ASSIGNMENT SUBMISSION FORM

Assignment Title: Classifying Internet Advertisements
Submitted by: Group 1
(Student name and group name)

CONTENTS

SUMMARY

Detecting Advertisements on the Web

This dataset represents a set of possible advertisements on Internet pages. The features encode the geometry of the image (if available) as well as phrases occuring in the URL, the image's URL and alt text, the anchor text, and words occuring near the anchor text. The task is to predict whether an image is an advertisement ("ad") or not ("nonad").
It uses innovative data from Nick Kushmerick. There are 3279 cases, each describing an image within an anchor tag in a HTML document. About 14% of these anchored images are banner advertisements, and the goal is to generate rules that predict whether an image is an ad. (Kushmerick's system *AdEater* uses this prediction to eliminate advertisement images and so speed up page downloading.)

This dataset is very high-dimensional -- there are 1558 attributes, about half the number of cases! These features include three numbers -- image height, width, and aspect ratio -- together with boolean features representing the presence or absence of phrases in the image caption, its *alt* tag, and the anchor, image, and base URLs. For example, the attribute ***ancurl*http+www*** has the value 1 if the URL referred to in the anchor contains *http* followed by *www* (ignoring punctuation). More than a quarter of the cases have unknown values for one or more of the attributes.

Given a set of *training instances* that are preclassified as being an advertisement (AD) or not (AD), the goal is to learn a *classifier* that maps instances to either AD or AD.

Dataset Used for A Program Called as ADEATER

AdEater is a fully implemented browsing assistant that automatically removes advertisement images from Internet pages. Unlike related systems that rely on hand-crafted rules, AdEater takes an inductive learning approach, automatically generating rules from training examples. Our experiments demonstrate that our approach is practical: the off-line training phase takes less than six minutes; on-line classification takes about 70 msec; and classification accuracy exceeds 97% given a modest set of training data.

Dataset Attributes

1-) Each image enclosed in an <A> tag is a candidate advertisement; non-anchor images are rarely advertisements, and are therefore ignored. Let Udest. be the URL to which the anchor points, and let Uimg be the image's URL.

2-) Three numeric features capture geometric infor-mation about the image: **height** , **width** , and **aspect ratio** (ratio of width to height). These features are drawn directly from the HTML file, not the image. Therefore, these features might be missing (indicated by " ? ") if the corresponding < IMG > tag does indicate the height or width. For example, no geometric features can be extracted for instance C.

3-) A single binary feature local? indicates whether Udest 's and U img 's servers are in the same Internet domain. For example, if Udest = a.host.com/-page.html, then local? is 1 for U img=b.host.-com/image.jpg, but 0 for U img=elsewhere.org/-picture.gif.

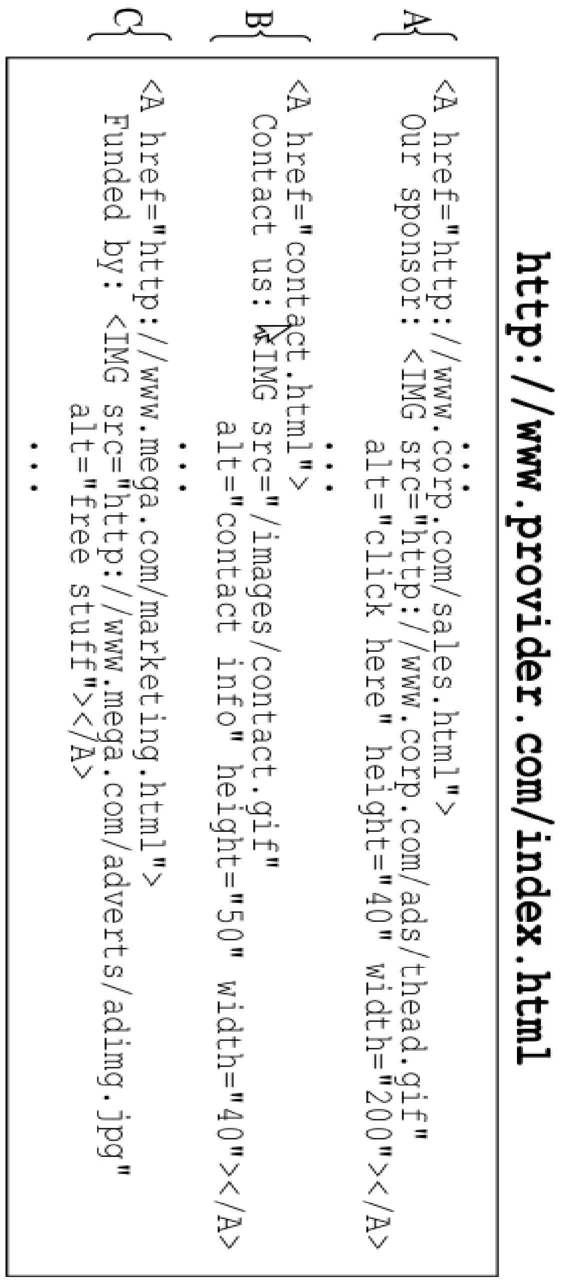
4-) An instance's caption is the words occurring in the enclosing <A> tag, ignoring punctuation and case. A set of binary features encode each caption word, each two-word phrase, and so on, through K -word phrases. Caption features are then dis- carded if the phrase occurs fewer than at M times in the training set. For example, the caption feature "funded + by" is 1 for instances whose caption contains this two-word phrase (instance C only, in the example). Note that the specific caption features generated depend on the particular training instances; feature vectors have a fixed width respect to a given set of training instances.

5-) An instance's alt text is the set of "alternate" words in the < IMG > tag. As with captions, the encoding contains one boolean feature for phrase of length each 1, 2, . . . , K that occurs at least M times.

6-) Additional sets of features are provided by the base URL Ubase , the destination URL Udest , and the image URL Uimg. For each of these URL s, one binary feature corresponds to the servername. Then, punctuation and case are discarded in the rest of the URL , and (like caption and alt text), a set of binary features encode phrase of length 1, 2, . . . , K that occurrs at least M times in the training set. One-word phrases are ignored if they are members of a stop list containing low-information terms such as "http", "www", "jpg","html" , etc

Note that the above procedure generates a family of encodings, one for each value of K (maximum phrase length) and M (minimum phrase count) . In the current implementation, K = 2 and M = 10. For the training data gathered as described in Sec. 2.2, the encoding consisted of 1558 features: height , width , aspect ratio, local?, 19 caption features,111 alt features,495 base URL features, 472 destination URL features, and 457 image URL features.

