Assignment 2: Mushroom Scenario

CIS 435

Section 56

Summer Quarter

School of Continuing Studies

Northwestern University

Daniel Prusinski

Business Intelligence Data Analyst

Target Corporation

Minneapolis, MN

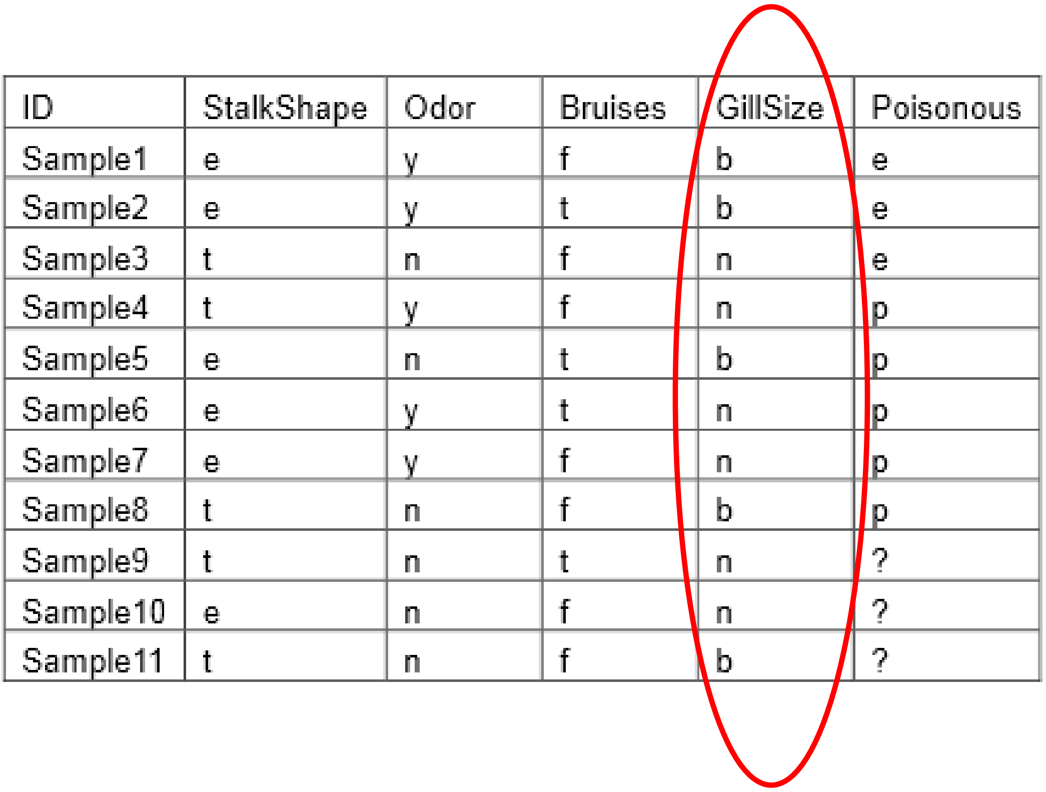
In Compliance with Master of Science Predictive Analytics

A.

1

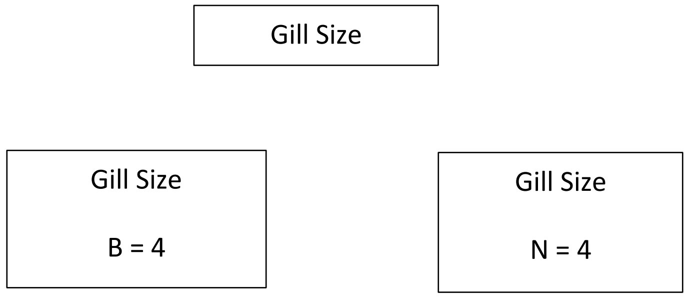
	Entropy of Poisonus		Entropy of Edible	
	(-5/8 * Log5/8) + (-3/8 * Log3/8)		(-3/8 * Log3/8) + (-5/8 * Log5/8)	
-0.625		0.625	0.375	-0.375
Log			Log	
-0.678072	Log		- 1.4150370	
0.423795	Entropy of Poisonous		0.5306389	
0.375		-0.375	0.625	-0.625
Log			Log	
- 1.4150370			-0.678072	
0.5306389	Entropy of Not Poisonous		0.423795	
0.9544339	Both Combined		0.9544339	

2. Gill Size, this can be seen visually from just looking at the data. This attribute gives you the most information gain. Info Gain is .04.



