== Classifier model (full training set) ===

J48 pruned tree
-----

Ba <= 0.27
|   Mg <= 2.41
|   |   K <= 0.03
|   |   |   Na <= 13.75: build wind non-float (3.0)
|   |   |   Na > 13.75: tableware (9.0)
|   |   |   K > 0.03
|   |   |   |   Na <= 13.49
|   |   |   |   |   RI <= 1.5241: containers (13.0/1.0)
|   |   |   |   |   RI > 1.5241: build wind non-float (3.0)
|   |   |   |   Na > 13.49: build wind non-float (7.0/1.0)
|   Mg > 2.41
|   |   Al <= 1.41
|   |   |   RI <= 1.51707
|   |   |   |   RI <= 1.51596: build wind float (3.0)
|   |   |   |   RI > 1.51596
|   |   |   |   |   Fe <= 0.12
|   |   |   |   |   |   Mg <= 3.54: vehic wind float (5.0)
|   |   |   |   |   |   Mg > 3.54
|   |   |   |   |   |   |   RI <= 1.51667: build wind non-float (2.0)
|   |   |   |   |   |   |   RI > 1.51667: vehic wind float (2.0)
|   |   |   |   |   |   |   Fe > 0.12: build wind non-float (2.0)
|   |   |   |   RI > 1.51707
|   |   |   |   |   K <= 0.23
|   |   |   |   |   |   Mg <= 3.34: build wind non-float (2.0)
|   |   |   |   |   |   Mg > 3.34
|   |   |   |   |   |   |   Si <= 72.64
|   |   |   |   |   |   |   |   Na <= 14.01: build wind float (14.0)
|   |   |   |   |   |   |   |   Na > 14.01
|   |   |   |   |   |   |   |   |   RI <= 1.52211
|   |   |   |   |   |   |   |   |   |   Na <= 14.32: vehic wind float (3.0)
|   |   |   |   |   |   |   |   |   |   Na > 14.32: build wind float (2.0)
|   |   |   |   |   |   |   |   |   |   RI > 1.52211: build wind float (3.0)
|   |   |   |   |   |   |   |   |   |   Si > 72.64: vehic wind float (3.0)
|   |   |   |   K > 0.23
|   |   |   |   |   Mg <= 3.75
|   |   |   |   |   |   Fe <= 0.14
|   |   |   |   |   |   |   RI <= 1.52043: build wind float (36.0)
|   |   |   |   |   |   |   RI > 1.52043: build wind non-float (2.0/1.0)
|   |   |   |   |   |   |   Fe > 0.14

```

Figura 30 – Resultados do ID3

```

| | | | | | | RI <= 1.52043: build wind float (36.0)
| | | | | | | RI > 1.52043: build wind non-float (2.0/1.0)
| | | | | | | Fe > 0.14
| | | | | | | Al <= 1.17: build wind non-float (5.0)
| | | | | | | Al > 1.17: build wind float (6.0/1.0)
| | | | | | | Mg > 3.75: build wind non-float (10.0)
| | | | | | | Al > 1.41
| | | | | | | Si <= 72.49
| | | | | | | Ca <= 8.28: build wind non-float (6.0)
| | | | | | | Ca > 8.28: vehic wind float (5.0/1.0)
| | | | | | | Si > 72.49
| | | | | | | RI <= 1.51732
| | | | | | | Fe <= 0.22: build wind non-float (30.0/1.0)
| | | | | | | Fe > 0.22
| | | | | | | RI <= 1.51629: build wind float (2.0)
| | | | | | | RI > 1.51629: build wind non-float (2.0)
| | | | | | | RI > 1.51732
| | | | | | | RI <= 1.51789: build wind float (3.0)
| | | | | | | RI > 1.51789: build wind non-float (2.0)
Ba > 0.27
| | | | | | | Si <= 70.16: build wind non-float (2.0/1.0)
| | | | | | | Si > 70.16: headlamps (27.0/1.0)

Number of Leaves :    30

Size of the tree :    59

Time taken to build model: 0.01 seconds

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

Correctly Classified Instances      143           66.8224 %
Incorrectly Classified Instances    71           33.1776 %
Kappa statistic                    0.55
Mean absolute error                 0.1026
Root mean squared error             0.2897
Relative absolute error             48.4507 %
Root relative squared error         89.2727 %
Total Number of Instances          214

=== Detailed Accuracy By Class ===

      TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
      0,714    0,174    0,667     0,714    0,690      0,532    0,806    0,667    build wind float
      0,618    0,181    0,653     0,618    0,635      0,443    0,768    0,606    build wind non-float
      0,353    0,046    0,400     0,353    0,375      0,325    0,766    0,251    vehic wind float

```

Figura 31 – Resultados do ID3

```

=== Detailed Accuracy By Class ===

      TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area  PRC Area  Class
      0,714    0,174    0,667      0,714    0,690      0,532    0,806    0,667    build wind float
      0,618    0,181    0,653      0,618    0,635      0,443    0,768    0,606    build wind non-float
      0,353    0,046    0,400      0,353    0,375      0,325    0,766    0,251    vehic wind float
      ?        0,000    ?          ?          ?          ?        ?        ?        vehic wind non-float
      0,769    0,010    0,833      0,769    0,800      0,788    0,872    0,575    containers
      0,778    0,029    0,538      0,778    0,636      0,629    0,930    0,527    tableware
      0,793    0,022    0,852      0,793    0,821      0,795    0,869    0,738    headlamps
Weighted Avg. 0,668    0,130    0,670      0,668    0,668      0,539    0,807    0,611

