Classificação
Recuperação de Informação

Renan Stephano Barbosa de Souza Rodrigues

Etapas

1- remover html tags:

```
public static String html2text(String html) {
    return Jsoup.parse(html).text();
}
```

http://stackoverflow.com/questions/240546/remove-html-tags-from-a-string

2 - retirar os stop words

http://stackoverflow.com/questions/35319544/removing-stopwords-java

remover usando expressão regular:

```
// instead of the ".....", add all your stopwords, separated by "|" // "\\b" is to account for word
 boundaries, i.e. not replace "his" in "this" // the "\\s?" is to suppress optional trailing white space
 Pattern p = Pattern.compile("\\b(I|this|its.....)\\b\\s?"); Matcher m = p.matcher("I love this phone, its
 super fast and there's so much new and cool things with jelly bean....but of recently I've seen some bugs.");
 String s = m.replaceAll(""); System.out.println(s);
```

Etapas

2- criar um bag of words

https://www.youtube.com/watch?v=jSZ9jQy1sfE

```
@relation relevates

@attribute Document string
@attribute relevante {yes, no}


@data

"... string... ", no
"... string... ", yes...
```

Etapas

3-Usar Principais classificadores

– Naïve bayes
– Decision tree (J48)
– SVM (SMO)
– LogisTc regression (logisTc)
– MulTlayer perceptron

5- treino de validação cruzada
https://www.youtube.com/watch?v=72LXnrT0qIY

Exemplo Seleção dos sites

http://www.lebes.com.br/

relevante

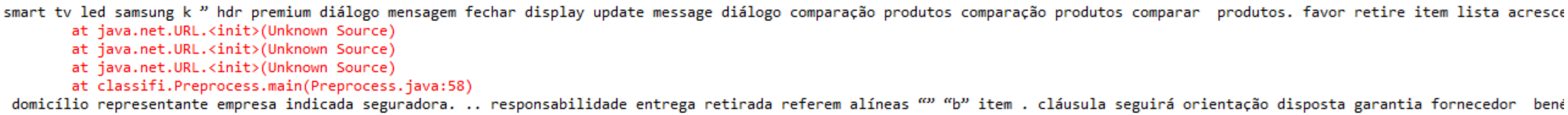
1. http://www.lebes.com.br/som-e-video/tvs-led
2. http://www.lebes.com.br/tv-led-24--semp-l24d27-conversor-digital/p
3. http://www.lebes.com.br/smart-tv-led-58-samsung-un58h5203agxzd-full-hd-com-conversor-digital-563357/p
4. http://www.lebes.com.br/tv-led-40--full-hd-samsung-hg40nd450bgxzd-usb-hdmi-modo-filme---bivolt-560432/p
5. http://www.lebes.com.br/smart-tv-led-32-hd-samsung-un32j4300-usb-hdmi-wi-fi-conversor-digital-558858/p
6. http://www.lebes.com.br/smart-tv-led-43--lg-43lh5700-full-hd-com-conversor-digital-565630/p
7. http://www.lebes.com.br/smart-tv-led-43--semp-toshiba-43l2500-full-hd-com-conversor-digital-565035/p
8. http://www.lebes.com.br/tv-samsung-led-32--hd-samsung-hg32nd450sgxzd-hdmi-usb---conversor-digital---bivolt-560431/p
9. http://www.lebes.com.br/tv-led-14--semp-le1473-hdmi-conversor-digital/p
10. http://www.lebes.com.br/smart-tv-led-semp-toshiba-tcl-49--l49s4900fs-full-hd---com-conversor-digital/p

não relevante

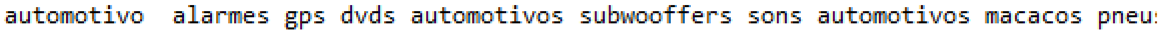
1. https://s3-sa-east-1.amazonaws.com/cdn.siteblindado.com/lp_aw/verifica-pt-br.html?url=www.lebes.com.br
2. http://www.ebit.com.br/lojas-lebes
3. http://www.lebes.com.br/brv

Problemas

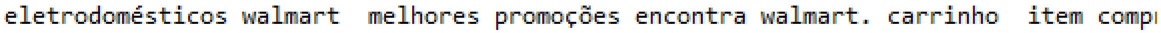
A maioria dos grandes portais estavam blindados.