Now do an example;



```
http://www.provider.com/index.html

...
<A href="http://www.corp.com/sales.html">
  Our sponsor: <IMG src="http://www.corp.com/ads/thead.gif"
    alt="click here" height="40" width="200"></A>
...
<A href="contact.html">
  Contact us: <IMG src="/images/contact.gif"
    alt="contact info" height="50" width="40"></A>
...
<A href="http://www.mega.com/marketing.html">
  Funded by: <IMG src="http://www.mega.com/adverts/adimg.jpg"
    alt="free stuff"></A>
...
```

A	B	C	Feature	
40	50	?	**height**	
200	40	?	**width**	
5.0	0.8	?	**aspect ratio**	
0	0	1	**local?**	
1	0	0	"our"	**caption** features
1	0	0	"sponsor"	
1	0	0	"our+sponsor"	
0	1	0	"contact"	
0	1	0	"us"	
0	1	0	"contact+us"	
0	0	1	"funded"	
0	0	1	"by"	
0	0	1	"funded+by"	
1	0	0	"free"	**alt** features
1	0	0	"stuff"	
1	0	0	"free+stuff"	
0	1	0	"contact"	
0	1	0	"info"	
0	1	0	"contact+info"	
0	0	1	"click"	
0	0	1	"here"	
0	0	1	"click+here"	
1	1	1	"www.provider.com"	U_{base} features
1	1	1	"index"	
1	1	1	"index+html"	
1	0	0	"www.corp.com"	U_{target} features
1	0	0	"sales"	
1	0	0	"sales+html"	
0	1	0	"contact"	
0	1	0	"contact+html"	
0	0	1	"www.mega.com"	
0	0	1	"marketing"	
0	0	1	"marketing+html"	
1	0	0	"www.corp.com"	U_{img} features
1	0	0	"ads"	
1	0	0	"ads+thead"	
1	0	0	"thead"	
1	0	0	"thead+gif"	
0	1	0	"images+contact"	
0	1	0	"images"	
0	1	0	"contact"	
0	1	0	"contact+gif"	
0	0	1	"www.mega.com"	
0	0	1	"adverts"	
0	0	1	"adimg"	
0	0	1	"adverts+adimg"	
0	0	1	"adimg+jpg"	
AD	**$\overline{AD}$**	**AD**	**Classification**	

Learning Rules

C4.5rules learns a set of rules, each a conjunction of tests together with a predicted classification if the tests are satisfied. For numeric features, tests are of the form *“fi* < t” or “fi > t” , where r is a constant real number. For binary features, tests are of the form “fi’ or “fi’. For our application, C4.5rules learned a set of 25 rules. Two representative examples are as follows:

- If aspect ratio > 4.5833 alt doesn’t contain “to” but does contain “click+here”, and Udest doesn’t contain “http+www”, then instance is an AD.
- H Ubase does not contain “messier”, and Udest contains the “redirect+cgi”, then instance is an AD.

Note that these are actual rules learned by C4.5rules: the rules have only been reformatted to make them easier to read, and the learning algorithm, not a person, identifies relevant phrases such as “click+here”.

DATAMINING APPROACH AND METHODS

We have also conducted a series of more objective experiments, using the standard machine learning “cross validation” methodology. We first randomly partitioned the gathered instances into a *training* set containing 90% of the instances and a *test* set containing the remainder. We then invoked C4.5rules on the training set, and measured the performance of the rules on the test set. We cross validated our results in this way ten times.
Averaging across the ten trials, we found that the learned rules have an accuracy of 97.1%. To further understand the limitations of our approach, we have also measured the system’s learning curve. A second experiment was designed to validate the particular features in our encoding

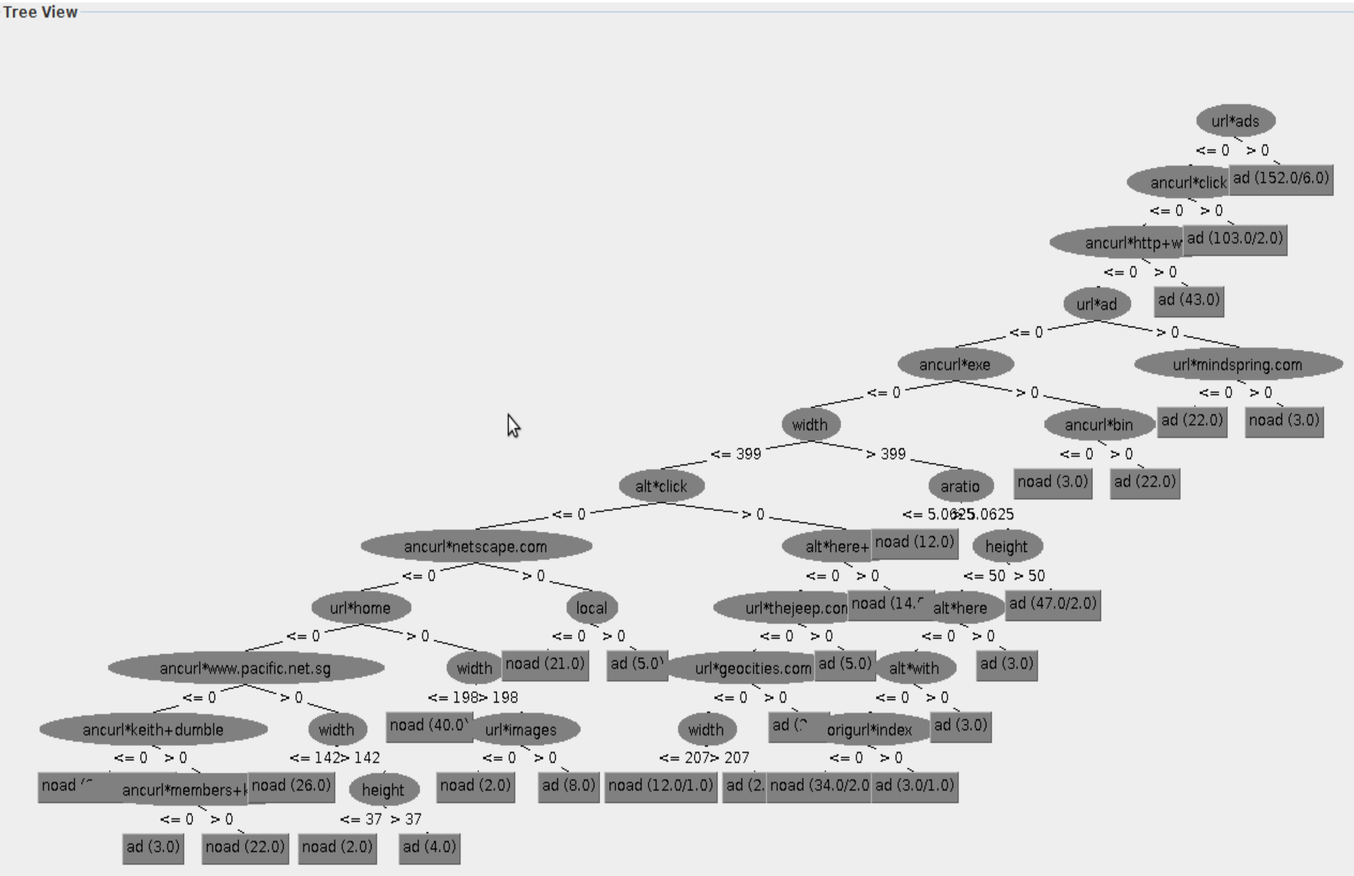
We mostly used cross valitadion 10 folds because we tested other options like use as full training set or percentage split, but the best result came out from 10 folds. Therefore we accepted it as standart and continued to test other classifiers.