=== Confusion Matrix ===

  a  b  c  d  e  f  g  <-- classified as
50 15  3  0  0  1  1 | a = build wind float
16 47  6  0  2  3  2 | b = build wind non-float
 5  5  6  0  0  1  0 | c = vehic wind float
 0  0  0  0  0  0  0 | d = vehic wind non-float
 0  2  0  0 10  0  1 | e = containers
 1  1  0  0  0  7  0 | f = tableware
 3  2  0  0  0  1 23 | g = headlamps

```

Figura 32 – Resultado Naive-Bayes

```

Time taken to build model: 0 seconds

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

Correctly Classified Instances      108          50.4673 %
Incorrectly Classified Instances    106          49.5327 %
Kappa statistic                    0.3276
Mean absolute error                 0.1445
Root mean squared error             0.329
Relative absolute error             68.2172 %
Root relative squared error        101.3623 %
Total Number of Instances          214

=== Detailed Accuracy By Class ===

      TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area  PRC Area  Class
      0,886    0,500    0,463      0,886    0,608      0,374    0,779    0,605    build wind float
      0,118    0,072    0,474      0,118    0,189      0,077    0,772    0,545    build wind non-float
      0,059    0,041    0,111      0,059    0,077      0,025    0,699    0,137    vehic wind float
      ?        0,000    ?          ?          ?          ?        ?        ?        vehic wind non-float
      0,231    0,030    0,333      0,231    0,273      0,239    0,876    0,392    containers
      0,889    0,029    0,571      0,889    0,696      0,698    0,990    0,731    tableware
      0,862    0,022    0,862      0,862    0,862      0,840    0,948    0,826    headlamps
Weighted Avg. 0,505    0,198    0,490      0,505    0,435      0,310    0,808    0,569

=== Confusion Matrix ===

  a  b  c  d  e  f  g  <-- classified as
62 1 4 0 0 2 1 | a = build wind float
56 9 4 0 3 3 1 | b = build wind non-float
15 0 1 0 0 1 0 | c = vehic wind float
 0  0  0  0  0  0  0 | d = vehic wind non-float
 0  9  0  0  3  0  1 | e = containers
 0  0  0  0  0  8  1 | f = tableware
 1  0  0  0  3  0 25 | g = headlamps

```


Figura 33 – Resultado KNN

```

Time taken to build model: 0 seconds

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

Correctly Classified Instances      149          69.6262 %
Incorrectly Classified Instances    65           30.3738 %
Kappa statistic                    0.5829
Mean absolute error                0.098
Root mean squared error            0.2707
Relative absolute error            46.265 %
Root relative squared error        83.4126 %
Total Number of Instances         214

=== Detailed Accuracy By Class ===

      TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area  PRC Area  Class
      0,800    0,208    0,651     0,800    0,718     0,566    0,847    0,681    build wind float
      0,671    0,130    0,739     0,671    0,703     0,554    0,822    0,731    build wind non-float
      0,235    0,036    0,364     0,235    0,286     0,245    0,730    0,200    vehic wind float
      ?        0,000    ?         ?         ?         ?         ?         ?         vehic wind non-float
      0,615    0,015    0,727     0,615    0,667     0,650    0,954    0,613    containers
      0,778    0,020    0,636     0,778    0,700     0,689    0,953    0,603    tableware
      0,793    0,016    0,885     0,793    0,836     0,814    0,868    0,805    headlamps
Weighted Avg.    0,696    0,121    0,695     0,696    0,691     0,580    0,842    0,670

=== Confusion Matrix ===

 a  b  c  d  e  f  g  <-- classified as
56  9  5  0  0  0  0 | a = build wind float
19 51  2  0  1  2  1 | b = build wind non-float
 9  4  4  0  0  0  0 | c = vehic wind float
 0  0  0  0  0  0  0 | d = vehic wind non-float
 0  3  0  0  8  0  2 | e = containers
 0  1  0  0  1  7  0 | f = tableware
 2  1  0  0  1  2 23 | g = headlamps

```

Como a partir do ID3 não foi possível identificar nenhum atributo à ser retirado. Não houve mudança na base de dados.

3.3 Hypothyroid

De maneira similar ao Cancer de Mama, esta base traz 3772 instancias e 30 atributos orientados ao Hipotireoidismo.

Figura 34 – Base Hypothyroid inicial

Current relation	
Relation: hypothyroid	Attributes: 30
Instances: 3772	Sum of weights: 3772

Attributes	
<input type="button" value="All"/>	<input type="button" value="None"/>
<input type="button" value="Invert"/>	<input type="button" value="Pattern"/>

No.	Name
1	<input type="checkbox"/> age
2	<input type="checkbox"/> sex
3	<input type="checkbox"/> on thyroxine
4	<input type="checkbox"/> query on thyroxine
5	<input type="checkbox"/> on antithyroid medication
6	<input type="checkbox"/> sick
7	<input type="checkbox"/> pregnant
8	<input type="checkbox"/> thyroid surgery
9	<input type="checkbox"/> I131 treatment
10	<input type="checkbox"/> query hypothyroid
11	<input type="checkbox"/> query hyperthyroid
12	<input type="checkbox"/> lithium
13	<input type="checkbox"/> goitre
14	<input type="checkbox"/> tumor
15	<input type="checkbox"/> hypopituitary
16	<input type="checkbox"/> psych
17	<input type="checkbox"/> TSH measured
18	<input type="checkbox"/> TSH
19	<input type="checkbox"/> T3 measured

3.3.1 Método ID3

Na mesma sequência para encontrar e configurar o método ID3, foi feito, obtendo os seguintes resultados.

Figura 35 – Resultado ID3

