CS 634 - DATA MINING

FINAL TERM PROJECT

Option 1 (Supervised Data Mining)
Support Vector Machines (Category 1)
and
Decision Tree (Category 3)

TO: DR. JASON T.L. WANG, PROFESSOR

DEPARTMENT OF COMPUTER SCIENCE

NEW JERSEY INSTITUTE OF TECHNOLOGY

DONE BY: CHARANPREET KAUR DHIR

UCID: 31478357

CKD22 @NJIT.EDU

OBJECTIVE

Implement 2 classification algorithm i.e. Decision tree and Random Forest tree on a dataset. And for our project I chose Breast Cancer Wisconsin (Prognostic) Data Set**.**

DATA SET FOR THE PROJET

URL : http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Prognostic%29

Title : Breast Cancer Wisconsin (Prognostic) Data Set

DATA SET INFORMATION :

Each record represents follow-up data for one breast cancer case. These are consecutive patients seen by Dr. Wolberg since 1984, and include only those cases exhibiting invasive breast cancer and no evidence of distant metastases at the time of diagnosis.

There are total of 699 breast cancer cases. Each case could be diagnosed as benign or malignant. For the diagnosis, each case has ten attributes listed in the table below. The last attribute in the table represents the class, or label with which the case was diagnosed.

Class distribution:

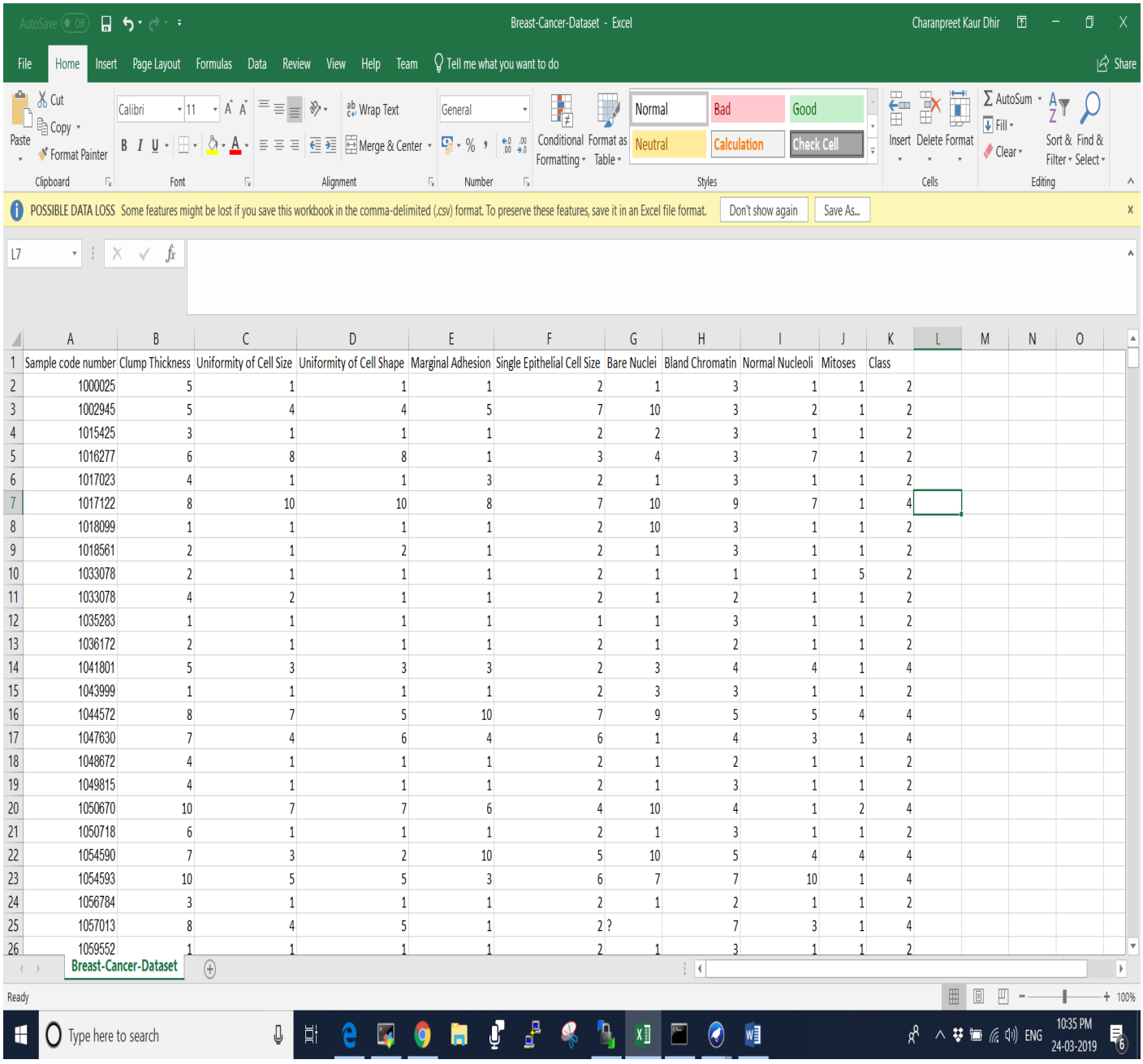
Benign: 458 (65.5%)
Malignant: 241 (34.5%)

ATTRIBUTE INFORMATION:

ATTRIBUTE	DOMAIN
Sample code number	id number
Clump Thickness	1 – 10
Uniformity of Cell Size	1 – 10
Uniformity of Cell Shape	1 – 10
Marginal Adhesion	1 – 10
Single Epithelial Cell Size	1 – 10
Bare Nuclei	1 – 10
Bland Chromatin	1 – 10
Normal Nucleoli	1 – 10
Mitoses	1 – 10
Class:	(2 for benign, 4 for malignant)

DATA IN CSV FILE:

The data is stored in CSV file displayed below and is used further in WEKA.