ID	StalkShape	Odor	Bruises	GillSize	Poisonous
Sample1	e	y	f	b	e
Sample2	e	y	t	b	e
Sample3	t	n	f	n	e
Sample4	t	y	f	n	p
Sample5	e	n	t	b	p
Sample6	e	y	t	n	p
Sample7	e	y	f	n	p
Sample8	t	n	f	b	p
Sample9	t	n	t	n	?
Sample10	e	n	f	n	?
Sample11	t	n	f	b	?

3. Build a decision tree using entropy calculation. Make predictions on the last three samples.
I am trying to predict the Poisonous, thus I cannot use that attribute in the decision tree.
I know that Gill Size is the best first split, therefore I will start with that.



Gill Size

Gill Size

B = 4

Gill Size

N = 4

	B	N
P	2	3
E	2	1

Entropy of B
(-4/8 * Log4/8)
-0.5 0.5
Log
-1
0.5 Entropy, this is a terrible value

Given B, the entropy for Poison is
(2/4 * Log2/4)
-0.5 0.5
Log
-1
0.5 Entropy, this is a terrible value
So, you basically have a 50/50 chance

Given N, the entropy for Poison is
(-3/4 * Log3/4)
-0.75 0.75
Log
-0.415
0.31125 Entropy, this shows that N is OK as an indicator of poison.

Given N, the entropy for Edible is
(-1/4 * Log1/4)
-0.25 0.25

Log

-2

0.5 This shows that N is not a good indicator for non-poison.

ODOR – This follows the same math as above:

