Supervised Learning on Bakary Data Using WEKA

Waleed Pervaiz

CSE 352

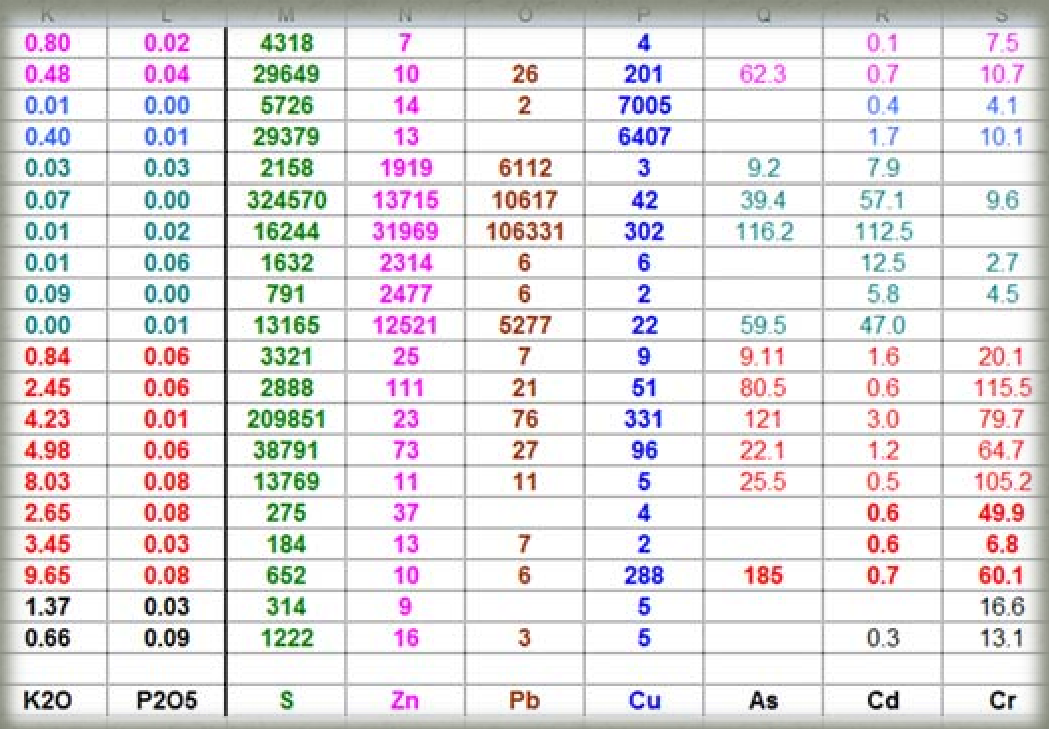
Outline

- Classification Tool: WEKA
- Waikato Environment for Knowledge Analysis by The University of Waikato.
- Available on the internet at:

  http://www.cs.waikato.ac.nz/~ml/weka/index.html

Raw Data

- The Raw data does give us a lot of information.
- However, in this form most of this information is useless and doesn't tell us anything.



K	L	M	N	O	P	Q	R	S
0.80	0.02	4318	7		4		0.1	7.5
0.48	0.04	29649	10	26	201	62.3	0.7	10.7
0.01	0.00	5726	14	2	7005		0.4	4.1
0.40	0.01	29379	13		6407		1.7	10.1
0.03	0.03	2158	1919	6112	3	9.2	7.9	
0.07	0.00	324570	13715	10617	42	39.4	57.1	9.6
0.01	0.02	16244	31969	106331	302	116.2	112.5	
0.01	0.06	1632	2314	6	6		12.5	2.7
0.09	0.00	791	2477	6	2		5.8	4.5
0.00	0.01	13165	12521	5277	22	59.5	47.0	
0.84	0.06	3321	25	7	9	9.11	1.6	20.1
2.45	0.06	2888	111	21	51	80.5	0.6	115.5
4.23	0.01	209851	23	76	331	121	3.0	79.7
4.98	0.06	38791	73	27	96	22.1	1.2	64.7
8.03	0.08	13769	11	11	5	25.5	0.5	105.2
2.65	0.08	275	37		4		0.6	49.9
3.45	0.03	184	13	7	2		0.6	6.8
9.65	0.08	652	10	6	288	185	0.7	60.1
1.37	0.03	314	9		5			16.6
0.66	0.09	1222	16	3	5		0.3	13.1
K2O	P2O5	S	Zn	Pb	Cu	As	Cd	Cr