Sample code number	Clump Thickness	Uniformity of Cell Size	Uniformity of Cell Shape	Marginal Adhesion	Single Epithelial Cell Size	Bare Nuclei	Bland Chromatin	Normal Nucleoli	Mitoses	Class
1000025	5	1	1	1	2	1	3	1	1	2
1002945	5	4	4	5	7	10	3	2	1	2
1015425	3	1	1	1	2	2	3	1	1	2
1016277	6	8	8	1	3	4	3	7	1	2
1017023	4	1	1	3	2	1	3	1	1	2
1017122	8	10	10	8	7	10	9	7	1	4
1018099	1	1	1	1	2	10	3	1	1	2
1018561	2	1	2	1	2	1	3	1	1	2
1033078	2	1	1	1	2	1	1	1	5	2
1033078	4	2	1	1	2	1	2	1	1	2
1035283	1	1	1	1	1	1	3	1	1	2
1036172	2	1	1	1	2	1	2	1	1	2
1041801	5	3	3	3	2	3	4	4	1	4
1043999	1	1	1	1	2	3	3	1	1	2
1044572	8	7	5	10	7	9	5	5	4	4
1047630	7	4	6	4	6	1	4	3	1	4
1048672	4	1	1	1	2	1	2	1	1	2
1049815	4	1	1	1	2	1	3	1	1	2
1050670	10	7	7	6	4	10	4	1	2	4
1050718	6	1	1	1	2	1	3	1	1	2
1054590	7	3	2	10	5	10	5	4	4	4
1054593	10	5	5	3	6	7	7	10	1	4
1056784	3	1	1	1	2	1	2	1	1	2
1057013	8	4	5	1	2	?	7	3	1	4
1059552	1	1	1	1	2	1	3	1	1	2

DECISION TREE (C4.5)

INTRODUCTION

A decision tree is a structure that includes a root node, branches, and leaf nodes. Each internal node denotes a test on an attribute, each branch denotes the outcome of a test, and each leaf node holds a class label. The topmost node in the tree is the root node. Good for categorical attributes with non-continuous values.

ALGORITHM

- The attribute with the highest information gain is selected at each level.
- Suppose there are two classes P (positive) and N (negative) where P contains p training examples and N contains n training examples

$$I(p,n) = -\frac{p}{p+n}\log_2\frac{p}{p+n} - \frac{n}{p+n}\log_2\frac{n}{p+n}$$

- Assume that using attribute A a set S will be partitioned into subsets $\{S_1, S_2, \ldots, S_v\}$
  - If S_i contains p_i examples of P and n_i examples of N, the entropy, or the expected information needed to classify objects in all subtree's S_i is

.

$$E(A) = \sum_{i=1}^{v}\frac{p_i + n_i}{p+n} I(p_i, n_i)$$

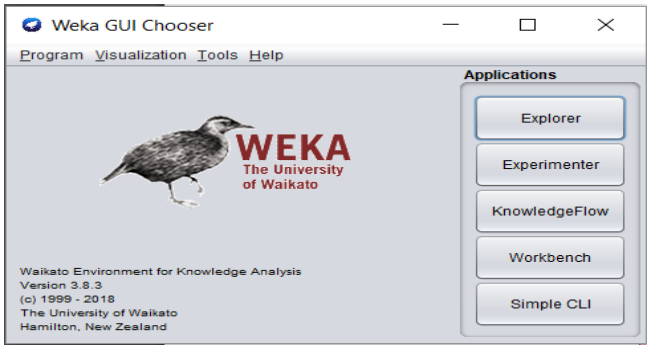
$$Gain(A) = I(p,n) - E(A)$$

SOFTWARE USED: WEKA

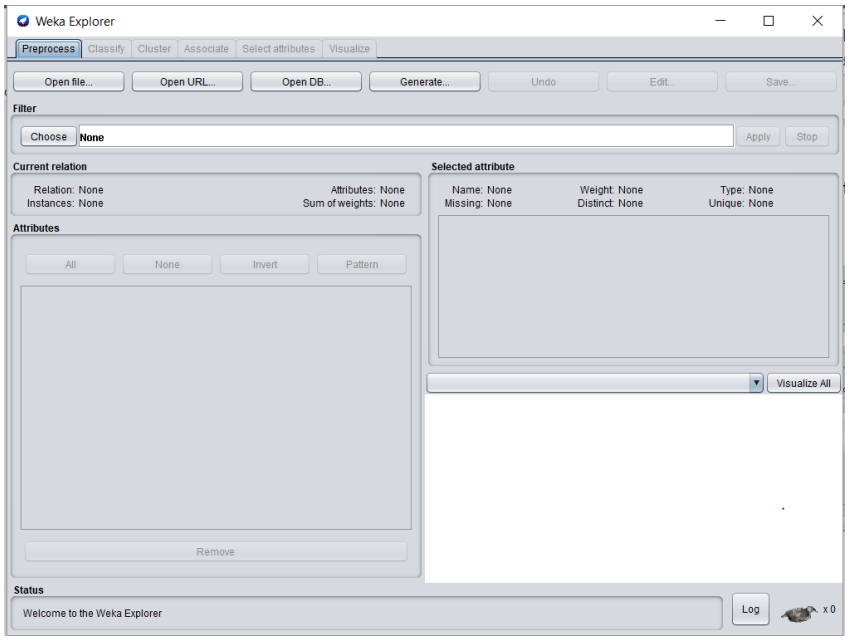
The ID3 algorithm classifies data with discrete values; in order for WEKA to recognize the dataset values as discrete a pre-processing phase must be done in the data before running the algorithm. In this phase the Discretization of the dataset is done, along with the filtering of the "Sample code number" attribute that is not needed for the classification.