	Y	N
E	2	1
P	3	2

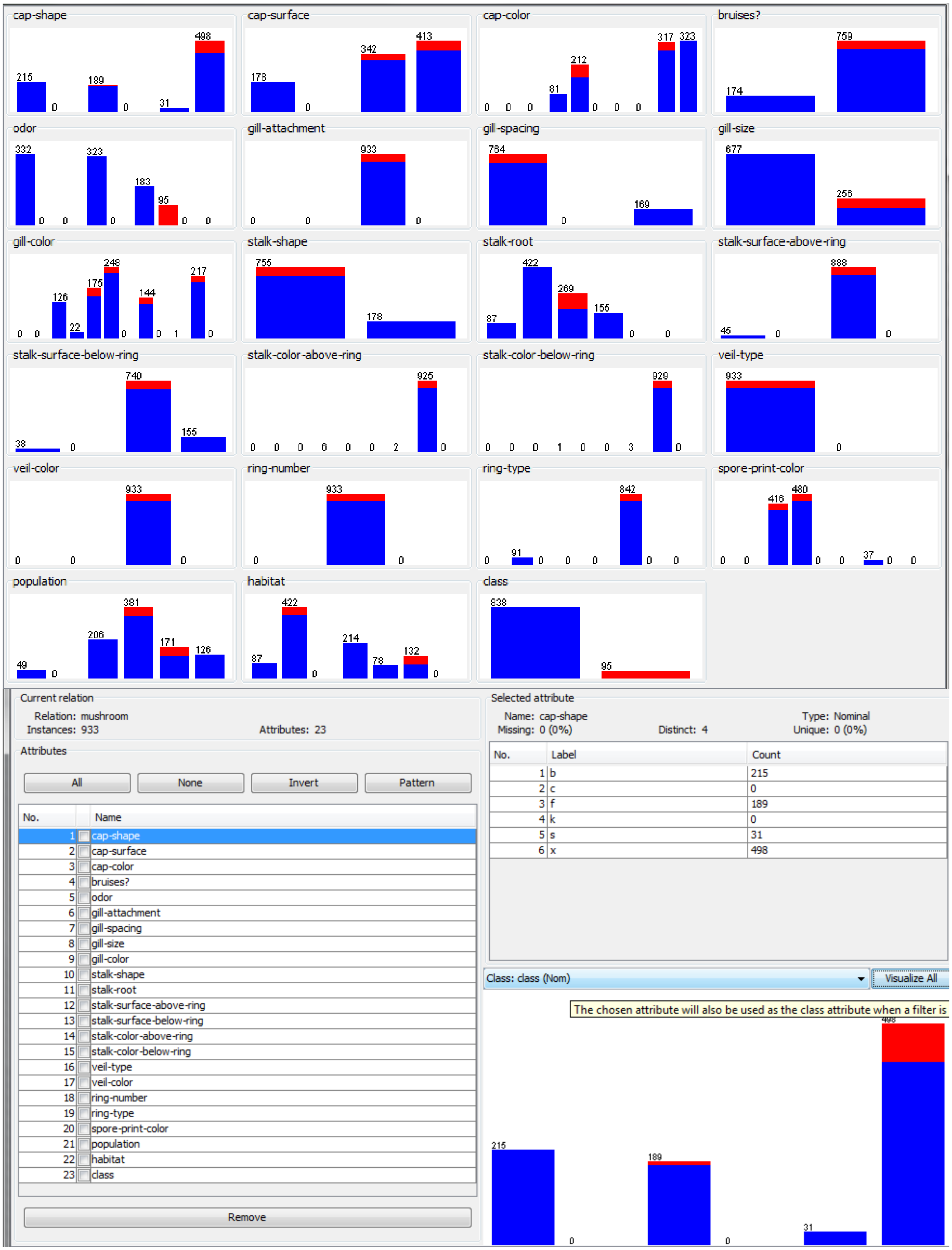
Odor Y = Edible =.52

= Poisonous =.44

Odor N = Edible =.52

=Poisonous = .38

Given these calculations, I would predict Sample 9 & 10 are edible. Sample 11 has a 50/50 shot, I would simply not eat it, I would air on the side of this being poisonous.



Current relation

Relation: mushroom
Instances: 933 Attributes: 23

Selected attribute

Name: cap-shape Type: Nominal
Missing: 0 (0%) Distinct: 4 Unique: 0 (0%)

No.	Label	Count
1	b	215
2	c	0
3	f	189
4	k	0
5	s	31
6	x	498

Attributes

All None Invert Pattern

No.	Name
1	□ cap-shape
2	□ cap-surface
3	□ cap-color
4	□ bruises?
5	□ odor
6	□ gill-attachment
7	□ gill-spacing
8	□ gill-size
9	□ gill-color
10	□ stalk-shape
11	□ stalk-root
12	□ stalk-surface-above-ring
13	□ stalk-surface-below-ring
14	□ stalk-color-above-ring
15	□ stalk-color-below-ring
16	□ veil-type
17	□ veil-color
18	□ ring-number
19	□ ring-type
20	□ spore-print-color
21	□ population
22	□ habitat
23	□ class

Remove

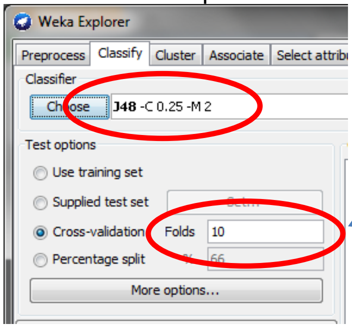
Class: class (Nom) Visualize All

The chosen attribute will also be used as the class attribute when a filter is

Preliminary Analysis: In this data set there are 933 examples represented in 23 attributes. Some attributes are binary, like bruises, gill-size, veil - type, stalk shape, and class, ye other attributes are categorical or continuous. I assume that the class attribute is the identifier of poisonous and non-poisonous. Visually, it is easy to see that the best first branch is gill size based on the vast number of instances that will be eliminated, 677 to be exact.

B.Weka exercise

Using mushroom Small.arff file as the input file, construct a decision tree using J48 algorithm.
Take the default parameter setting, and use cross-validation with the 10 folds option.



Classifier
J48 -C 0.25 -M 2

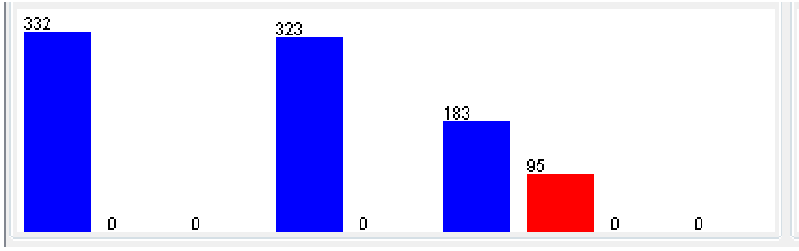
Test options
○ Use training set
○ Supplied test set
● Cross-validation Folds 10
○ Percentage split % 66

What is the result tree? What do you think about this tree? Are you satisfied with the result? Why or why not?