Data Preparation

- To prepare data for pre-processing the following steps were taken.
- Any attributes that have missing data (i.e. more than 20%) will be removed.
- The following attributes were thus removed:
  - Pb
  - As
  - Cd
  - Ni
  - Sc
  - Co
  - Li
  - Mo

Data Preparation

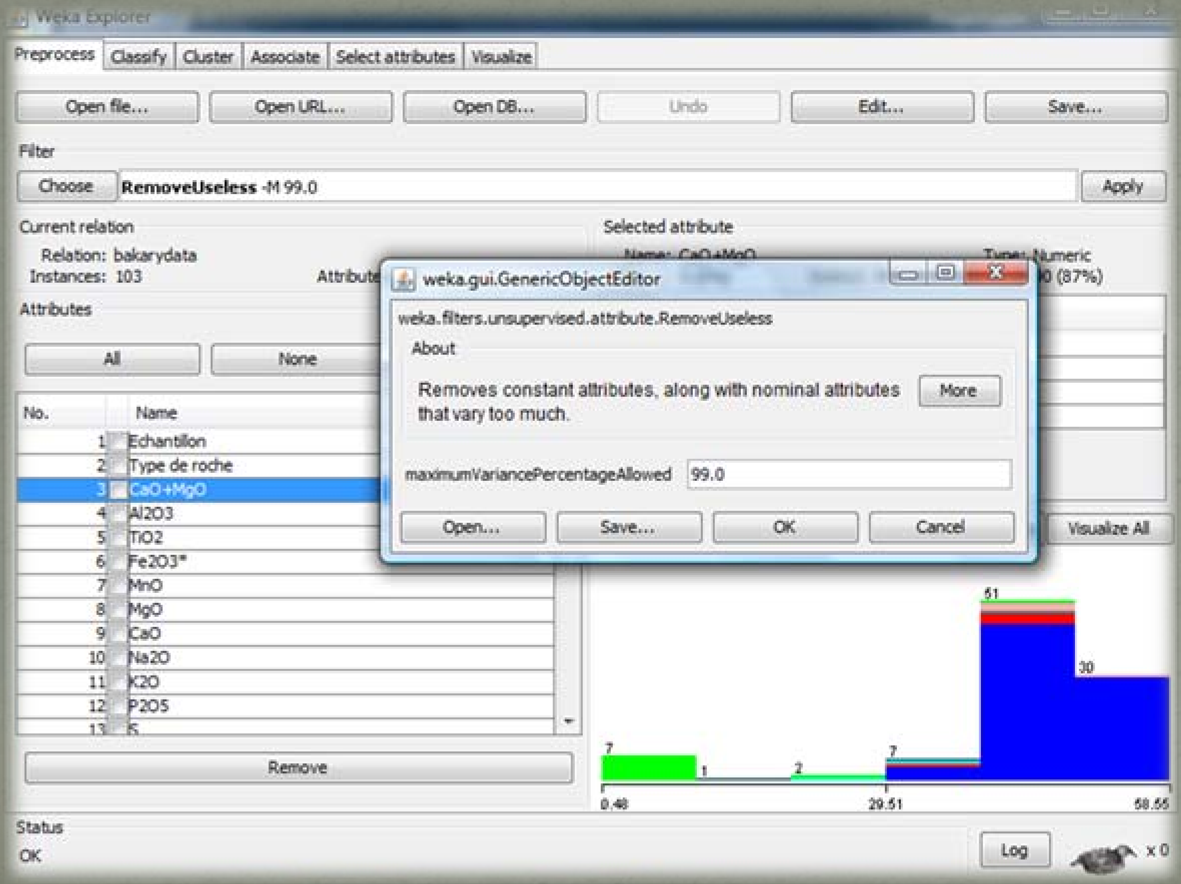
- All other attributes that are missing values were filled in with their averages (mean).
- Missing values for the following attributes were inserted:
  - TiO2 (Carbonates) – Mean: 0.005 (Inserted at E58 and E62)
  - P2O5 (Carbonates) – Mean: 0.74 (Inserted at L33)
  - S (Carbonates) – Mean: 423 (Inserted at M52, M66, M75)
  - Zn (Carbonates) – Mean: 16 (Inserted at N46, N53, N67)
  - Cu (Carbonates) – Mean: 3 (Inserted at O37, O43, O46, O48, O52, O55, O57, O60, O63, O67)
  - Cr (Galene) – Mean: 9.6 (Inserted at P85, P87)
  - Cr (Spahlerite) – Mean: 3.6 (Inserted at P90)
  - V (Carbonates) – Mean: 5.2 (Inserted at Q43, Q48, Q51, Q52, Q60, Q62, Q66, Q71, Q75)
  - V (Galene) – Mean: 2.5 (Inserted at Q86)
  - V (Spahlerite) – Mean: 9.4 (Inserted at Q90)