STEPS :

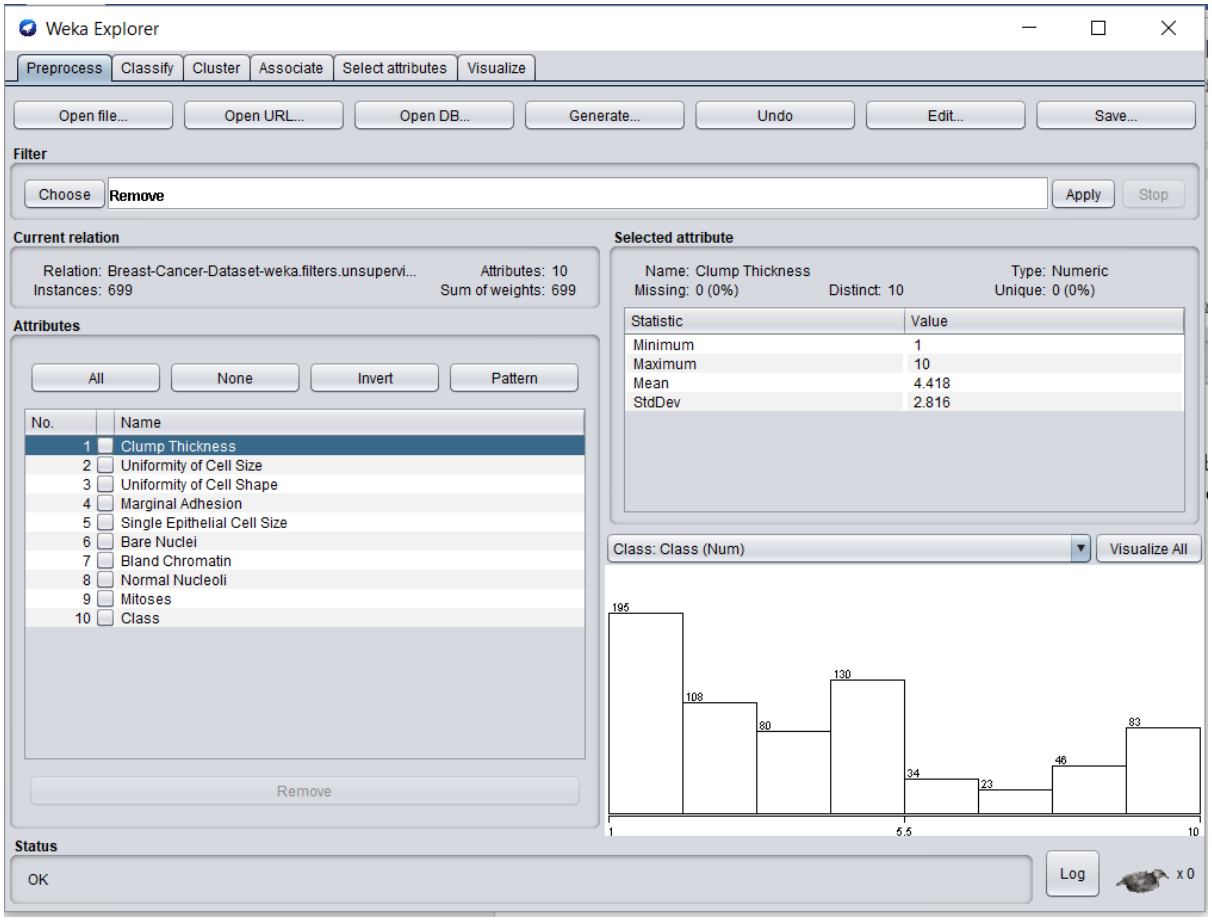
- ✓ We open the WEKA tool and click in the "Explorer" button



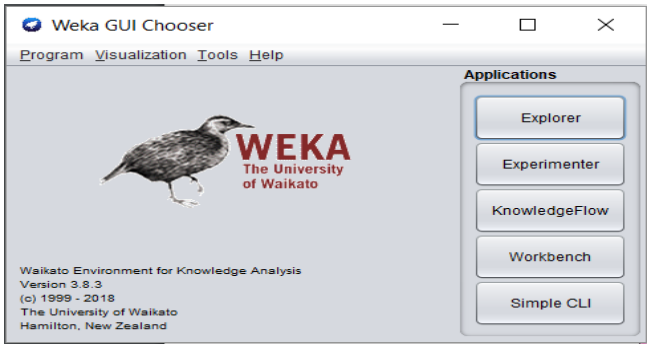
✓ Click on “OPEN FILE” and select the Breast Cancer Data Set csv file stored on the disk.



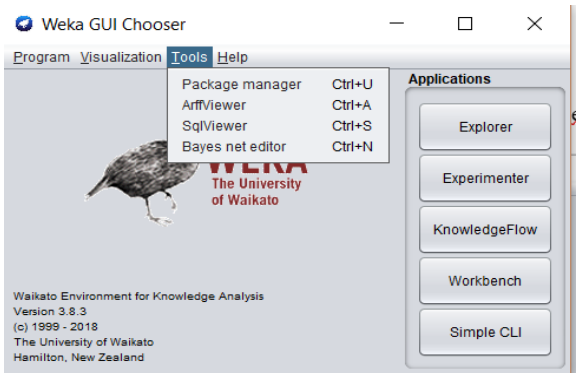
✓ We remove the first attribute “Sample Code Number” attribute by clicking in the Choose button located under the word Filter, and then we choose the option filters -> unsupervised -> attribute -> remove



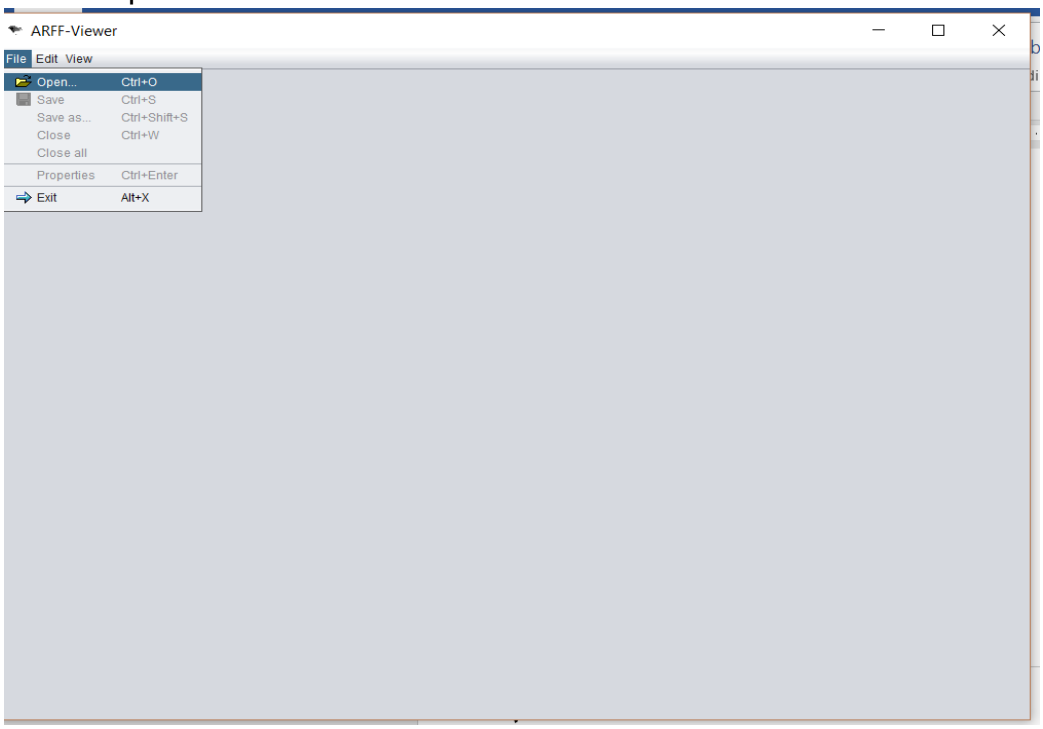
How to create .arff file using weka?