```

Test mode:      10-fold cross-validation

=== Classifier model (full training set) ===

J48 pruned tree
-----

TSH <= 6: negative (3366.31/2.0)
TSH > 6
|   FTI <= 64
|   |   TSH measured = t
|   |   |   T4U measured = t
|   |   |   |   thyroid surgery = f
|   |   |   |   |   T3 <= 2.3: primary_hypothyroid (82.7)
|   |   |   |   |   T3 > 2.3
|   |   |   |   |   |   TSH <= 15: negative (2.06/0.06)
|   |   |   |   |   |   TSH > 15: primary_hypothyroid (3.24)
|   |   |   |   |   |   thyroid surgery = t
|   |   |   |   |   |   |   TT4 <= 49: negative (3.0)
|   |   |   |   |   |   |   TT4 > 49: primary_hypothyroid (2.0)
|   |   |   |   |   |   |   |   T4U measured = f: compensated_hypothyroid (7.08/2.62)
|   |   |   |   |   |   |   |   TSH measured = f: negative (6.24)
|   |   FTI > 64
|   |   |   on thyroxine = f
|   |   |   |   TSH measured = t
|   |   |   |   |   thyroid surgery = f
|   |   |   |   |   |   TT4 <= 150
|   |   |   |   |   |   |   TT4 <= 48
|   |   |   |   |   |   |   |   T4U measured = t: negative (2.0/1.0)
|   |   |   |   |   |   |   |   T4U measured = f: primary_hypothyroid (3.04/0.04)
|   |   |   |   |   |   |   |   TT4 > 48: compensated_hypothyroid (191.5/3.06)
|   |   |   |   |   |   |   |   TT4 > 150: negative (9.16/0.16)
|   |   |   |   |   |   |   |   thyroid surgery = t: negative (6.74)
|   |   |   |   |   |   |   |   TSH measured = f: negative (30.75)
|   |   |   |   |   on thyroxine = t: negative (56.17)

Number of Leaves :      15

Size of the tree :      29

Time taken to build model: 0.02 seconds

```

Figura 36 – Resultado ID3

```

Time taken to build model: 0.02 seconds

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

Correctly Classified Instances      3756          99.5758 %
Incorrectly Classified Instances    16           0.4242 %
Kappa statistic                    0.9707
Mean absolute error                 0.003
Root mean squared error            0.0414
Relative absolute error             4.1612 %
Root relative squared error        21.7445 %
Total Number of Instances          3772

=== Detailed Accuracy By Class ===

      TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
      0,999    0,021    0,998    0,999    0,998    0,979    0,993    0,999    negative
      0,985    0,002    0,970    0,985    0,977    0,976    0,999    0,964    compensated_hypothyroid
      0,937    0,001    0,957    0,937    0,947    0,946    1,000    0,988    primary_hypothyroid
      0,000    0,000    ?        0,000    ?        ?        0,197    0,000    secondary_hypothyroid
Weighted Avg.  0,996    0,019    ?        0,996    ?        ?        0,993    0,996