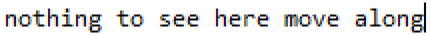
```
smart tv led samsung k ” hdr premium diálogo mensagem fechar display update message diálogo comparação produtos comparação produtos comparar  produtos. favor retire item lista acresce
	at java.net.URL.<init>(Unknown Source)
	at java.net.URL.<init>(Unknown Source)
	at java.net.URL.<init>(Unknown Source)
	at classifi.Preprocess.main(Preprocess.java:58)
 domicílio representante empresa indicada seguradora. .. responsabilidade entrega retirada referem alíneas “” “b” item . cláusula seguirá orientação disposta garantia fornecedor  bené
```



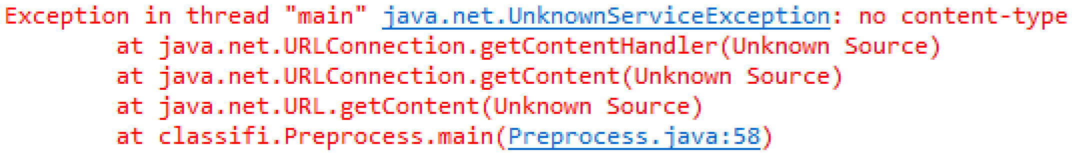
```
automotivo  alarmes gps dvds automotivos subwooffers sons automotivos macacos pneu
```



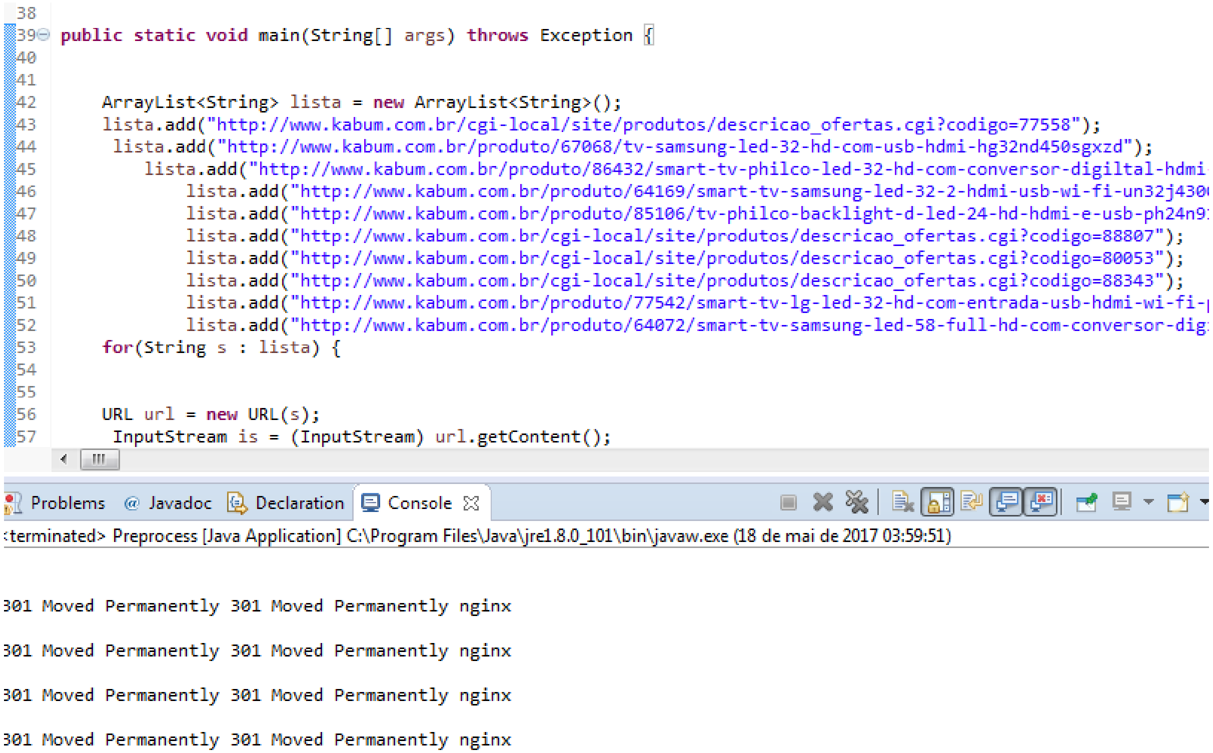
```
eletrodomésticos walmart  melhores promoções encontra walmart. carrinho  item comp
```



```
nothing to see here move along
```



```
Exception in thread "main" java.net.UnknownServiceException: no content-type
	at java.net.URLConnection.getContentHandler(Unknown Source)
	at java.net.URLConnection.getContent(Unknown Source)
	at java.net.URL.getContent(Unknown Source)
	at classifi.Preprocess.main(Preprocess.java:58)
```