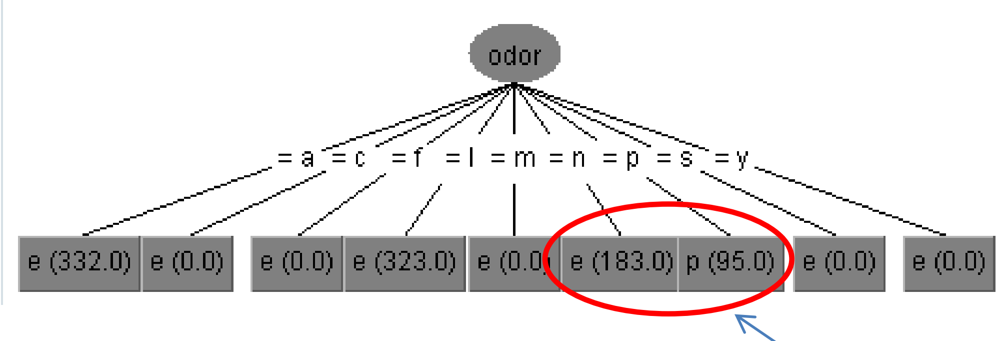
```
Scheme:weka.classifiers.trees.J48 -C 0.25 -M 2
Test mode:10-fold cross-validation
=== Classifier model (full training set) ===

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

odor = a: e (332.0)
odor = c: e (0.0)
odor = f: e (0.0)
odor = l: e (323.0)
odor = m: e (0.0)
odor = n: e (183.0)
odor = p: p (95.0)
odor = s: e (0.0)
odor = y: e (0.0)

Number of Leaves  :      9

Size of the tree :       10

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

Correctly Classified Instances        933             100      %
Incorrectly Classified Instances        0               0      %
Kappa statistic                         1
Mean absolute error                     0
Root mean squared error                   0
Relative absolute error                 0      %
Root relative squared error               0      %
Total Number of Instances             933
```

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

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 1         0          1          1         1          1        e
                 1         0          1          1         1          1        p
Weighted Avg.    1         0          1          1         1          1