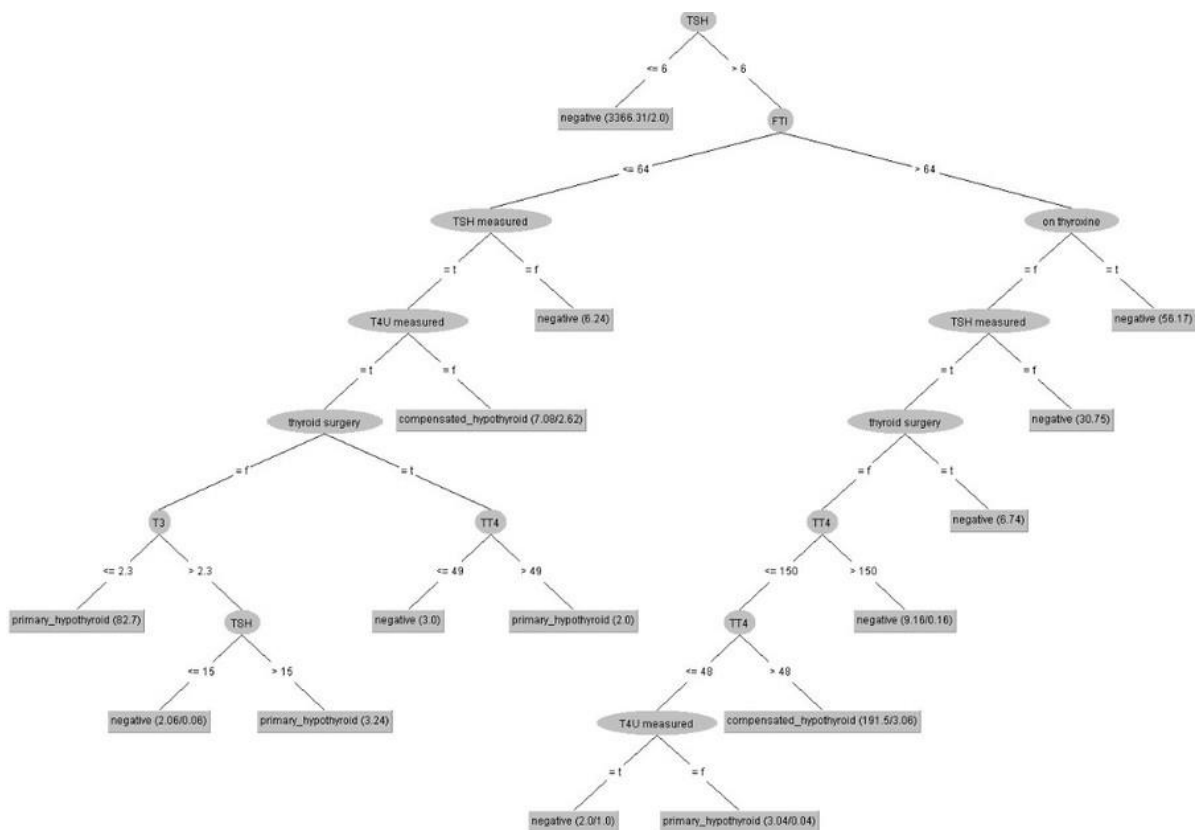
=== Confusion Matrix ===

  a  b  c  d  <-- classified as
3476 3  2  0 |  a = negative
  1 191 2  0 |  b = compensated_hypothyroid
  3  3 89  0 |  c = primary_hypothyroid
  2  0  0  0 |  d = secondary_hypothyroid

```

Obtemos também a árvore da base.

Figura 37 – Árvore ID3



3.3.2 Método Naive-Bayes

Seguinte as orientações e formas para fazer a configuração, façamos de maneira igual, obtendo os seguintes resultados.

Figura 38 – Resultado Naive-Bayes

```
Time taken to build model: 0.01 seconds

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

Correctly Classified Instances      3629           96.2089 %
Incorrectly Classified Instances    143           3.7911 %
Kappa statistic                    0.7027
Mean absolute error                 0.0276
Root mean squared error            0.1218
Relative absolute error             37.8655 %
Root relative squared error        63.9425 %
Total Number of Instances         3772

=== Detailed Accuracy By Class ===
```

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0,992	0,354	0,971	0,992	0,982	0,735	0,976	0,998	negative
	0,474	0,006	0,814	0,474	0,599	0,607	0,963	0,710	compensated_hypothyroid
	0,874	0,005	0,814	0,874	0,843	0,839	0,997	0,916	primary_hypothyroid
	0,000	0,000	?	0,000	?	?	0,109	0,000	secondary_hypothyroid
Weighted Avg.	0,962	0,327	?	0,962	?	?	0,975	0,980	

```

=== Confusion Matrix ===
  a  b  c  d  <-- classified as
3454 10 17  0 | a = negative
100  92  2  0 | b = compensated_hypothyroid
1    11 83  0 | c = primary_hypothyroid
2     0  0  0 | d = secondary_hypothyroid

```

3.3.3 Método KNN

Igualmente a realização dos metodos anteriores, fizemos neste, obtendo os seguintes resultados.

Figura 39 – Resultado KNN

```

=== Classifier model (full training set) ===

IB1 instance-based classifier
using 3 nearest neighbour(s) for classification

Time taken to build model: 0.01 seconds

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

Correctly Classified Instances      3516           93.2131 %
Incorrectly Classified Instances    256           6.7869 %
Kappa statistic                    0.3392
Mean absolute error                 0.0463
Root mean squared error            0.1766
Relative absolute error             63.524 %
Root relative squared error        92.7572 %
Total Number of Instances         3772

=== Detailed Accuracy By Class ===

```

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0,991	0,735	0,942	0,991	0,966	0,407	0,746	0,960	negative
	0,093	0,010	0,333	0,093	0,145	0,154	0,661	0,128	compensated_hypothyroid
	0,516	0,002	0,891	0,516	0,653	0,672	0,879	0,637	primary_hypothyroid
	0,000	0,000	?	0,000	?	?	0,901	0,003	secondary_hypothyroid
Weighted Avg.	0,932	0,679	?	0,932	?	?	0,745	0,909	

```

=== Confusion Matrix ===

```

	a	b	c	d	<-- classified as
3449	30	2	0	0	a = negative
172	18	4	0	0	b = compensated_hypothyroid
40	6	49	0	0	c = primary_hypothyroid
2	0	0	0	0	d = secondary_hypothyroid

Figura 40 – Resultado KNN (K=3)

```

Test mode: 10-fold cross-validation

=== Classifier model (full training set) ===

IB1 instance-based classifier
using 3 nearest neighbour(s) for classification

Time taken to build model: 0 seconds

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

Correctly Classified Instances      3568      94.5917 %
Incorrectly Classified Instances    204      5.4083 %
Kappa statistic                    0.4846
Mean absolute error                 0.0333
Root mean squared error             0.1609
Relative absolute error             45.6497 %
Root relative squared error         84.484 %
Total Number of Instances          3772