J48 TREE – CROSS-VALIDATION 10 FOLD

```
=== Run information ===

Scheme:weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:     ad
Instances:    3279
Attributes:   1559
[list of attributes omitted]
```

```
Test mode:10-fold cross-validation

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

J48 pruned tree
------------------

url*ads <= 0
|   ancurl*click <= 0
|   |   ancurl*http+www <= 0
|   |   |   url*ad <= 0
|   |   |   |   ancurl*exe <= 0
|   |   |   |   |   width <= 399
|   |   |   |   |   |   alt*click <= 0
|   |   |   |   |   |   |   ancurl*netscape.com <= 0
|   |   |   |   |   |   |   |   url*home <= 0
|   |   |   |   |   |   |   |   |   ancurl*www.pacific.net.sg <= 0
|   |   |   |   |   |   |   |   |   |   ancurl*keith+dumble <= 0: noad (2661.0/40.0)
|   |   |   |   |   |   |   |   |   |   ancurl*keith+dumble > 0
|   |   |   |   |   |   |   |   |   |   |   ancurl*members+keith <= 0: ad (3.0)
|   |   |   |   |   |   |   |   |   |   |   ancurl*members+keith > 0: noad (22.0)
|   |   |   |   |   |   |   |   |   ancurl*www.pacific.net.sg > 0
|   |   |   |   |   |   |   |   |   |   width <= 142: noad (26.0)
|   |   |   |   |   |   |   |   |   |   width > 142
|   |   |   |   |   |   |   |   |   |   |   height <= 37: noad (2.0)
|   |   |   |   |   |   |   |   |   |   |   height > 37: ad (4.0)
|   |   |   |   |   |   |   |   url*home > 0
|   |   |   |   |   |   |   |   |   width <= 198: noad (40.0)
|   |   |   |   |   |   |   |   |   width > 198
|   |   |   |   |   |   |   |   |   |   url*images <= 0: noad (2.0)
|   |   |   |   |   |   |   |   |   |   url*images > 0: ad (8.0)
|   |   |   |   |   |   |   ancurl*netscape.com > 0
|   |   |   |   |   |   |   |   local <= 0: noad (21.0)
|   |   |   |   |   |   |   |   local > 0: ad (5.0)
|   |   |   |   |   |   alt*click > 0
|   |   |   |   |   |   |   alt*here+to <= 0
|   |   |   |   |   |   |   |   url*thejeep.com <= 0
|   |   |   |   |   |   |   |   |   url*geocities.com <= 0
|   |   |   |   |   |   |   |   |   |   width <= 207: noad (12.0/1.0)
|   |   |   |   |   |   |   |   |   |   width > 207: ad (2.0)
|   |   |   |   |   |   |   |   |   url*geocities.com > 0: ad (2.0)
|   |   |   |   |   |   |   |   url*thejeep.com > 0: ad (5.0)
|   |   |   |   |   |   |   alt*here+to > 0: noad (14.0)
|   |   |   |   |   width > 399
|   |   |   |   |   |   aratio <= 5.0625: noad (12.0)
|   |   |   |   |   |   aratio > 5.0625
|   |   |   |   |   |   |   height <= 50
|   |   |   |   |   |   |   |   alt*here <= 0
|   |   |   |   |   |   |   |   |   alt*with <= 0
|   |   |   |   |   |   |   |   |   |   origurl*index <= 0: noad (34.0/2.0)
|  |  |  |  |  |  |  |  |  |  |  origurl*index > 0: ad (3.0/1.0)
|  |  |  |  |  |  |  |  |  |  alt*with > 0: ad (3.0)
|  |  |  |  |  |  |  |  |  alt*here > 0: ad (3.0)
|  |  |  |  |  |  |  |  height > 50: ad (47.0/2.0)
|  |  |  |  ancurl*exe > 0
|  |  |  |  |  ancurl*bin <= 0: noad (3.0)
|  |  |  |  |  ancurl*bin > 0: ad (22.0)
|  |  |  url*ad > 0
|  |  |  |  url*mindspring.com <= 0: ad (22.0)
|  |  |  |  url*mindspring.com > 0: noad (3.0)
|  |  ancurl*http+www > 0: ad (43.0)
|  ancurl*click > 0: ad (103.0/2.0)
url*ads > 0: ad (152.0/6.0)

Number of Leaves  :   29

Size of the tree :        57


Time taken to build model: 19.85 seconds

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

Correctly Classified Instances        3184               97.1028 %
Incorrectly Classified Instances        95                2.8972 %
Kappa statistic                          0.875
Mean absolute error                      0.0469
Root mean squared error                  0.166
Relative absolute error                 19.4789 %
Root relative squared error             47.8314 %
Total Number of Instances             3279

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.852     0.01       0.935     0.852     0.892      0.916    ad
                 0.99      0.148      0.976     0.99      0.983      0.916    noad
Weighted Avg.    0.971     0.129      0.971     0.971     0.97       0.916

=== Confusion Matrix ===

    a    b   <-- classified as
  391   68 |    a = ad
   27 2793 |    b = noad
```