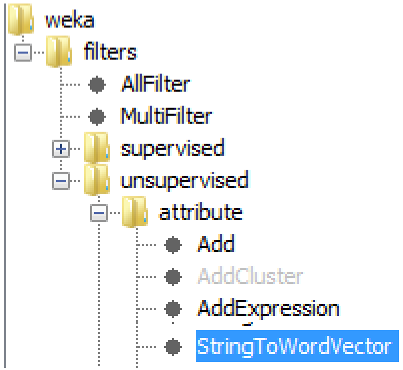


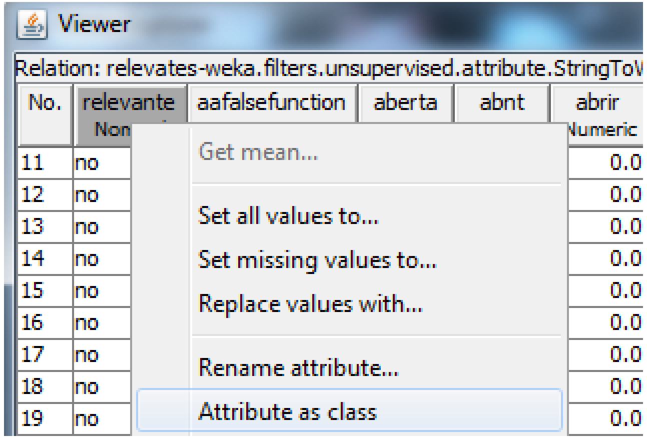
```java

public static void main(String[] args) throws Exception {


    ArrayList<String> lista = new ArrayList<String>();
    lista.add("http://www.kabum.com.br/cgi-local/site/produtos/descricao_ofertas.cgi?codigo=77558");
     lista.add("http://www.kabum.com.br/produto/67068/tv-samsung-led-32-hd-com-usb-hdmi-hg32nd450sgxzd");
        lista.add("http://www.kabum.com.br/produto/86432/smart-tv-philco-led-32-hd-com-conversor-digiltal-hdmi
            lista.add("http://www.kabum.com.br/produto/64169/smart-tv-samsung-led-32-2-hdmi-usb-wi-fi-un32j430
            lista.add("http://www.kabum.com.br/produto/85106/tv-philco-backlight-d-led-24-hd-hdmi-e-usb-ph24n9
            lista.add("http://www.kabum.com.br/cgi-local/site/produtos/descricao_ofertas.cgi?codigo=88807");
            lista.add("http://www.kabum.com.br/cgi-local/site/produtos/descricao_ofertas.cgi?codigo=80053");
            lista.add("http://www.kabum.com.br/cgi-local/site/produtos/descricao_ofertas.cgi?codigo=88343");
            lista.add("http://www.kabum.com.br/produto/77542/smart-tv-lg-led-32-hd-com-entrada-usb-hdmi-wi-fi-
            lista.add("http://www.kabum.com.br/produto/64072/smart-tv-samsung-led-58-full-hd-com-conversor-dig
    for(String s : lista) {


    URL url = new URL(s);
     InputStream is = (InputStream) url.getContent();
```