=== Detailed Accuracy By Class ===

```

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0,997	0,632	0,950	0,997	0,973	0,559	0,739	0,958	negative
	0,160	0,003	0,738	0,160	0,263	0,330	0,640	0,228	compensated_hypothyroid
	0,705	0,002	0,882	0,705	0,784	0,784	0,910	0,747	primary_hypothyroid
	0,000	0,000	?	0,000	?	?	0,900	0,003	secondary_hypothyroid
Weighted Avg.	0,946	0,584	?	0,946	?	?	0,739	0,915	

```

=== Confusion Matrix ===

  a   b   c   d  <-- classified as
3470   6   5   0 |  a = negative
 160  31   3   0 |  b = compensated_hypothyroid
  23   5  67   0 |  c = primary_hypothyroid
   1   0   1   0 |  d = secondary_hypothyroid

```

3.3.4 Comparação dos Resultados

Figura 41 – Resultado ID3

```

Test mode:      10-fold cross-validation

=== Classifier model (full training set) ===

J48 pruned tree
-----

TSH <= 6: negative (3366.31/2.0)
TSH > 6
|   FII <= 64
|   |   TSH measured = t
|   |   |   T4U measured = t
|   |   |   |   thyroid surgery = f
|   |   |   |   |   T3 <= 2.3: primary_hypothyroid (82.7)
|   |   |   |   |   T3 > 2.3
|   |   |   |   |   |   TSH <= 15: negative (2.06/0.06)
|   |   |   |   |   |   TSH > 15: primary_hypothyroid (3.24)
|   |   |   |   |   |   thyroid surgery = t
|   |   |   |   |   |   TT4 <= 49: negative (3.0)
|   |   |   |   |   |   TT4 > 49: primary_hypothyroid (2.0)
|   |   |   |   |   T4U measured = f: compensated_hypothyroid (7.08/2.62)
|   |   |   |   TSH measured = f: negative (6.24)
|   |   FII > 64
|   |   |   on thyroxine = f
|   |   |   |   TSH measured = t
|   |   |   |   |   thyroid surgery = f
|   |   |   |   |   |   TT4 <= 150
|   |   |   |   |   |   |   TT4 <= 48
|   |   |   |   |   |   |   |   T4U measured = t: negative (2.0/1.0)
|   |   |   |   |   |   |   |   T4U measured = f: primary_hypothyroid (3.04/0.04)
|   |   |   |   |   |   |   |   TT4 > 48: compensated_hypothyroid (191.5/3.06)
|   |   |   |   |   |   |   |   TT4 > 150: negative (9.16/0.16)
|   |   |   |   |   |   |   thyroid surgery = t: negative (6.74)
|   |   |   |   |   TSH measured = f: negative (30.75)
|   |   |   on thyroxine = t: negative (56.17)

Number of Leaves :      15

Size of the tree :      29

Time taken to build model: 0.02 seconds

```

Figura 42 – Resultado ID3

```

Time taken to build model: 0.02 seconds

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

Correctly Classified Instances      3756          99.5758 %
Incorrectly Classified Instances    16           0.4242 %
Kappa statistic                    0.9707
Mean absolute error                 0.003
Root mean squared error             0.0414
Relative absolute error             4.1612 %
Root relative squared error         21.7445 %
Total Number of Instances          3772

=== Detailed Accuracy By Class ===

          TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area  PRC Area  Class
          0,999   0,021   0,998     0,999   0,998     0,979   0,993     0,999   negative
          0,985   0,002   0,970     0,985   0,977     0,976   0,999     0,964   compensated_hypothyroid
          0,937   0,001   0,957     0,937   0,947     0,946   1,000     0,988   primary_hypothyroid
          0,000   0,000   ?         0,000   ?         ?       0,197     0,000   secondary_hypothyroid
Weighted Avg.   0,996   0,019   ?         0,996   ?         ?       0,993     0,996

=== Confusion Matrix ===

  a   b   c   d  <-- classified as
3476  3   2   0 |  a = negative
  1 191   2   0 |  b = compensated_hypothyroid
  3   3  89   0 |  c = primary_hypothyroid
  2   0   0   0 |  d = secondary_hypothyroid

```

Figura 43 – Resultado Naive-Bayes

```

Time taken to build model: 0.01 seconds

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

Correctly Classified Instances      3629          96.2089 %
Incorrectly Classified Instances    143           3.7911 %
Kappa statistic                    0.7027
Mean absolute error                 0.0276
Root mean squared error             0.1218
Relative absolute error             37.8655 %
Root relative squared error         63.9425 %
Total Number of Instances          3772

=== Detailed Accuracy By Class ===

          TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area  PRC Area  Class
          0,992   0,354   0,971     0,992   0,982     0,735   0,976     0,998   negative
          0,474   0,006   0,814     0,474   0,599     0,607   0,963     0,710   compensated_hypothyroid
          0,874   0,005   0,814     0,874   0,843     0,839   0,997     0,916   primary_hypothyroid
          0,000   0,000   ?         0,000   ?         ?       0,109     0,000   secondary_hypothyroid
Weighted Avg.   0,962   0,327   ?         0,962   ?         ?       0,975     0,980