TREE



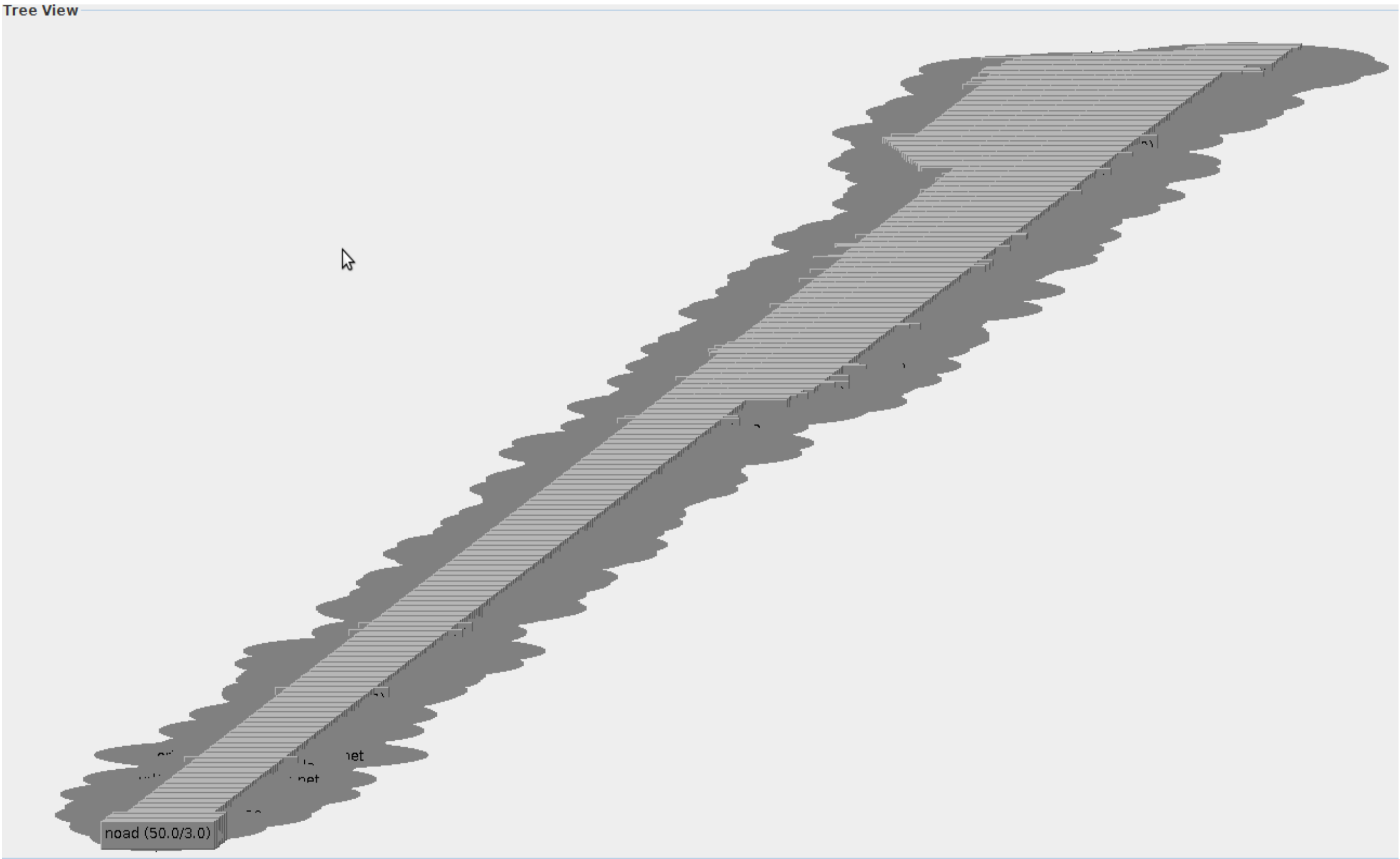
J48-Graft-10-Fold

Grafting adds nodes to the decision trees to increase the predictive accuracy. In the grafted j48 , new branches are added in the place of a single leaf or graft within leaves.

```
=== Run information ===

Scheme:weka.classifiers.trees.J48graft -C 0.25 -M 2
Relation:     ad
Instances:    3279
Attributes:   1559
[list of attributes omitted]
Test mode:10-fold cross-validation

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

Number of Leaves  :   1360

Size of the tree :        2719
```

```
Time taken to build model: 23.8 seconds

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

Correctly Classified Instances        3187               97.1943 %
Incorrectly Classified Instances        92                2.8057 %
Kappa statistic                          0.8779
Mean absolute error                      0.0461
Root mean squared error                  0.1642
Relative absolute error                 19.1335 %
Root relative squared error             47.3118 %
Total Number of Instances             3279

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.845     0.007      0.949     0.845     0.894      0.911    ad
                 0.993     0.155      0.975     0.993     0.984      0.911    noad
Weighted Avg.    0.972     0.134      0.972     0.972     0.971      0.911

=== Confusion Matrix ===

    a    b   <-- classified as
  388   71 |    a = ad
   21 2799 |    b = noad
```

TREE



NAIVE BAYES

```
=== Run information ===

Scheme:weka.classifiers.bayes.NaiveBayes
Relation:     ad
Instances:    3279
Attributes:   1559
[list of attributes omitted]
Test mode:10-fold cross-validation

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

Naive Bayes Classifier

Time taken to build model: 1.21 seconds

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

Correctly Classified Instances        3152               96.1269 %
Incorrectly Classified Instances       127                3.8731 %
```

```
Kappa statistic                    0.8277
Mean absolute error                  0.0394
Root mean squared error                0.1913
Relative absolute error             16.3682 %
Root relative squared error           55.1228 %
Total Number of Instances            3279

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.782     0.01       0.93      0.782     0.85       0.942    ad
                 0.99      0.218      0.965     0.99      0.978      0.943    noad
Weighted Avg.    0.961     0.189      0.96      0.961     0.96       0.943

=== Confusion Matrix ===

    a    b   <-- classified as
  359  100 |    a = ad
   27 2793 |    b = noad
```