<terminated> Preprocess [Java Application] C:\Program Files\Java\jre1.8.0_101\bin\javaw.exe (18 de mai de 2017 03:59:51)

```
301 Moved Permanently 301 Moved Permanently nginx

301 Moved Permanently 301 Moved Permanently nginx

301 Moved Permanently 301 Moved Permanently nginx

301 Moved Permanently 301 Moved Permanently nginx
```

Criação do bag of words *(Usando o Weka)

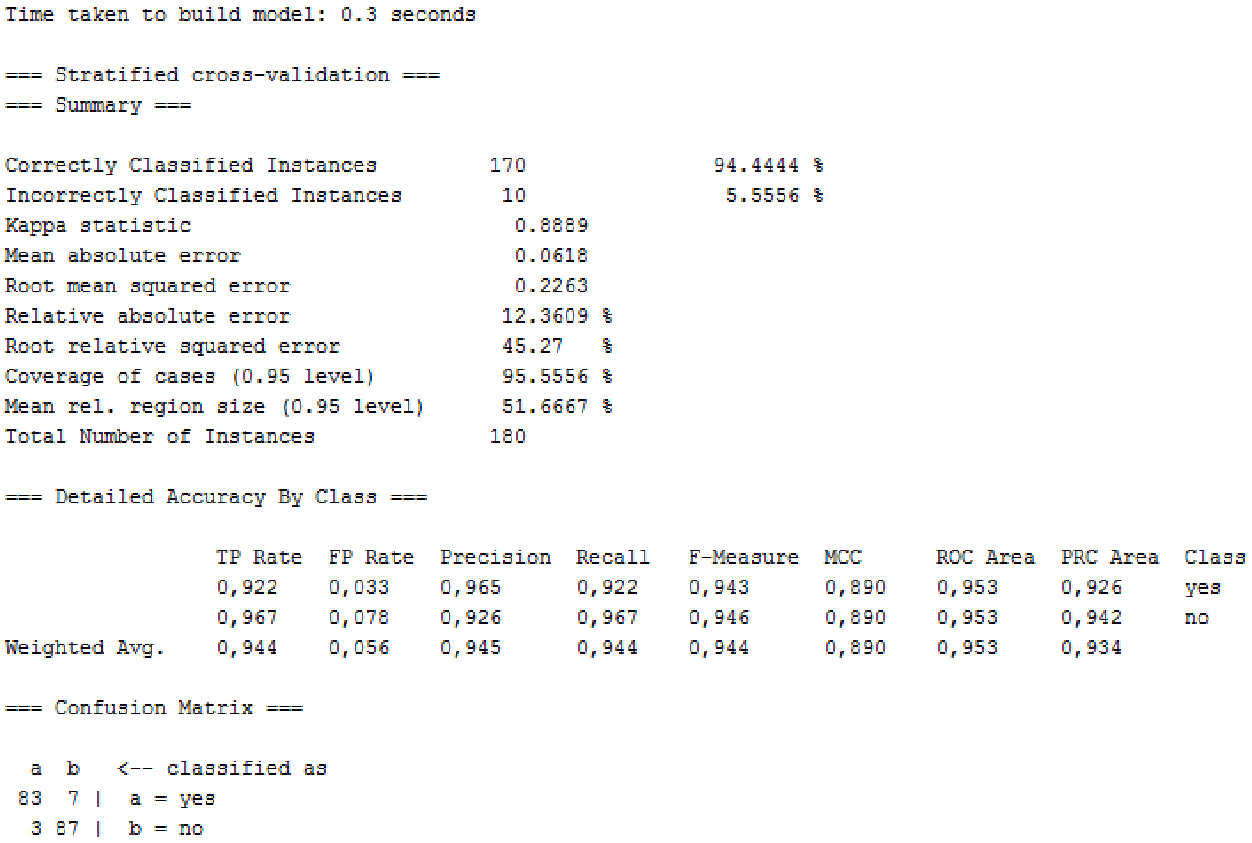




Relation: relevates-weka.filters.unsupervised.attribute.StringToW

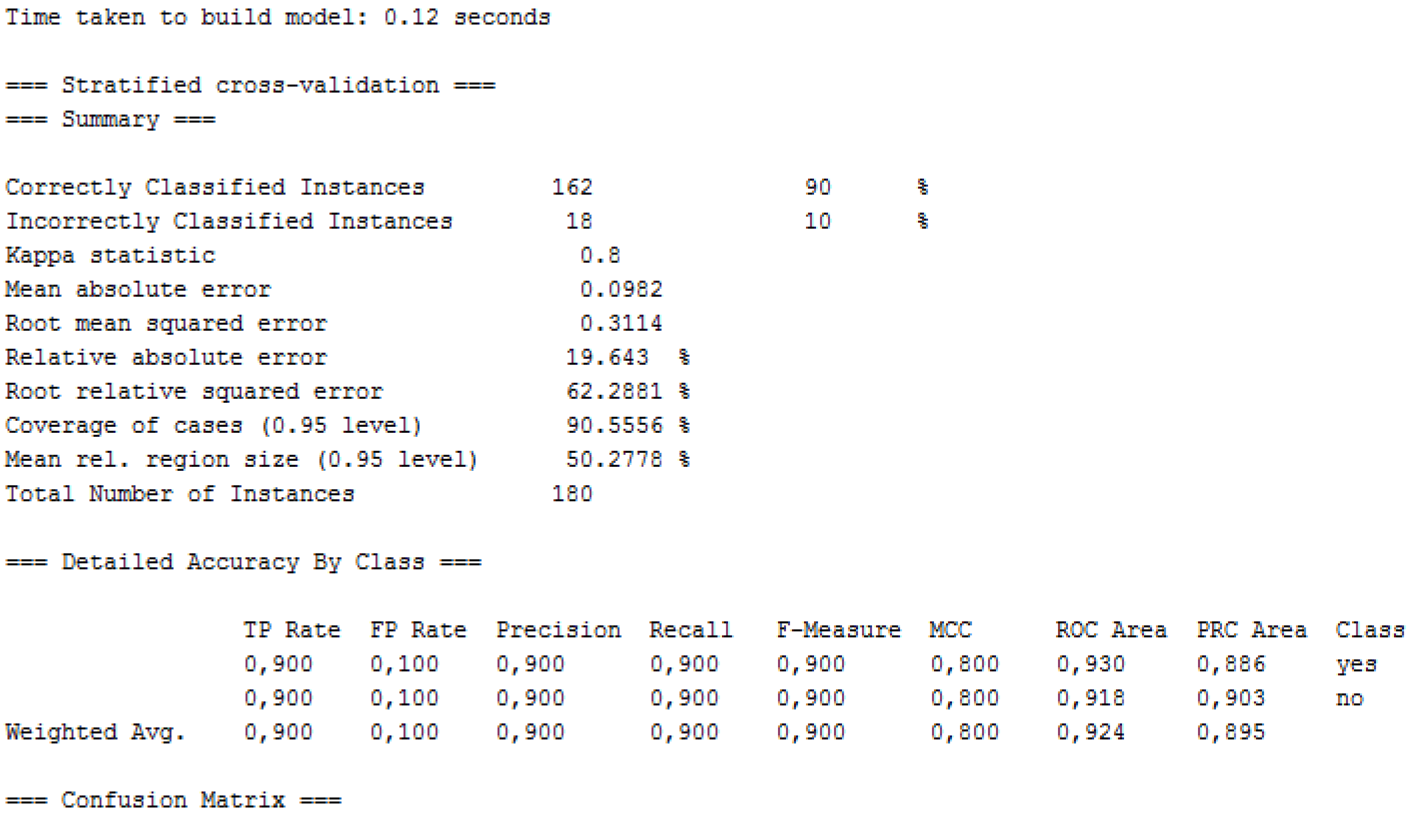
No.	relevante Nom	aafalsefunction	aberta	abnt	abrir Numeric
11	no				0.0
12	no				0.0
13	no				0.0
14	no				0.0
15	no				0.0
16	no				0.0
17	no				0.0
18	no				0.0
19	no				0.0