=== Confusion Matrix ===

  a   b   c   d  <-- classified as
3454 10 17   0 |  a = negative
 100  92  2   0 |  b = compensated_hypothyroid
  1  11  83   0 |  c = primary_hypothyroid
  2   0   0   0 |  d = secondary_hypothyroid

```


Figura 44 – Resultado KNN

```

=== Classifier model (full training set) ===

IB1 instance-based classifier
using 3 nearest neighbour(s) for classification

Time taken to build model: 0.01 seconds

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

Correctly Classified Instances      3516           93.2131 %
Incorrectly Classified Instances    256           6.7869 %
Kappa statistic                    0.3392
Mean absolute error                 0.0463
Root mean squared error             0.1766
Relative absolute error             63.524 %
Root relative squared error         92.7572 %
Total Number of Instances          3772

=== Detailed Accuracy By Class ===

```

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0,991	0,735	0,942	0,991	0,966	0,407	0,746	0,960	negative
	0,093	0,010	0,333	0,093	0,145	0,154	0,661	0,128	compensated_hypothyroid
	0,516	0,002	0,891	0,516	0,653	0,672	0,879	0,637	primary_hypothyroid
	0,000	0,000	?	0,000	?	?	0,901	0,003	secondary_hypothyroid
Weighted Avg.	0,932	0,679	?	0,932	?	?	0,745	0,909	