- ✓ Click on tools
- ✓ Click on ArffViewer



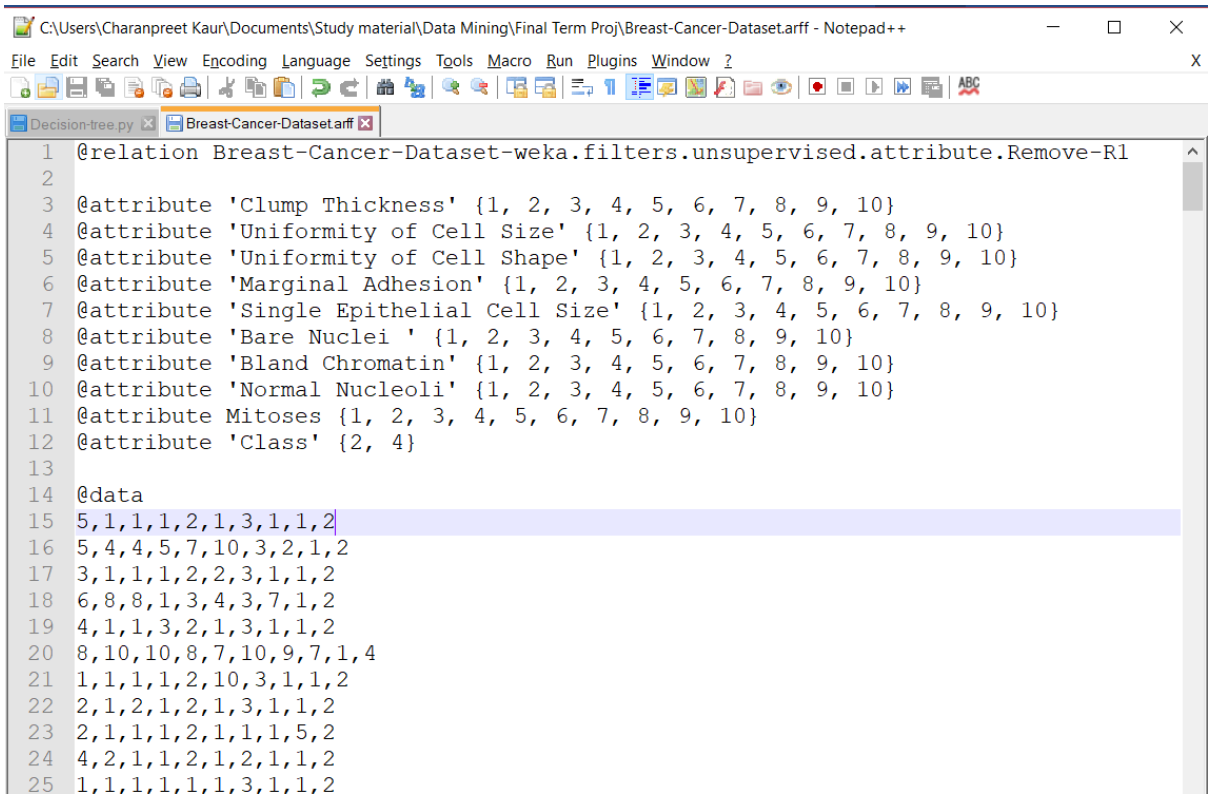
- ✓ File -> Open -> Select the csv data set file



- ✓ Now save it in arff format by file -> save as
- ✓ Open the saved file in notepad++