J48



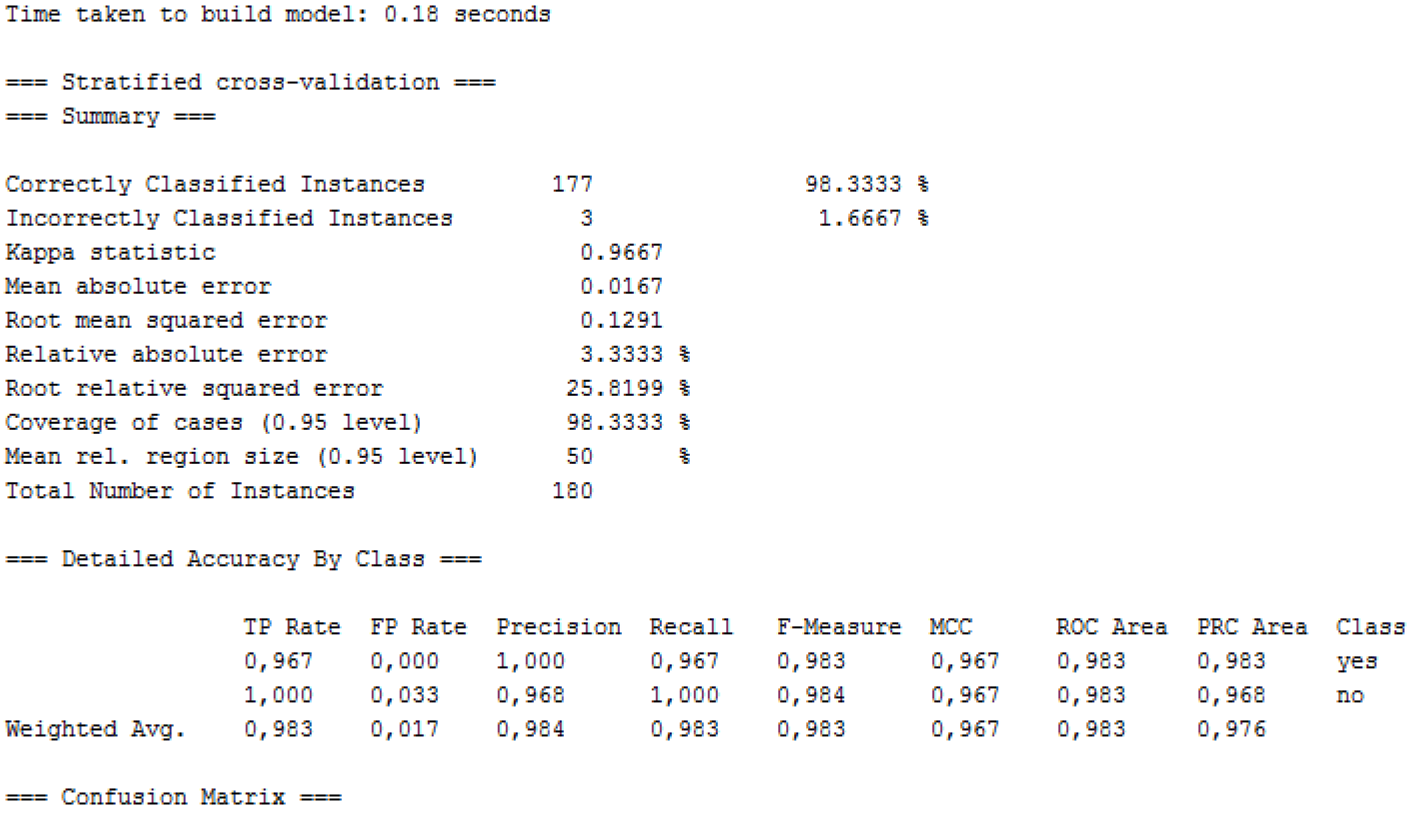
```
Time taken to build model: 0.3 seconds

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

Correctly Classified Instances         170               94.4444 %
Incorrectly Classified Instances        10                5.5556 %
Kappa statistic                          0.8889
Mean absolute error                      0.0618
Root mean squared error                  0.2263
Relative absolute error                 12.3609 %
Root relative squared error             45.27   %
Coverage of cases (0.95 level)          95.5556 %
Mean rel. region size (0.95 level)      51.6667 %
Total Number of Instances              180

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,922    0,033    0,965      0,922    0,943      0,890    0,953     0,926     yes
                 0,967    0,078    0,926      0,967    0,946      0,890    0,953     0,942     no
Weighted Avg.    0,944    0,056    0,945      0,944    0,944      0,890    0,953     0,934

=== Confusion Matrix ===

  a  b   <-- classified as
 83  7 |  a = yes
  3 87 |  b = no
```