```

=== Confusion Matrix ===

  a   b   c   d  <-- classified as
3449  30   2   0 |  a = negative
 172  18   4   0 |  b = compensated_hypothyroid
  40   6  49   0 |  c = primary_hypothyroid
   2   0   0   0 |  d = secondary_hypothyroid

```

Após as classificações obtivemos a base limpa

Figura 45 – Base Limpa

Current relation

Relation: hypothyroid-weka.filters.unsupervise... Attributes: 10
 Instances: 3772 Sum of weights: 3772

Attributes

All None Invert Pattern

No.		Name
1	<input checked="" type="checkbox"/>	on thyroxine
2	<input type="checkbox"/>	thyroid surgery
3	<input type="checkbox"/>	TSH measured
4	<input type="checkbox"/>	TSH
5	<input type="checkbox"/>	T3
6	<input type="checkbox"/>	TT4
7	<input type="checkbox"/>	T4U measured
8	<input type="checkbox"/>	T4U
9	<input type="checkbox"/>	FTI
10	<input type="checkbox"/>	Class

Após a limpeza tanto o Naive-Bayes quanto o KNN teve uma melhora na classificação, Naive foi de 95% para 96% o KNN se manteve nos mesmos 93%.

3.4 Ionosphaera

Esta base traz 351 instancias e 35 atributos, é um conjunto de dados sobre radares que mostram evidências de algum tipo de estrutura na ionosfera.

Figura 46 – Base Ionosphaera inicial.

Current relation
Relation: ionosphere
Instances: 351

Attributes: 35
Sum of weights: 351

Attributes

AllNoneInvertPattern

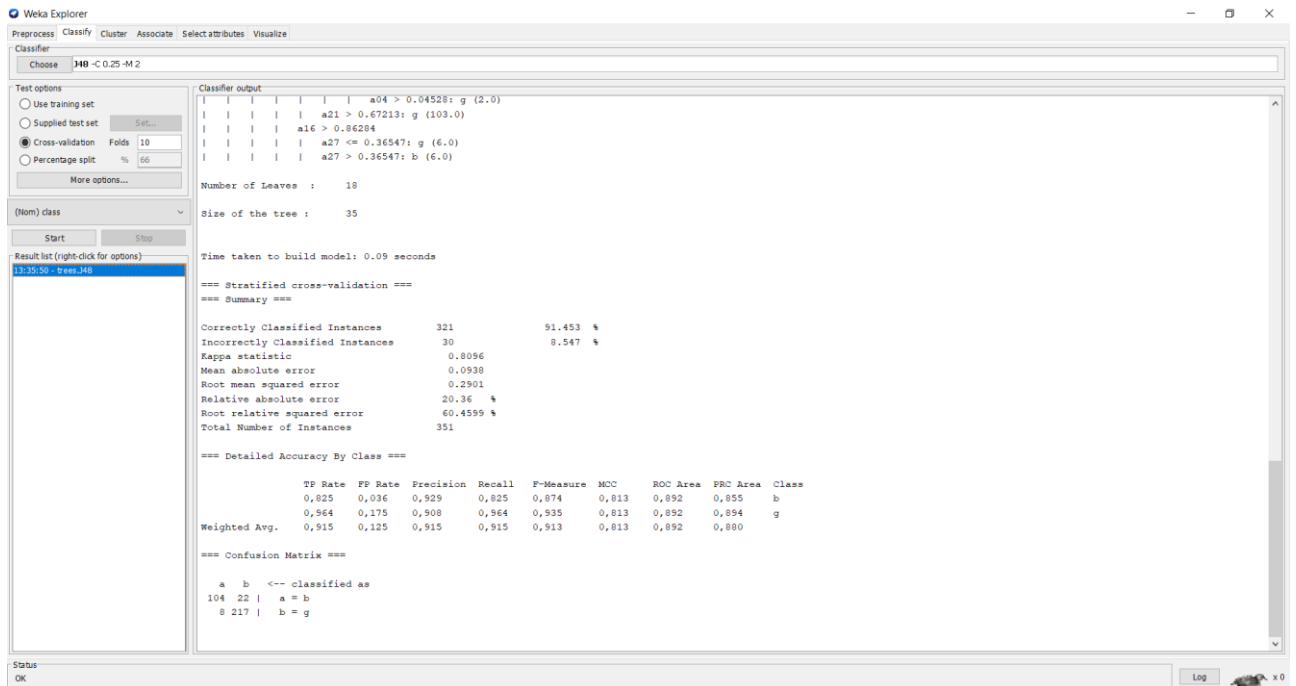
No.	Name
1	a01
2	a02
3	a03
4	a04
5	a05
6	a06
7	a07
8	a08
9	a09
10	a10
11	a11
12	a12
13	a13
14	a14
15	a15
16	a16
17	a17
18	a18
19	a19
20	a20
21	a21
22	a22
23	a23
24	a24
25	a25
26	a26
27	a27
28	a28
29	a29
30	a30

Remove

3.4.1 Método ID3

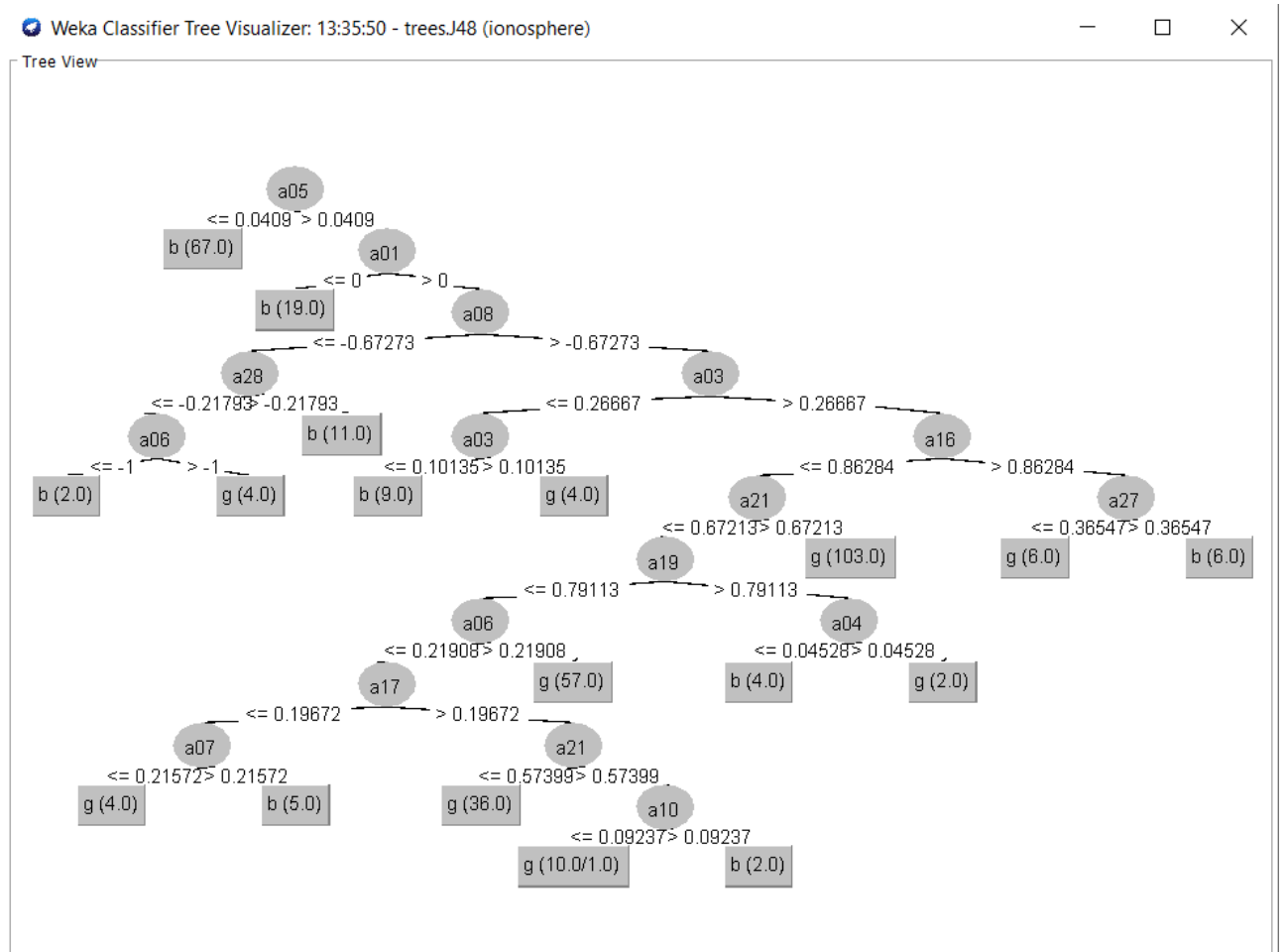
Na mesma sequencia para encontrar e configurar o método ID3, foi feito, obtendo os seguintes resultados.

Figura 47 – Base Ionosphaera ID3.



Obtemos também a árvore da base.

Figura 48 – Base Ionosfera árvore.



3.4.2 Método Naive-Bayes

Da mesma forma da base anterior, fazemos a mesma seleção, e a mesma configuração, obtendo os seguintes resultados.

Figura 49 – Base Ionosphaera Naive-Bayes.