META ATTRIBUTES SELECTED 10 FOLD

```
=== Run information ===

Scheme:weka.classifiers.meta.AttributeSelectedClassifier -E
"weka.attributeSelection.CfsSubsetEval " -S "weka.attributeSelection.BestFirst -D 1 -N 5" -W
weka.classifiers.trees.J48 -- -C 0.25 -M 2
Relation:     ad
Instances:    3279
Attributes:   1559
[list of attributes omitted]
Test mode:10-fold cross-validation

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

AttributeSelectedClassifier:



=== Attribute Selection on all input data ===

Search Method:
	Best first.
	Start set: no attributes
	Search direction: forward
	Stale search after 5 node expansions
	Total number of subsets evaluated: 23269
	Merit of best subset found:    0.503
```

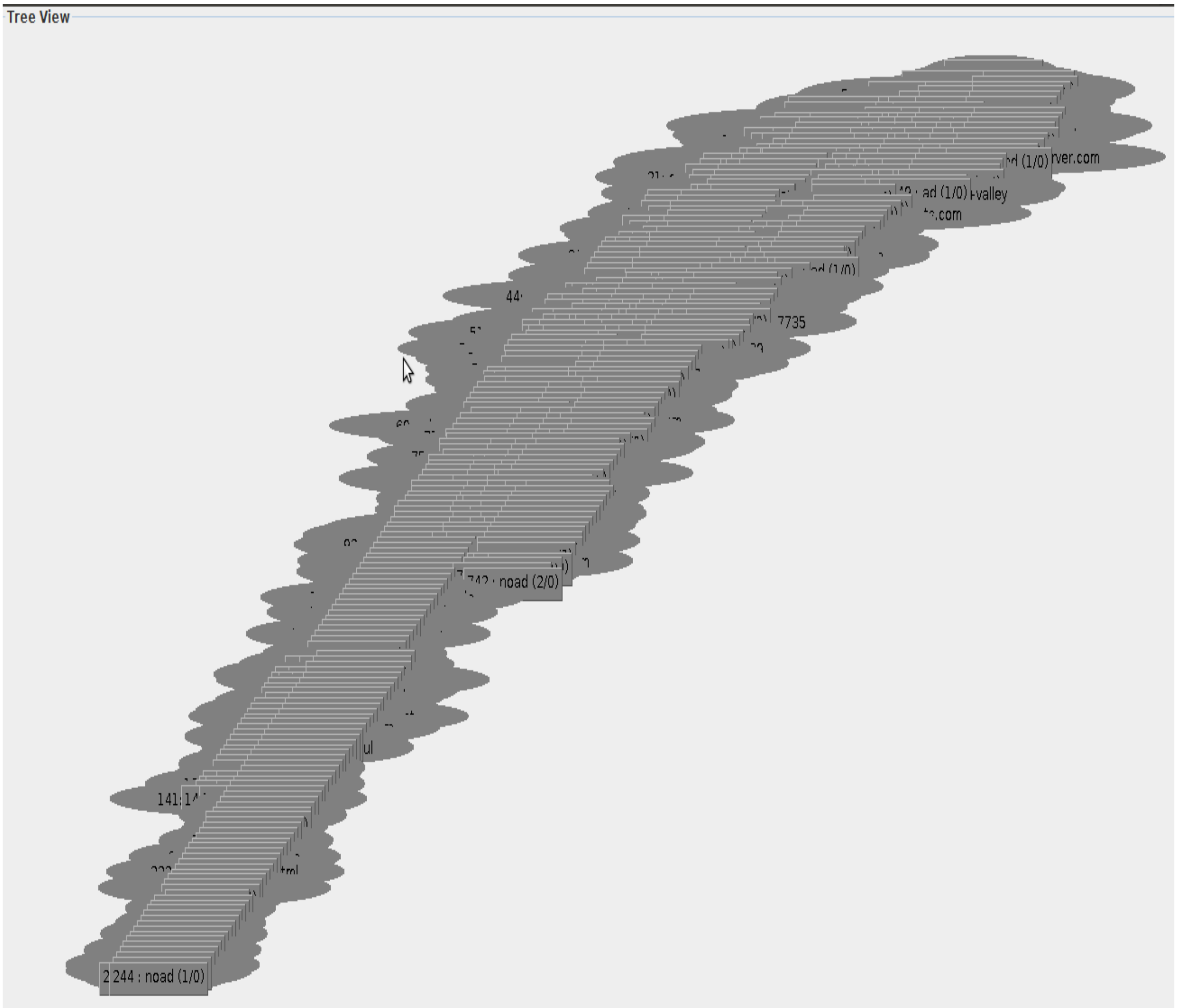
```
Attribute Subset Evaluator (supervised, Class (nominal): 1559 class):
	CFS Subset Evaluator
	Including locally predictive attributes

Selected attributes: 2,3,40,178,253,266,352,399,442,810,875,959,969,1023,1173,1230,1244,1279,1400,1460,1484,1530,1538,1547 : 24
                    width
                    aratio
                    url*pics
                    url*sjsu.edu
                    url*athens+8774
                    url*aol.com
                    url*ads
                    url*ad
                    url*icon
                    origurl*cats
                    origurl*bin
                    ancurl*mirror
                    ancurl*redirect
                    ancurl*adclick
                    ancurl*links
                    ancurl*http+www
                    ancurl*com
                    ancurl*bin
                    ancurl*click
                    alt*us
                    alt*click
                    alt*award
                    alt*home
                    caption*page


Header of reduced data:
@relation 'ad-weka.filters.unsupervised.attribute.Remove-V-R2-3,40,178,253,266,352,399,442,810,875,959,969,1023,1173,1230,1244,1279,1400,1460,1484,1530,1538,1547,1559'

@attribute width numeric
@attribute aratio numeric
@attribute url*pics numeric
@attribute url*sjsu.edu numeric
@attribute url*athens+8774 numeric
@attribute url*aol.com numeric
@attribute url*ads numeric
@attribute url*ad numeric
@attribute url*icon numeric
@attribute origurl*cats numeric
@attribute origurl*bin numeric
@attribute ancurl*mirror numeric
```

```
@attribute ancurl*redirect numeric
@attribute ancurl*adclick numeric
@attribute ancurl*links numeric
@attribute ancurl*http+www numeric
@attribute ancurl*com numeric
@attribute ancurl*bin numeric
@attribute ancurl*click numeric
@attribute alt*us numeric
@attribute alt*click numeric
@attribute alt*award numeric
@attribute alt*home numeric
@attribute caption*page numeric
@attribute class {ad,noad}

@data


Classifier Model
J48 pruned tree
------------------

url*ads <= 0
|   ancurl*click <= 0
|   |   ancurl*http+www <= 0
|   |   |   url*ad <= 0
|   |   |   |   width <= 399
|   |   |   |   |   alt*click <= 0
|   |   |   |   |   |   ancurl*bin <= 0
|   |   |   |   |   |   |   width <= 146: noad (2121.0/30.0)
|   |   |   |   |   |   |   width > 146
|   |   |   |   |   |   |   |   aratio <= 2.775: noad (277.0/2.0)
|   |   |   |   |   |   |   |   aratio > 2.775
|   |   |   |   |   |   |   |   |   aratio <= 3
|   |   |   |   |   |   |   |   |   |   width <= 200: ad (5.0)
|   |   |   |   |   |   |   |   |   |   width > 200: noad (3.0/1.0)
|   |   |   |   |   |   |   |   |   aratio > 3
|   |   |   |   |   |   |   |   |   |   aratio <= 6.4814: noad (103.0/1.0)
|   |   |   |   |   |   |   |   |   |   aratio > 6.4814
|   |   |   |   |   |   |   |   |   |   |   aratio <= 6.6666: ad (7.0)
|   |   |   |   |   |   |   |   |   |   |   aratio > 6.6666: noad (55.0/4.0)
|   |   |   |   |   |   ancurl*bin > 0
|   |   |   |   |   |   |   ancurl*com <= 0: noad (216.0/9.0)
|   |   |   |   |   |   |   ancurl*com > 0
|   |   |   |   |   |   |   |   width <= 37: noad (9.0)
|   |   |   |   |   |   |   |   width > 37: ad (6.0)
|   |   |   |   |   alt*click > 0
|   |   |   |   |   |   ancurl*bin <= 0: noad (34.0/9.0)
|   |   |   |   |   |   ancurl*bin > 0: ad (3.0)
|   |   |   |   width > 399
|   |   |   |   |   aratio <= 18
|  |  |  |  |  |  aratio <= 5.0625: noad (12.0)
|  |  |  |  |  |  aratio > 5.0625: ad (81.0/11.0)
|  |  |  |  |  aratio > 18: noad (24.0)
|  |  |  url*ad > 0
|  |  |  |  aratio <= 0.4058: noad (3.0)
|  |  |  |  aratio > 0.4058: ad (22.0)
|  |  ancurl*http+www > 0: ad (43.0)
|  ancurl*click > 0: ad (103.0/2.0)
url*ads > 0: ad (152.0/6.0)

Number of Leaves  :   20

Size of the tree :        39


Time taken to build model: 7.26 seconds

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

Correctly Classified Instances        3177               96.8893 %
Incorrectly Classified Instances       102                3.1107 %
Kappa statistic                          0.8628
Mean absolute error                      0.0533
Root mean squared error                  0.1702
Relative absolute error                 22.1106 %
Root relative squared error             49.0671 %
Total Number of Instances             3279

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.819     0.007      0.952     0.819     0.881      0.918    ad
                 0.993     0.181      0.971     0.993     0.982      0.918    noad
Weighted Avg.    0.969     0.156      0.969     0.969     0.968      0.918

=== Confusion Matrix ===

    a    b   <-- classified as
  376   83 |    a = ad
   19 2801 |    b = noad
```