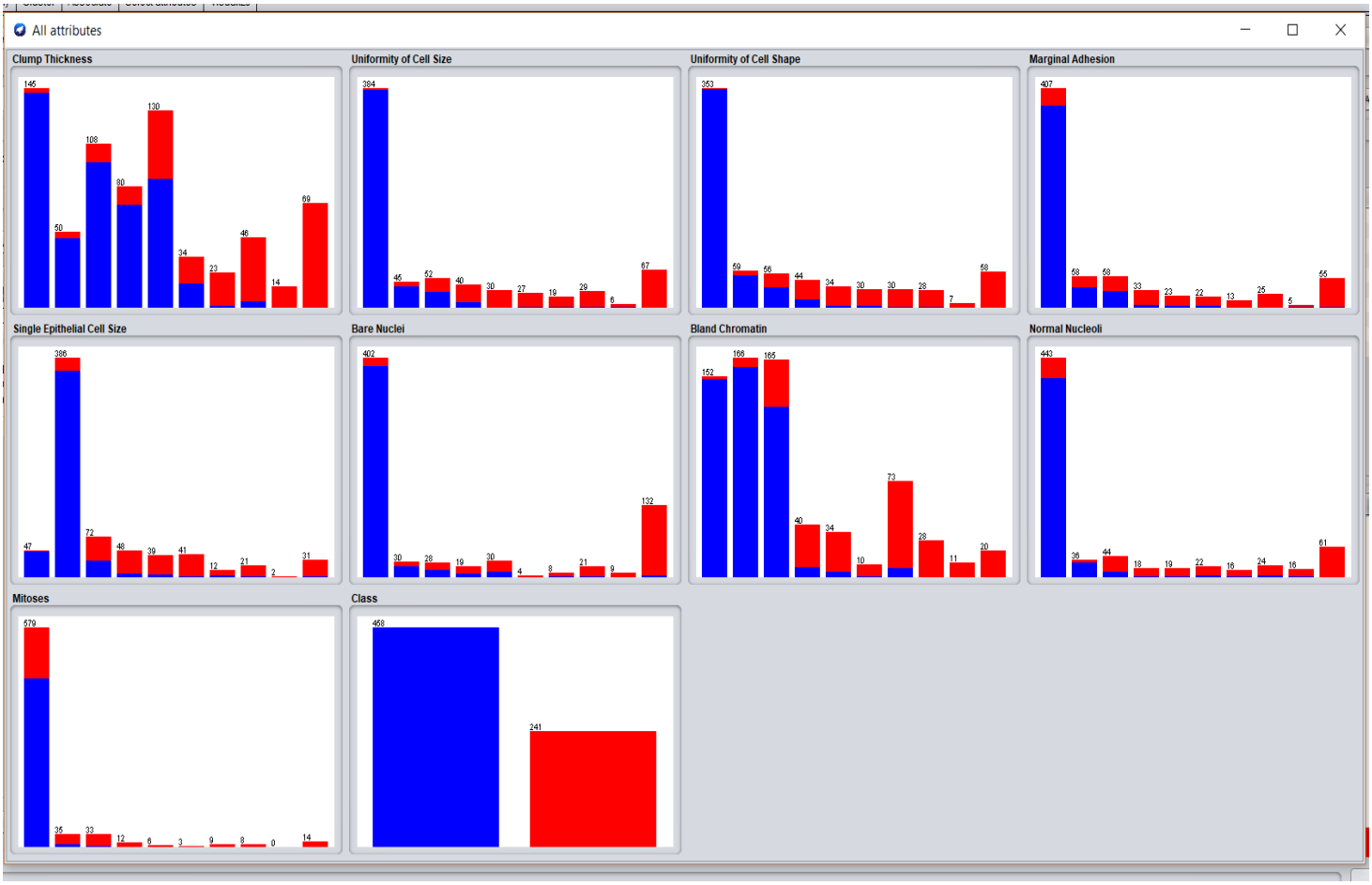
```
@relation Breast-Cancer-Dataset-weka.filters.unsupervised.attribute.Remove-R1

@attribute 'Clump Thickness' numeric
@attribute 'Uniformity of Cell Size' numeric
@attribute 'Uniformity of Cell Shape' numeric
@attribute 'Marginal Adhesion' numeric
@attribute 'Single Epithelial Cell Size' numeric
@attribute 'Bare Nuclei ' numeric
@attribute 'Bland Chromatin' numeric
@attribute 'Normal Nucleoli' numeric
@attribute Mitoses numeric
@attribute Class numeric

@data
5,1,1,1,2,1,3,1,1,2
5,4,4,5,7,10,3,2,1,2
3,1,1,1,2,2,3,1,1,2
6,8,8,1,3,4,3,7,1,2
4,1,1,3,2,1,3,1,1,2
8,10,10,8,7,10,9,7,1,4
1,1,1,1,2,10,3,1,1,2
2,1,2,1,2,1,3,1,1,2
```

- ✓ Now we need to Discretize the dataset: For this we need to go to the arrf file and open it as text. The numeric type must be replaced by the group of discrete numbers that the attribute can have. In our dataset case, all the attributes can have a value between 1-10, so the numeric word will be replaced by {1, 2, 3, 4, 5, 6, 7, 8, 9, 10}. The Class attribute can only have 2 and 4 as values so its discrete numbers will be {2, 4}.



- ✓ After this step we reload our arff file, and now the data will be displayed with different colors depending on the class which the case belongs. Blue for benign cases and Red for malign cases.