```
Time taken to build model: 0.01 seconds

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

Correctly Classified Instances      290           82.6211 %
Incorrectly Classified Instances    61           17.3789 %
Kappa statistic                    0.6394
Mean absolute error                 0.1736
Root mean squared error             0.3935
Relative absolute error             37.7001 %
Root relative squared error         82.0203 %
Total Number of Instances          351

=== Detailed Accuracy By Class ===
```

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0,865	0,196	0,712	0,865	0,781	0,648	0,935	0,917	b
	0,804	0,135	0,914	0,804	0,856	0,648	0,935	0,958	g
Weighted Avg.	0,826	0,157	0,842	0,826	0,829	0,648	0,935	0,943	

```

=== Confusion Matrix ===
  a  b  <-- classified as
109 17 |  a = b
 44 181 |  b = g

```

3.4.3 Método KNN

Igualmente à base anterior, é feito a seleção e configuração, obtendo os seguintes resultados.

Figura 49 – Resultado KNN (K=3)

```

=== Classifier model (full training set) ===

IB1 instance-based classifier
using 3 nearest neighbour(s) for classification

Time taken to build model: 0 seconds

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

Correctly Classified Instances      304           86.6097 %
Incorrectly Classified Instances    47           13.3903 %
Kappa statistic                    0.6878
Mean absolute error                 0.1441
Root mean squared error             0.3321
Relative absolute error             31.3017 %
Root relative squared error         69.2323 %
Total Number of Instances          351

=== Detailed Accuracy By Class ===

                TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                0,659    0,018    0,954      0,659    0,779      0,712    0,885    0,837     b
                0,982    0,341    0,837      0,982    0,904      0,712    0,885    0,890     g
Weighted Avg.   0,866    0,225    0,879      0,866    0,859      0,712    0,885    0,871

=== Confusion Matrix ===

  a  b  <-- classified as
83 43 |  a = b
 4 221 |  b = g

```

3.4.4 Limpando a base através dos resultados relevantes do ID3 obtemos.

Figura 50 – Resultado base limpa

Current relation

Relation: ionosphere-weka.filters.unsupervised.attribute.Remove-R2,9,11-15,18,20,22-26,29-34

Instances: 351

Attributes: 15

Sum of weights: 351

Attributes

AllNoneInvertPattern

No.		Name
1	<input checked="" type="checkbox"/>	a01
2	<input type="checkbox"/>	a03
3	<input type="checkbox"/>	a04
4	<input type="checkbox"/>	a05
5	<input type="checkbox"/>	a06
6	<input type="checkbox"/>	a07
7	<input type="checkbox"/>	a08
8	<input type="checkbox"/>	a10
9	<input type="checkbox"/>	a16
10	<input type="checkbox"/>	a17
11	<input type="checkbox"/>	a19
12	<input type="checkbox"/>	a21
13	<input type="checkbox"/>	a27
14	<input type="checkbox"/>	a28
15	<input type="checkbox"/>	class

Sobrando 15 atributos e 351 instâncias.

3.4.5 Método Naive-Bayes

Após retirar os atributos e rodar o Naive-Bayes obtém o seguinte resultado.

Figura 50 – Naive-bayes base limpa.

```
Time taken to build model: 0 seconds

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

Correctly Classified Instances      318          90.5983 %
Incorrectly Classified Instances    33           9.4017 %
Kappa statistic                    0.7968
Mean absolute error                 0.0984
Root mean squared error             0.2794
Relative absolute error             21.3708 %
Root relative squared error         58.2495 %
Total Number of Instances          351

=== Detailed Accuracy By Class ===
```

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0,881	0,080	0,860	0,881	0,871	0,797	0,959	0,953	b
	0,920	0,119	0,932	0,920	0,926	0,797	0,959	0,972	g
Weighted Avg.	0,906	0,105	0,907	0,906	0,906	0,797	0,959	0,965	

```

=== Confusion Matrix ===

  a    b  <-- classified as
111  15 |   a = b
 18 207 |   b = g

```

3.4.6 Método KNN

Após retirar os atributos e rodar o KNN com k=3 obtém o seguinte resultado.

Figura 51 – KNN com base limpa.

```

=== Classifier model (full training set) ===

IB1 instance-based classifier
using 3 nearest neighbour(s) for classification

Time taken to build model: 0 seconds

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

Correctly Classified Instances      309           88.0342 %
Incorrectly Classified Instances    42           11.9658 %
Kappa statistic                    0.7246
Mean absolute error                 0.1305
Root mean squared error             0.3061
Relative absolute error             28.3365 %
Root relative squared error         63.8035 %
Total Number of Instances          351

=== Detailed Accuracy By Class ===

                TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                0,706    0,022    0,947      0,706    0,809      0,741    0,911    0,872     b
                0,978    0,294    0,856      0,978    0,913      0,741    0,911    0,914     g
Weighted Avg.   0,880    0,196    0,889      0,880    0,876      0,741    0,911    0,899

=== Confusion Matrix ===

  a  b  <-- classified as
89  37 |  a = b
 5 220 |  b = g

```

Podemos observar que a que após a limpeza tanto Naive-Bayes e o KNN teve uma melhoria na classificação. O KNN foi de 87% para 88% e o Naive-Bayes foi de 82% para 90%. E o ID3 teve a maior classificação com 91%.