Naive Bayer



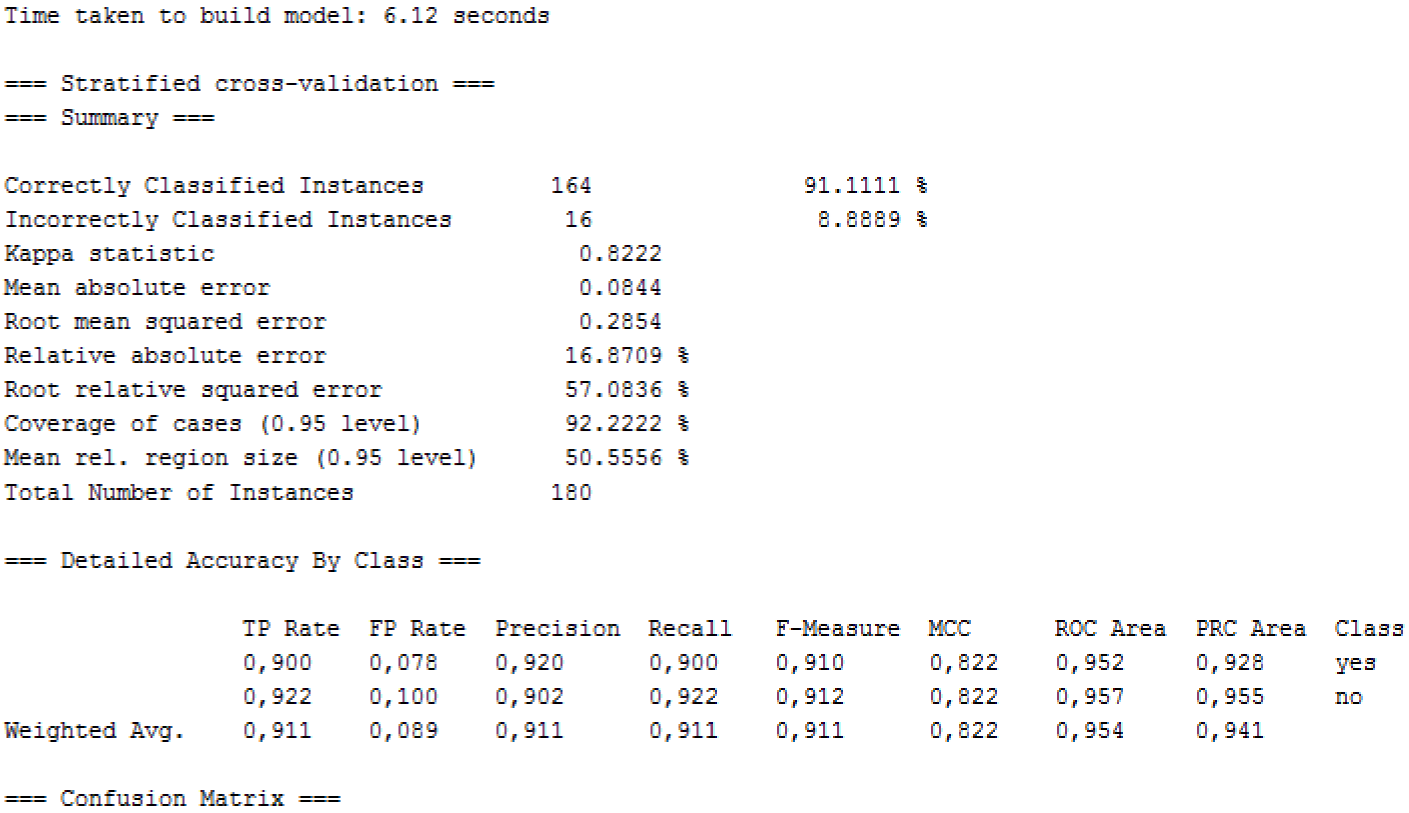
```
Time taken to build model: 0.12 seconds

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

Correctly Classified Instances         162               90      %
Incorrectly Classified Instances        18               10      %
Kappa statistic                          0.8
Mean absolute error                      0.0982
Root mean squared error                  0.3114
Relative absolute error                 19.643  %
Root relative squared error             62.2881 %
Coverage of cases (0.95 level)          90.5556 %
Mean rel. region size (0.95 level)      50.2778 %
Total Number of Instances              180

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,900    0,100    0,900      0,900    0,900      0,800    0,930     0,886     yes
                 0,900    0,100    0,900      0,900    0,900      0,800    0,918     0,903     no
Weighted Avg.    0,900    0,100    0,900      0,900    0,900      0,800    0,924     0,895

=== Confusion Matrix ===
```