- ✓ The data for all attributes is shown below

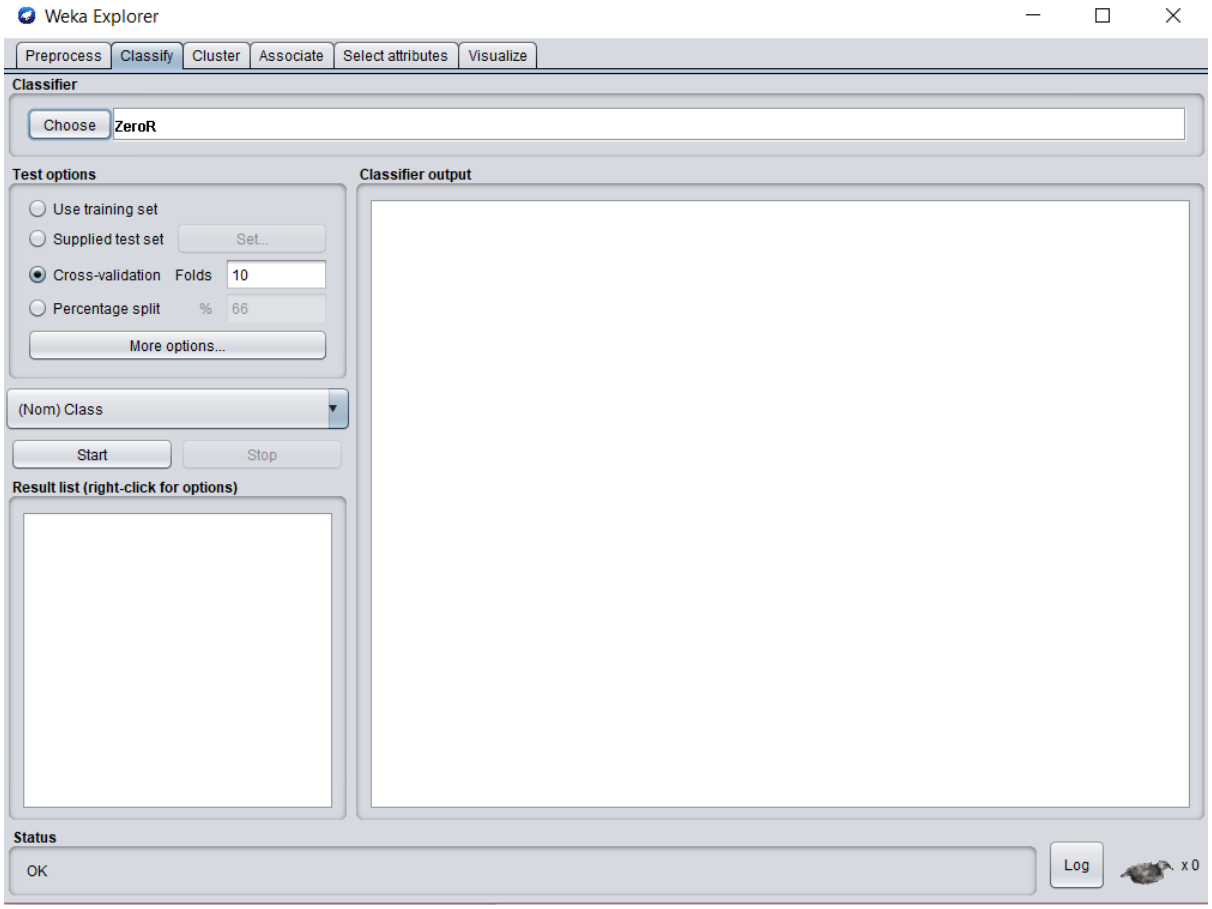


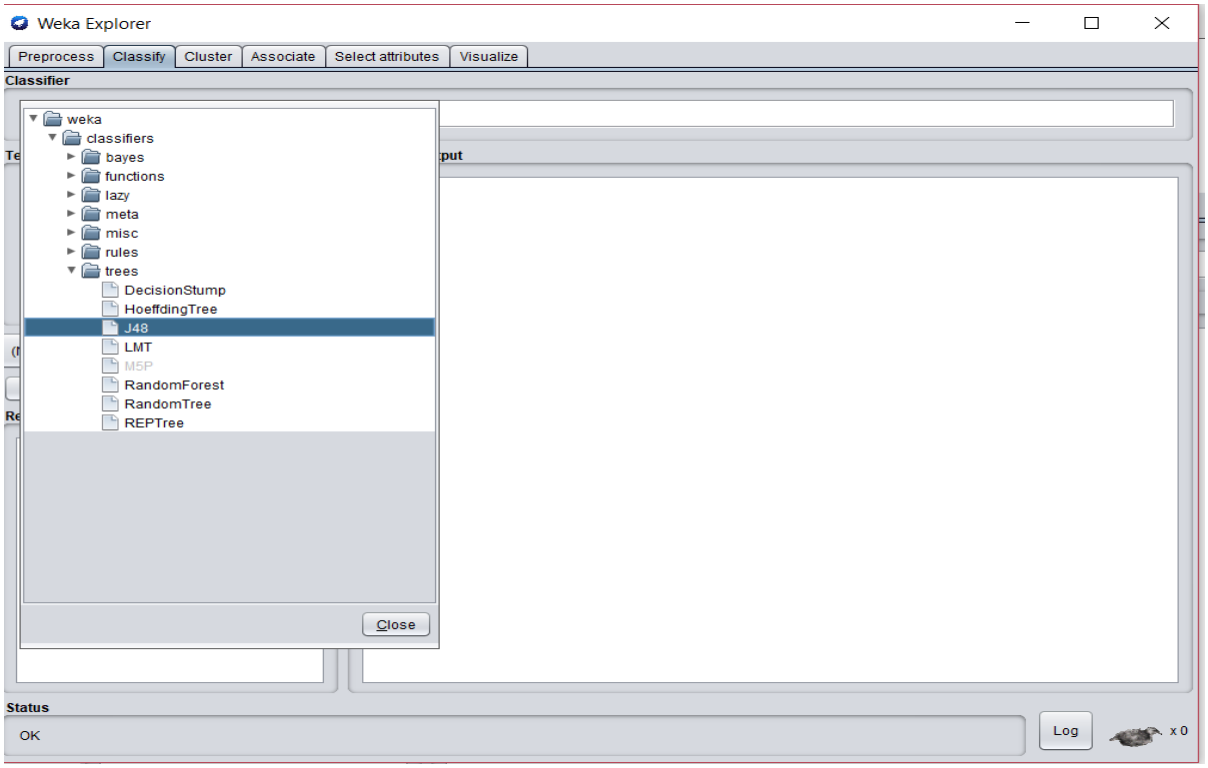
ALGORITHM WORKING:

- ✓ Choose the Algorithm: Select the Classify tab. Then click on the Choose button below the text “Classifier” and next to the text field. Then, in the list of algorithm displayed we proceed to choose the option classifiers-> trees -> J48.