RANDOM TREE CLASSIFIER

```
=== Run information ===

Scheme:weka.classifiers.trees.RandomTree -K 0 -M 1.0 -S 1
Relation:     ad
Instances:    3279
Attributes:   1559
[list of attributes omitted]
Test mode:10-fold cross-validation

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


RandomTree
==========

Size of the tree : 1145

Time taken to build model: 1.64 seconds

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

Correctly Classified Instances        3170               96.6758 %
Incorrectly Classified Instances       109                3.3242 %
Kappa statistic                          0.8605
Mean absolute error                      0.0337
Root mean squared error                  0.1818
Relative absolute error                 13.9744 %
Root relative squared error             52.3911 %
Total Number of Instances             3279

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.869     0.017      0.891     0.869     0.88       0.926    ad
                 0.983     0.131      0.979     0.983     0.981      0.926    noad
Weighted Avg.    0.967     0.115      0.966     0.967     0.967      0.926

=== Confusion Matrix ===

    a    b   <-- classified as
  399   60 |    a = ad
   49 2771 |    b = noad
```

TREE



Tree View
rver.com
ad (1/0)
valley
.com
7735
742 : noad (2/0)
141
2 244 : noad (1/0)

IBK instance-based classifier k=1 used as training set

```
=== Run information ===

Scheme:weka.classifiers.lazy.IBk -K 1 -W 0 -A "weka.core.neighboursearch.LinearNNSearch
-A \"weka.core.EuclideanDistance -R first-last\""
Relation:     ad
Instances:    3279
Attributes:   1559
[list of attributes omitted]
Test mode:evaluate on training data

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

IB1 instance-based classifier
using 1 nearest neighbour(s) for classification


Time taken to build model: 0.03 seconds

=== Evaluation on training set ===
=== Summary ===

Correctly Classified Instances        3276               99.9085 %
Incorrectly Classified Instances         3                0.0915 %
Kappa statistic                          0.9962
Mean absolute error                      0.0015
Root mean squared error                  0.0256
Relative absolute error                  0.6366 %
Root relative squared error              7.374  %
Total Number of Instances             3279

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.998     0.001     0.996      0.998     0.997          1        ad
                 0.999     0.002     1          0.999     0.999          1        noad
Weighted Avg.    0.999     0.002     0.999      0.999     0.999          1

=== Confusion Matrix ===

    a    b   <-- classified as
  458    1 |    a = ad
    2 2818 |    b = noad
```