=== Confusion Matrix ===

   a   b   <-- classified as
 838   0 |   a = e
   0  95 |   b = p
```

J48 used Odor as the first branch. At first glance, this looks like the best attributes to classify the instances.



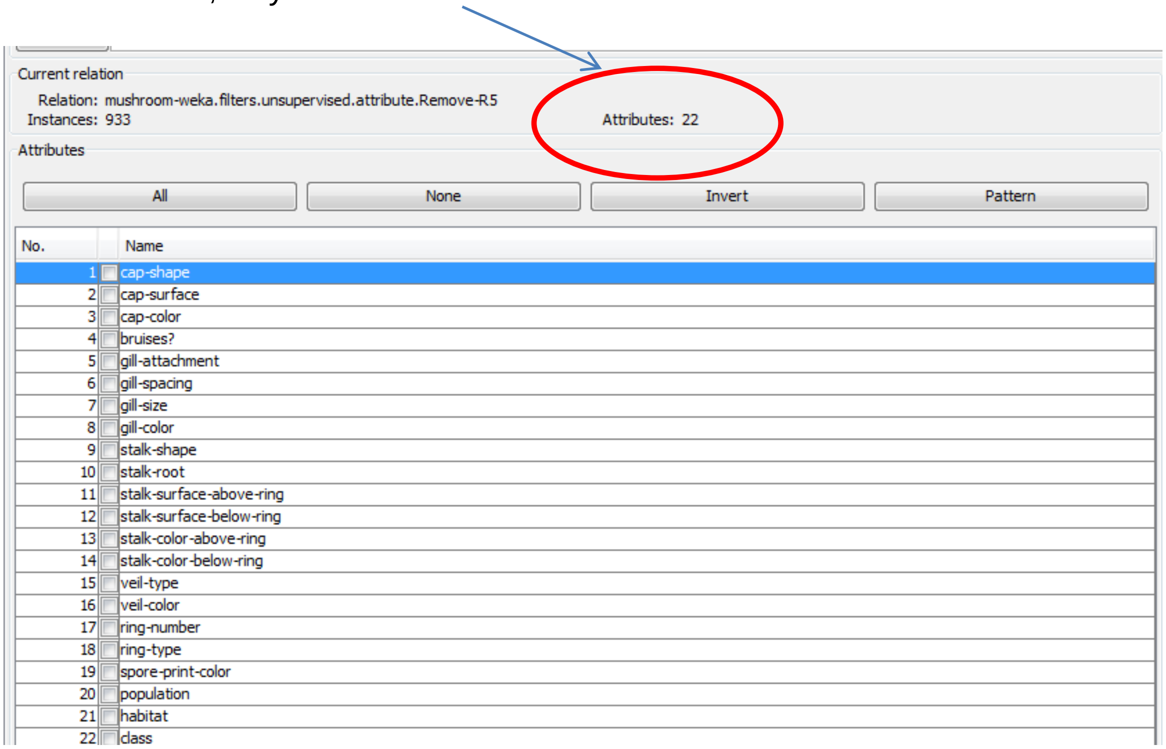
The decision tree below was generated from J48 with 10 folds. In my opinion, odor is very subjective and the margin for interpretation is too large to warrant using this attribute, especially when risking one's life is on the line.



Preliminarily, this tree looks good, but it appears to be over fit. Every possible category seems to have a node, which is why I am apprehensive to how this would perform on non-training data. For someone to distinguish between 9 different smells is unrealistic, especially when odor and n and p are so close to one another.

2. Remove the “odor” attribute and reconstruct the tree.

Odor removed, only 22 attributes.



Current relation

Relation: mushroom-weka.filters.unsupervised.attribute.Remove-R5

Instances: 933

Attributes: 22

Attributes

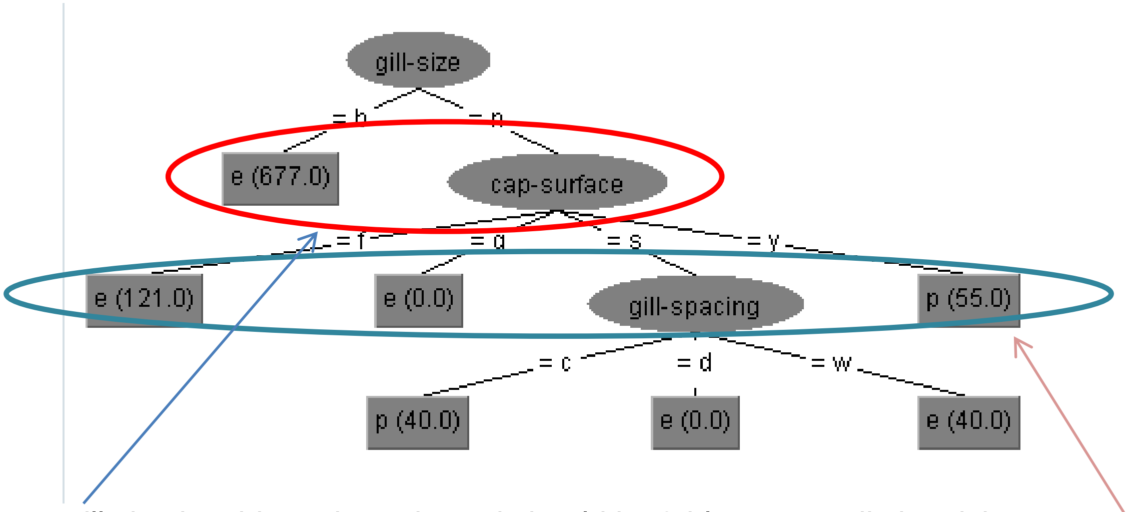
No.		Name
1	□	cap-shape
2	□	cap-surface
3	□	cap-color
4	□	bruises?
5	□	gill-attachment
6	□	gill-spacing
7	□	gill-size
8	□	gill-color
9	□	stalk-shape
10	□	stalk-root
11	□	stalk-surface-above-ring
12	□	stalk-surface-below-ring
13	□	stalk-color-above-ring
14	□	stalk-color-below-ring
15	□	veil-type
16	□	veil-color
17	□	ring-number
18	□	ring-type
19	□	spore-print-color
20	□	population
21	□	habitat
22	□	class

What is this second tree?