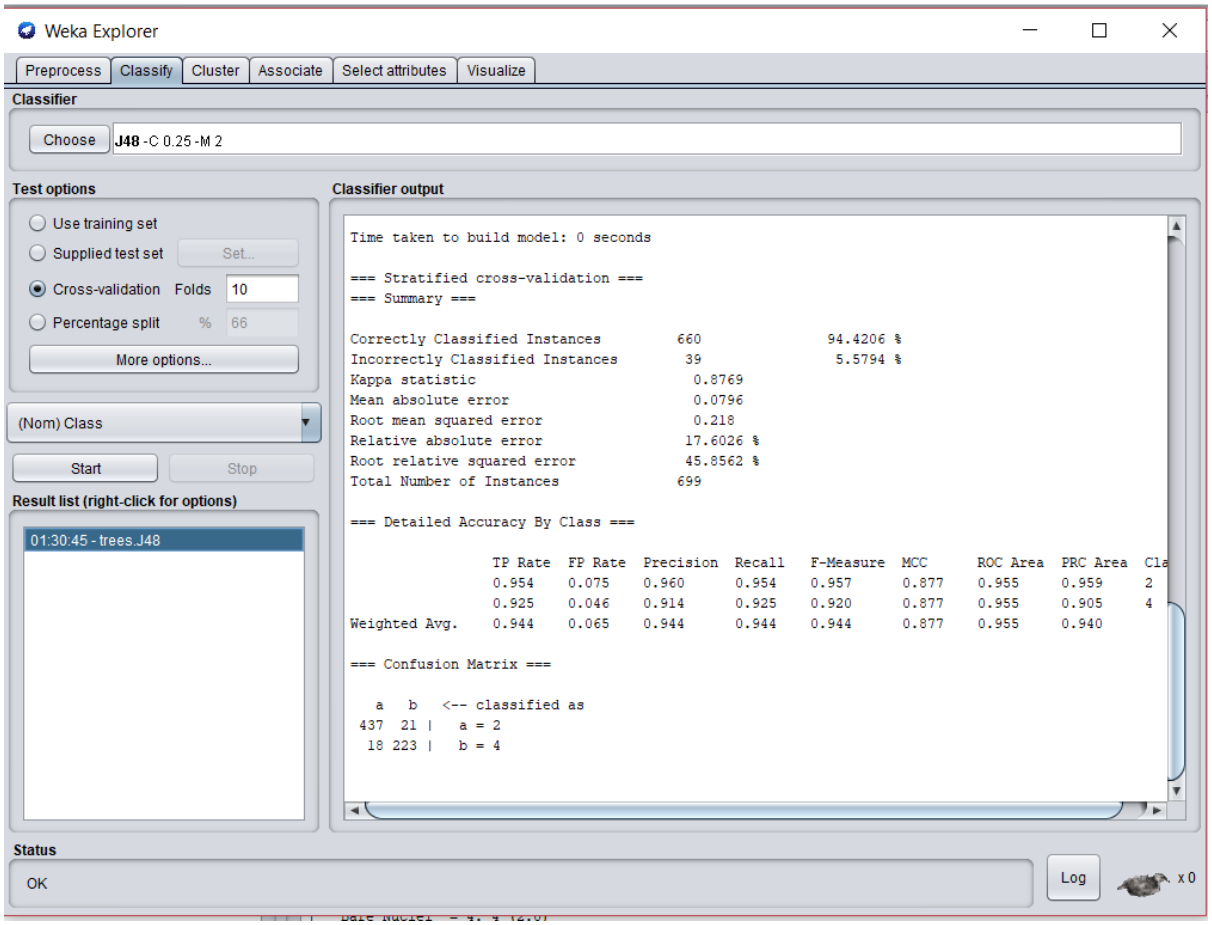
WHY J48?

In WEKA the ID3 algorithm only operates on nominal attributes; and because the attributes of our dataset are numeric the J48 algorithm will be used instead. It must be clear that the ID3 algorithm is the precursor of the J48 algorithm so won’t be any alteration in the results because of this change.

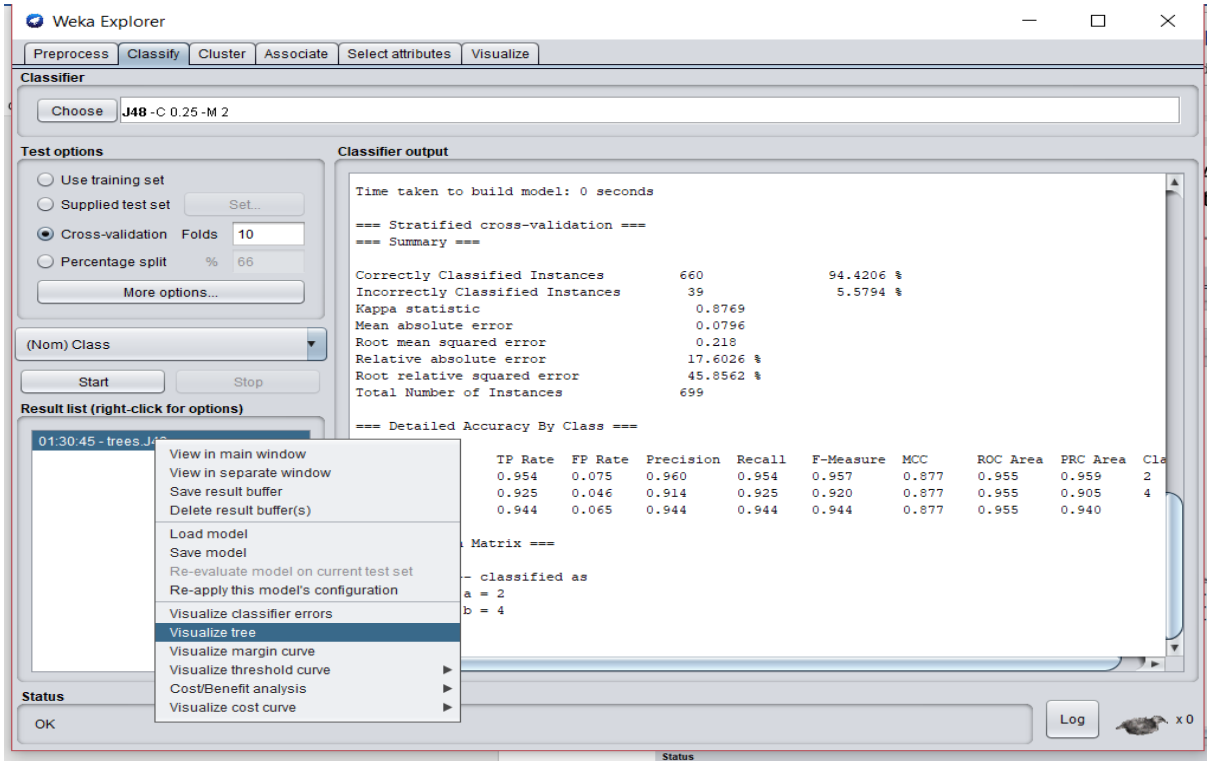

Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
ZeroR
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
Classifier output
(Nom) Class
Start
Stop
Result list (right-click for options)
Status
OK
Log
x 0