IBK instance-based classifier K=1, 10-fold cross-validation

```
=== Run information ===

Scheme:weka.classifiers.lazy.IBk -K 1 -W 0 -A "weka.core.neighboursearch.LinearNNSearch
-A \"weka.core.EuclideanDistance -R first-last\""
Relation:     ad
Instances:    3279
Attributes:   1559
[list of attributes omitted]
Test mode:10-fold cross-validation

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

IB1 instance-based classifier
using 1 nearest neighbour(s) for classification


Time taken to build model: 0.03 seconds

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

Correctly Classified Instances        3164               96.4928 %
Incorrectly Classified Instances       115                3.5072 %
Kappa statistic                          0.8493
Mean absolute error                      0.0358
Root mean squared error                  0.186
Relative absolute error                 14.8386 %
Root relative squared error             53.6157 %
Total Number of Instances             3279

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.834     0.014      0.908     0.834     0.869      0.939    ad
                 0.986     0.166      0.973     0.986     0.98       0.939    noad
Weighted Avg.    0.965     0.144      0.964     0.965     0.964      0.939

=== Confusion Matrix ===

    a    b   <-- classified as
  383   76 |    a = ad
   39 2781 |    b = noad
```

J48-full training set

```
=== Run information ===
Scheme:weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:     ad
Instances:    3279
Attributes:   1559

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

J48 pruned tree
url*ads <= 0
|   ancurl*click <= 0
|   |   ancurl*http+www <= 0
|   |   |   url*ad <= 0
|   |   |   |   ancurl*exe <= 0
|   |   |   |   |   width <= 399
|   |   |   |   |   |   alt*click <= 0
|   |   |   |   |   |   |   ancurl*netscape.com <= 0
|   |   |   |   |   |   |   |   url*home <= 0
|   |   |   |   |   |   |   |   |   ancurl*www.pacific.net.sg <= 0
|   |   |   |   |   |   |   |   |   |   ancurl*keith+dumble <= 0: noad (2661.0/40.0)
|   |   |   |   |   |   |   |   |   |   ancurl*keith+dumble > 0
|   |   |   |   |   |   |   |   |   |   |   ancurl*members+keith <= 0: ad (3.0)
|   |   |   |   |   |   |   |   |   |   |   ancurl*members+keith > 0: noad (22.0)
|   |   |   |   |   |   |   |   |   ancurl*www.pacific.net.sg > 0
|   |   |   |   |   |   |   |   |   |   width <= 142: noad (26.0)
|   |   |   |   |   |   |   |   |   |   width > 142
|   |   |   |   |   |   |   |   |   |   |   height <= 37: noad (2.0)
|   |   |   |   |   |   |   |   |   |   |   height > 37: ad (4.0)
|   |   |   |   |   |   |   |   url*home > 0
|   |   |   |   |   |   |   |   |   width <= 198: noad (40.0)
|   |   |   |   |   |   |   |   |   width > 198
|   |   |   |   |   |   |   |   |   |   url*images <= 0: noad (2.0)
|   |   |   |   |   |   |   |   |   |   url*images > 0: ad (8.0)
|   |   |   |   |   |   |   ancurl*netscape.com > 0
|   |   |   |   |   |   |   |   local <= 0: noad (21.0)
|   |   |   |   |   |   |   |   local > 0: ad (5.0)
|   |   |   |   |   |   alt*click > 0
|   |   |   |   |   |   |   alt*here+to <= 0
|   |   |   |   |   |   |   |   url*thejeep.com <= 0
|   |   |   |   |   |   |   |   |   url*geocities.com <= 0
|   |   |   |   |   |   |   |   |   |   width <= 207: noad (12.0/1.0)
|   |   |   |   |   |   |   |   |   |   width > 207: ad (2.0)
|   |   |   |   |   |   |   |   |   url*geocities.com > 0: ad (2.0)
|   |   |   |   |   |   |   |   url*thejeep.com > 0: ad (5.0)
|   |   |   |   |   |   |   alt*here+to > 0: noad (14.0)
|   |   |   |   |   width > 399
|   |   |   |   |   |   aratio <= 5.0625: noad (12.0)
|   |   |   |   |   |   aratio > 5.0625
|   |   |   |   |   |   |   height <= 50
|   |   |   |   |   |   |   |   |   alt*here <= 0
|   |   |   |   |   |   |   |   |   |   alt*with <= 0
|   |   |   |   |   |   |   |   |   |   |   origurl*index <= 0: noad (34.0/2.0)
|   |   |   |   |   |   |   |   |   |   |   origurl*index > 0: ad (3.0/1.0)
|   |   |   |   |   |   |   |   |   |   alt*with > 0: ad (3.0)
|   |   |   |   |   |   |   |   |   alt*here > 0: ad (3.0)
|   |   |   |   |   |   |   |   height > 50: ad (47.0/2.0)
|   |   |   |   ancurl*exe > 0
|   |   |   |   |   ancurl*bin <= 0: noad (3.0)
|   |   |   |   |   ancurl*bin > 0: ad (22.0)
|   |   |   url*ad > 0
|   |   |   |   url*mindspring.com <= 0: ad (22.0)
|   |   |   |   url*mindspring.com > 0: noad (3.0)
|   |   ancurl*http+www > 0: ad (43.0)
|   ancurl*click > 0: ad (103.0/2.0)
url*ads > 0: ad (152.0/6.0)

Number of Leaves  :   29

Size of the tree :        57


Time taken to build model: 31.28 seconds

=== Evaluation on training set ===
=== Summary ===

Correctly Classified Instances        3225               98.3532 %
Incorrectly Classified Instances        54                1.6468 %
Kappa statistic                          0.9295
Mean absolute error                      0.032
Root mean squared error                  0.1265
Relative absolute error                 13.2918 %
Root relative squared error             36.4699 %
Total Number of Instances             3279

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.906     0.004      0.974     0.906     0.939      0.957    ad
                 0.996     0.094      0.985     0.996     0.99       0.957    noad
Weighted Avg.    0.984     0.081      0.983     0.984     0.983      0.957


=== Confusion Matrix ===

    a    b   <-- classified as
  416   43 |    a = ad
   11 2809 |    b = noad
```