```
Scheme:weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:     mushroom-weka.filters.unsupervised.attribute.Remove-R5
Instances:    933
Attributes:   22
              cap-shape
              cap-surface
              cap-color
              bruises?
              gill-attachment
              gill-spacing
              gill-size
              gill-color
              stalk-shape
              stalk-root
              stalk-surface-above-ring
              stalk-surface-below-ring
              stalk-color-above-ring
              stalk-color-below-ring
              veil-type
              veil-color
              ring-number
              ring-type
              spore-print-color
              population
              habitat
```

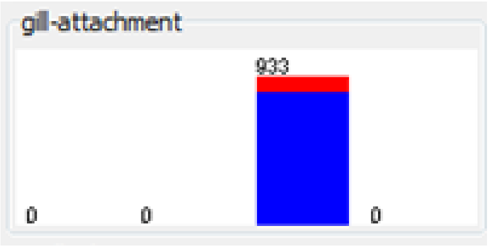
```
          class
Test mode:10-fold cross-validation

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

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

gill-size = b: e (677.0)
gill-size = n
|   cap-surface = f: e (121.0)
|   cap-surface = g: e (0.0)
|   cap-surface = s
|   |   gill-spacing = c: p (40.0)
|   |   gill-spacing = d: e (0.0)
|   |   gill-spacing = w: e (40.0)
|   cap-surface = y: p (55.0)

Number of Leaves  :      7

Size of the tree :           10


Time taken to build model: 0 seconds

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

Correctly Classified Instances         933              100      %
Incorrectly Classified Instances         0                0      %
Kappa statistic                          1
Mean absolute error                          0
Root mean squared error                          0
Relative absolute error                      0      %
Root relative squared error                    0      %
Total Number of Instances              933

=== Detailed Accuracy By Class ===

            TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
              1         0         1          1        1          1        e
              1         0         1          1        1          1        p
Weighted Avg.    1         0         1          1        1          1