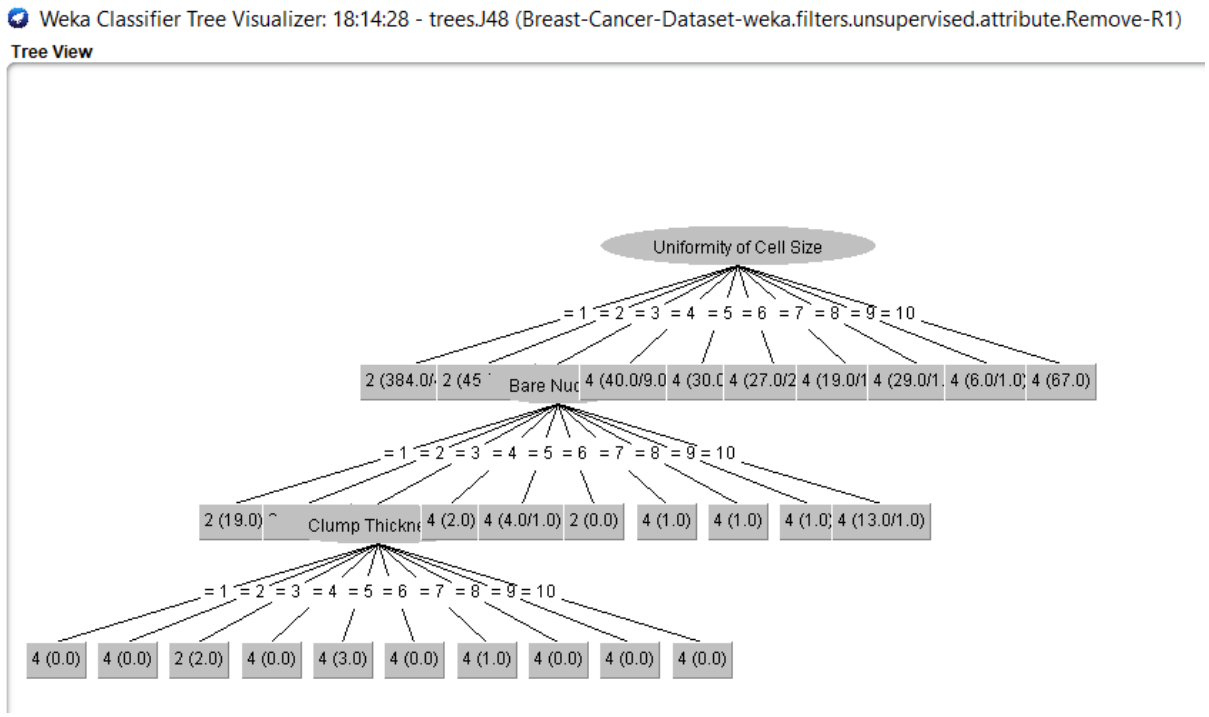
- ✓ According to the requirements we need to do the evaluation using the 10-Fold Cross Validation, we need to set this method in the Test options area by choosing the Cross-validation option. Also, we need to set the number of folds as ten. To run the algorithm, we click on the **Start** button and WEKA will proceed to make the evaluations in the dataset and present the results in the Classifier output area.



- ✓ To view the decision tree of the dataset make a right click on the result description displayed in the Result list area. Then select the option Visualize tree, a new window will appear with the decision tree display on it



DECISION TREE FOR THE DATASET:



Number of Leaves: 28

Size of the Tree: 31

Classifier Output:

=== Run information ===

Scheme: weka.classifiers.trees.J48 -C 0.25 -M 2
Relation: Breast-Cancer-Dataset-weka.filters.unsupervised.attribute.Remove-R1
Instances: 699
Attributes: 10
Clump Thickness
Uniformity of Cell Size
Uniformity of Cell Shape
Marginal Adhesion
Single Epithelial Cell Size
Bare Nuclei
Bland Chromatin
Normal Nucleoli
Mitoses
Class
Test mode: 10-fold cross-validation

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

J48 pruned tree

Uniformity of Cell Size = 1: 2 (384.0/4.0)

Uniformity of Cell Size = 2: 2 (45.0/8.0)

Uniformity of Cell Size = 3

| Bare Nuclei = 1: 2 (19.0)

| Bare Nuclei = 2: 2 (5.0/1.0)

| Bare Nuclei = 3

| | Clump Thickness = 1: 4 (0.0)

| | Clump Thickness = 2: 4 (0.0)

| | Clump Thickness = 3: 2 (2.0)

| | Clump Thickness = 4: 4 (0.0)

| | Clump Thickness = 5: 4 (3.0)

| | Clump Thickness = 6: 4 (0.0)

| | Clump Thickness = 7: 4 (1.0)

| | Clump Thickness = 8: 4 (0.0)

| | Clump Thickness = 9: 4 (0.0)

| | Clump Thickness = 10: 4 (0.0)

| Bare Nuclei = 4: 4 (2.0)

| Bare Nuclei = 5: 4 (4.0/1.0)

| Bare Nuclei = 6: 2 (0.0)

| Bare Nuclei = 7: 4 (1.0)

| Bare Nuclei = 8: 4 (1.0)

| Bare Nuclei = 9: 4 (1.0)

| Bare Nuclei = 10: 4 (13.0/1.0)

Uniformity of Cell Size = 4: 4 (40.0/9.0)

Uniformity of Cell Size = 5: 4 (30.0)

Uniformity of Cell Size = 6: 4 (27.0/2.0)

Uniformity of Cell Size = 7: 4 (19.0/1.0)

Uniformity of Cell Size = 8: 4 (29.0/1.0)

Uniformity of Cell Size = 9: 4 (6.0/1.0)

Uniformity of Cell Size = 10: 4 (67.0)

Number of Leaves : 28

Size of the tree : 31

Time taken to build model: 0 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	660	94.4206 %
Incorrectly Classified Instances	39	5.5794 %
Kappa statistic	0.8769	
Mean absolute error	0.0796	
Root mean squared error	0.218	
Relative absolute error	17.6026 %	
Root relative squared error	45.8562 %	
Total Number of Instances	699	

=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0.954	0.075	0.960	0.954	0.957	0.877	0.955	0.959	2
0.925	0.046	0.914	0.925	0.920	0.877	0.955	0.905	4

Weighted Avg. 0.944 0.065 0.944 0.944 0.944 0.877 0.955 0.940

=== Confusion Matrix ===

```
   a   b   <-- classified as
 437  21 |   a = 2
  18 223 |   b = 4
```