CONCLUSION

	J48	J48-Grafted	Naive Bayes	Random Tree	Attributes Selected
Correctly Classified Instances	**97.1028 %**	**97.1943 %**	**96.1269 %**	**96.6758 %**	**96.8893 %**
Incorrectly Classified Instances	**2.8972 %**	**2.8057 %**	**3.8731 %**	**3.3242 %**	**3.1107 %**
TP/ad	**0.852**	**0.845**	**0.782**	**0.869**	**0.819**
TP/noad	**0.99**	**0.993**	**0.99**	**0.983**	**0.993**
FP/ad	**0.01**	**0.007**	**0.01**	**0.017**	**0.007**
FP/noad	**0.148**	**0.155**	**0.218**	**0.131**	**0.181**
Precision/ad	**0.935**	**0.949**	**0.93**	**0.891**	**0.952**
Precision/ noad	**0.976**	**0.975**	**0.965**	**0.979**	**0.971**
Time	**19.85 seconds**	**23.8 seconds**	**1.21 seconds**	**1.64 seconds**	**7.26 seconds**
Number of Leaves	**29**	**1360**	-	-	-
Size of the tree	**57**	**2719**	-	1145	-

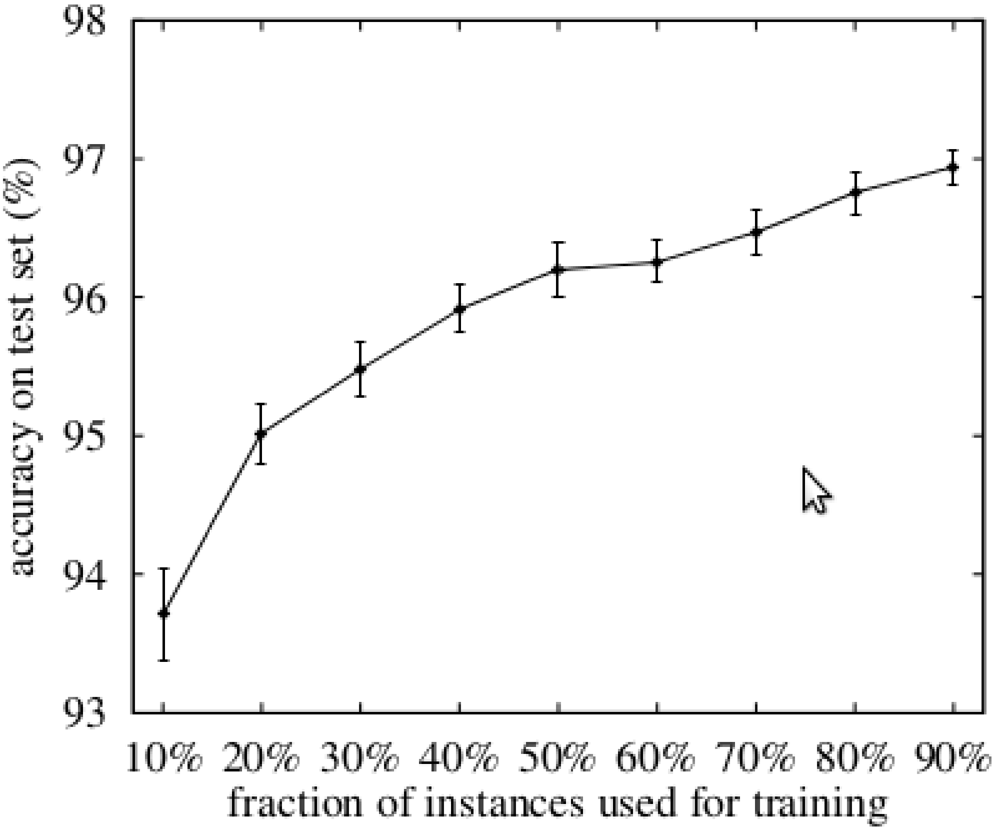
	J48	IBK k=1, 10 fold	J48-traning set	IBK k=1, use as training set
Correctly Classified Instances	**97.1028 %**	**97.1943 %**	**98.3532 %**	**99.9085 %**
Incorrectly Classified Instances	**2.8972 %**	**2.8057 %**	**1.6468 %**	**0.0915 %**
TP/ad	**0.852**	**0.845**	**0.906**	**0.998**
TP/noad	**0.99**	**0.993**	**0.996**	**0.999**
FP/ad	**0.01**	**0.007**	**0.004**	**0.001**
FP/noad	**0.148**	**0.155**	**0.094**	**0.002**
Precision/ad	**0.935**	**0.949**	**0.974**	**0.996**
Precision/ noad	**0.976**	**0.975**	**0.985**	**1**

Nicholas Kushmerick's has tested his dataset with J48 and now his comments:
“Our experiments demonstrate that our approach is practical: the off-line training phase takes less than six minutes; on-line classification takes about 70 msec; and classification accuracy exceeds 97% given a modest set of training data.”

“To calculate a learning curve for our system, we gave the learning algorithm 10%, 20%, . . . , 90% of the training data, and then calculated 10-fold cross-validated accuracy on the remainder. Fig. 4(a) shows the results, along with 95% confidence intervals after ten repetitions of this process. The observed accuracy asymptotically approaches the 97.1% figure reported earlier, and exceeds 93% with just 10% of the training data.”

“We have also conducted a series of more objective experiments, using the standard machine learning “cross validation” methodology. We first randomly partitioned the gathered instances into a *training* set containing 90% of the instances and a *test* set containing the remainder. We then invoked C4.5rules on the training set, and measured the performance of the rules on the test set. We cross validated our results in this way ten times.
Averaging across the ten trials, we found that the learned rules have an accuracy of 97.1%. “

Actually we think Kushmerick's thoughts are right because we have tested other classifiers and still J48 classifier was the best. Cross validation works great with this dataset. Best k-fold is 10. You can see the k-fold chart here.



(a)

REFERENCES

1- Learning to remove Internet advertisements-Nicholas Kushmerick Department of Computer Science, University College Dublin, Dublin 4, Ireland
http://www.sc.ehu.es/ccwbayes/docencia/mmcc/docs/lecturas-clasificacion/abstracts-resumir/kushmerick99learning.pdf

2- Internet Advertisements Data Set
https://archive.ics.uci.edu/ml/datasets/Internet+Advertisements