SMO



```
Time taken to build model: 0.18 seconds

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

Correctly Classified Instances         177               98.3333 %
Incorrectly Classified Instances         3                1.6667 %
Kappa statistic                          0.9667
Mean absolute error                      0.0167
Root mean squared error                  0.1291
Relative absolute error                  3.3333 %
Root relative squared error             25.8199 %
Coverage of cases (0.95 level)          98.3333 %
Mean rel. region size (0.95 level)      50      %
Total Number of Instances              180

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,967    0,000    1,000      0,967    0,983      0,967    0,983     0,983     yes
                 1,000    0,033    0,968      1,000    0,984      0,967    0,983     0,968     no
Weighted Avg.    0,983    0,017    0,984      0,983    0,983      0,967    0,983     0,976

=== Confusion Matrix ===
```

Logist Model



```
Time taken to build model: 6.12 seconds

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

Correctly Classified Instances         164               91.1111 %
Incorrectly Classified Instances        16                8.8889 %
Kappa statistic                          0.8222
Mean absolute error                      0.0844
Root mean squared error                  0.2854
Relative absolute error                 16.8709 %
Root relative squared error             57.0836 %
Coverage of cases (0.95 level)          92.2222 %
Mean rel. region size (0.95 level)      50.5556 %
Total Number of Instances              180

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,900    0,078    0,920      0,900    0,910      0,822    0,952     0,928     yes
                 0,922    0,100    0,902      0,922    0,912      0,822    0,957     0,955     no
Weighted Avg.    0,911    0,089    0,911      0,911    0,911      0,822    0,954     0,941

=== Confusion Matrix ===
```