=== Confusion Matrix ===

   a   b   <-- classified as
 838   0 |   a = e
   0  95 |   b = p
```

This tree is based off of the “Gill-size, Cap-surface, and then Gill-spacing attributes, and has 7 leaves.



Gill size is a binary branch, and should be fairly easy to distinguish.
Cap – Surface has edible on one side of the spectrum and poisonous on the other, which should be relatively easy to distinguish.
Gill-Spacing is a 50/50 chance, but the spectrum is wide spread, so one could reason as to what the edible gill size is which would strengthen the odds.

3. What do you think about a gill-attachment attribute? Would it impact the tree building if you remove this variable?



The gill attachment variable is pointless because all the instances fall within the same category. If this variable was removed, there would be no negative impact to the tree.

4. Keep: gill-attachment, gill-color, stalk-color-above-ring, stalk-color-below-ring, veil-type, veil-color, and ring-number. Now,construct the tree. What tree have you obtained?
Preliminary analysis:

Gill Attachment – This adds no value, see notes above.

Gill Cover – There are only 148 instances that can be removed with this attribute, and it too does not look like solid variable.

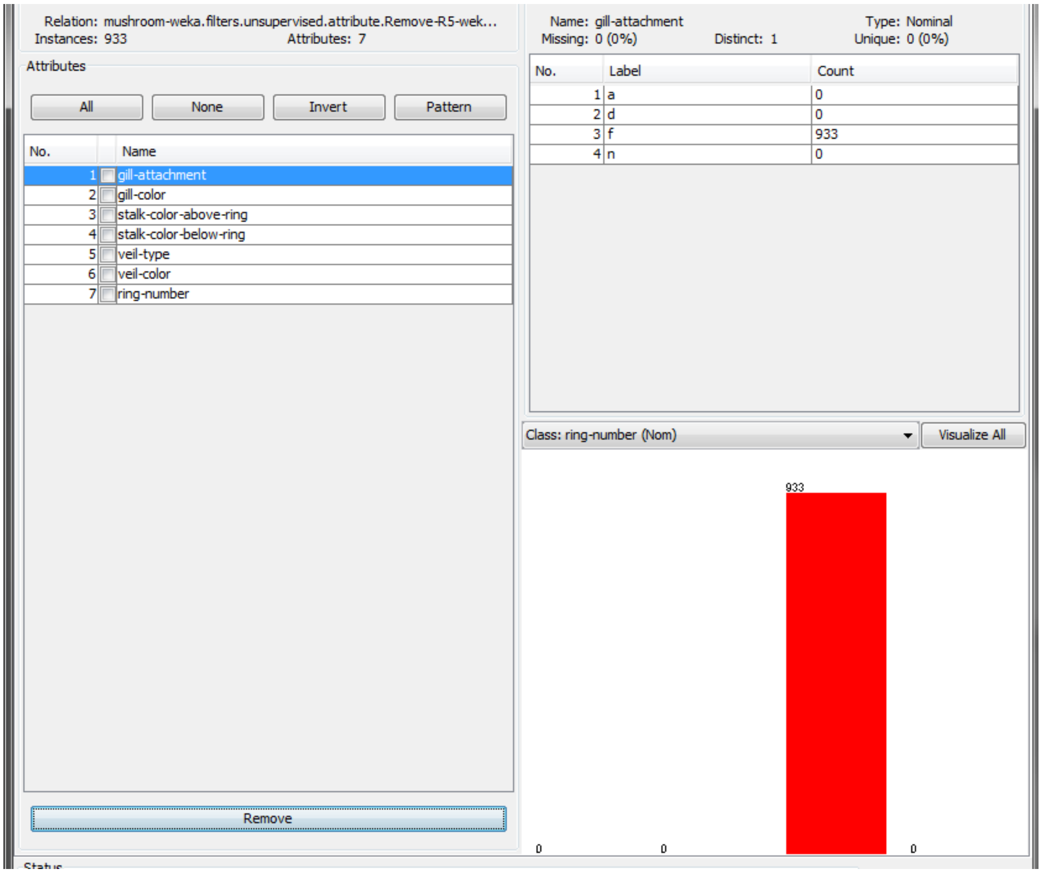
Stalk-Color-Above-Ring – Only 2 instances are in one category and 925 are in the other. This variable is not a good classifier either.

Stalk-Color-Below-Ring – This variable has 929 instances that are within one category, and this is a poor variable.

Veil type, Ring-Number, and Veil Cover are also complete variables where all the instances are in one category.

From the variables listed above, I suspect that the decision tree from these variables would be very poor based on how poorly they delineate poisonous mushrooms from edible.

Initial step of grouping all the variables:



Relation: mushroom-weka.filters.unsupervised.attribute.Remove-R5-wek...
Instances: 933 Attributes: 7

No.	Name
1	gill-attachment
2	gill-color
3	stalk-color-above-ring
4	stalk-color-below-ring
5	veil-type
6	veil-color
7	ring-number

Name: gill-attachment Type: Nominal
Missing: 0 (0%) Distinct: 1 Unique: 0 (0%)

No.	Label	Count
1	a	0
2	d	0
3	f	933
4	n	0

Class: ring-number (Nom)

Result:

=== Run information ===

Scheme:weka.classifiers.trees.J48 -C 0.25 -M 2
Relation: mushroom-weka.filters.unsupervised.attribute.Remove-R5-weka.filters.unsupervised.attribute.Remove-R1-4,6-7,9-12,18-22
Instances: 933
Attributes: 7

    gill-attachment
    gill-color
    stalk-color-above-ring
    stalk-color-below-ring
    veil-type
    veil-color

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

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

J48 pruned tree
------------------
: o (933.0)

Number of Leaves  :       1

Size of the tree :        1


Time taken to build model: 0 seconds

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

Correctly Classified Instances         933              100      %
Incorrectly Classified Instances         0                0      %
Kappa statistic                          1
Mean absolute error                      0
Root mean squared error                  0
Relative absolute error                  0      %
Root relative squared error              0      %
Total Number of Instances              933

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0         0         0         0         0          ?        n
                 1         0         1         1         1          ?        o
                 0         0         0         0         0          ?        t
Weighted Avg.    1         0         1         1         1          0

=== Confusion Matrix ===

   a   b   c   <-- classified as
   0   0   0 |   a = n
   0 933   0 |   b = o
   0   0   0 |   c = t
```

As I suspected, this tree shows no helpful information in regard to discerning the edible form poisonous mushrooms.