Data Preparation

- For easier reading, class values were replaced with simpler values.
- The following values were changed:
  - R. carbonatées changed to C1
  - Pyrite changed to C2
  - Chalcopyrites changed to C3
  - Galène changed to C4
  - Spahlerite changed to C5
  - Sédiments terrigènes changed to C6

Data Preparation

- Using WEKA, we remove any noisy data that may unnecessarily skew our data and results.



Weka Explorer

Filter

RemoveUseless -M 99.0

Current relation

Relation: bakarydata

Instances: 103

Selected attribute

Name: CaO+MgO

Type: Numeric

0 (87%)

Attributes

No.	Name
1	Echantillon
2	Type de roche
3	CaO+MgO
4	Al2O3
5	TiO2
6	Fe2O3*
7	MnO
8	MgO
9	CaO
10	Na2O
11	K2O
12	P2O5
13	S

weka.gui.GenericObjectEditor

weka.filters.unsupervised.attribute.RemoveUseless

About

Removes constant attributes, along with nominal attributes that vary too much.

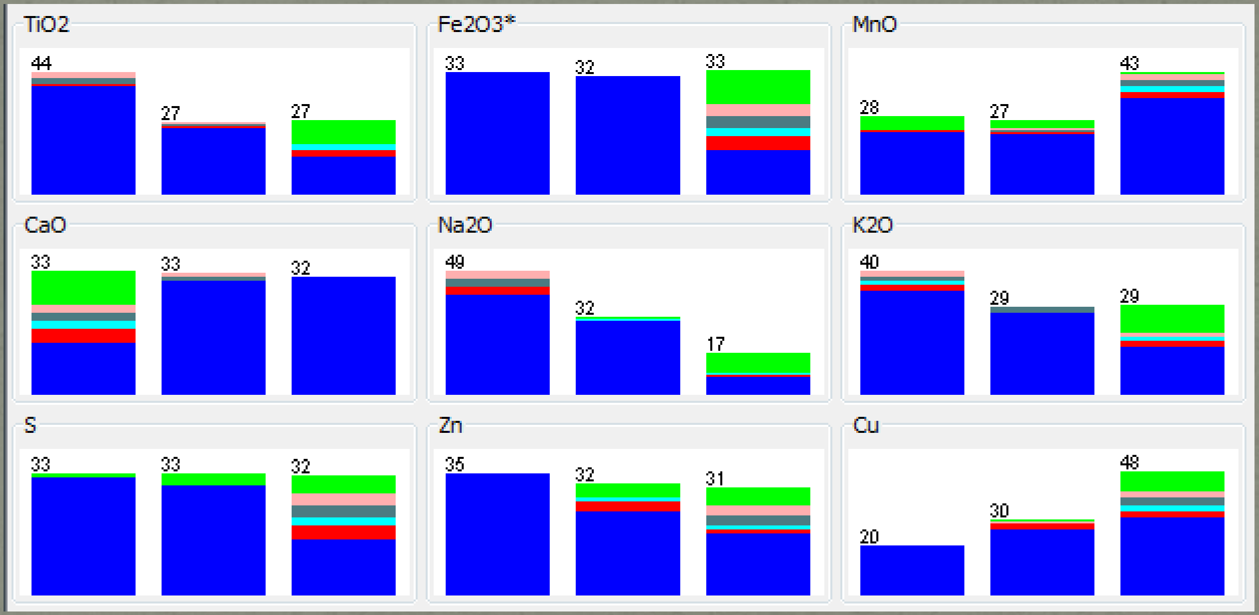
maximumVariancePercentageAllowed 99.0

Status

OK

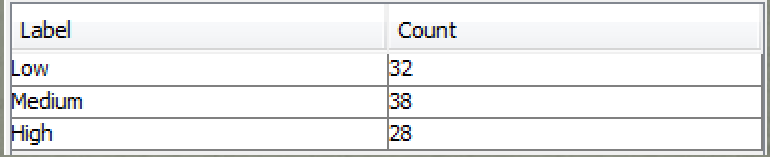
Discretization

- With all the missing data filled in, the noisy data eliminated, we discretize the data using the WEKA tool. (3 equal frequency bins)



Discretization

- Values in the bins were then replaced by specific words:
  - Low
  - Medium
  - High



Label	Count
Low	32
Medium	38
High	28

- This helps in understanding data better.

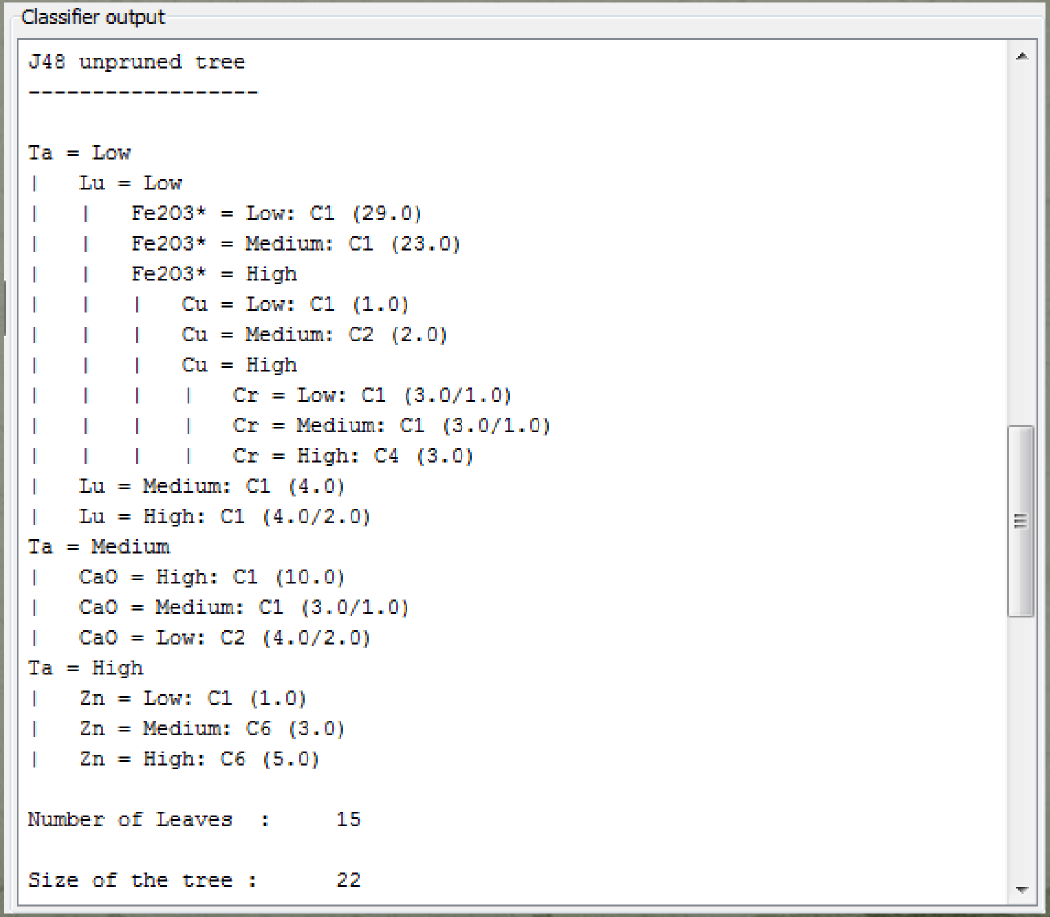
- Decision Tree algorithms will still work with these non-numerical values.

Experiments

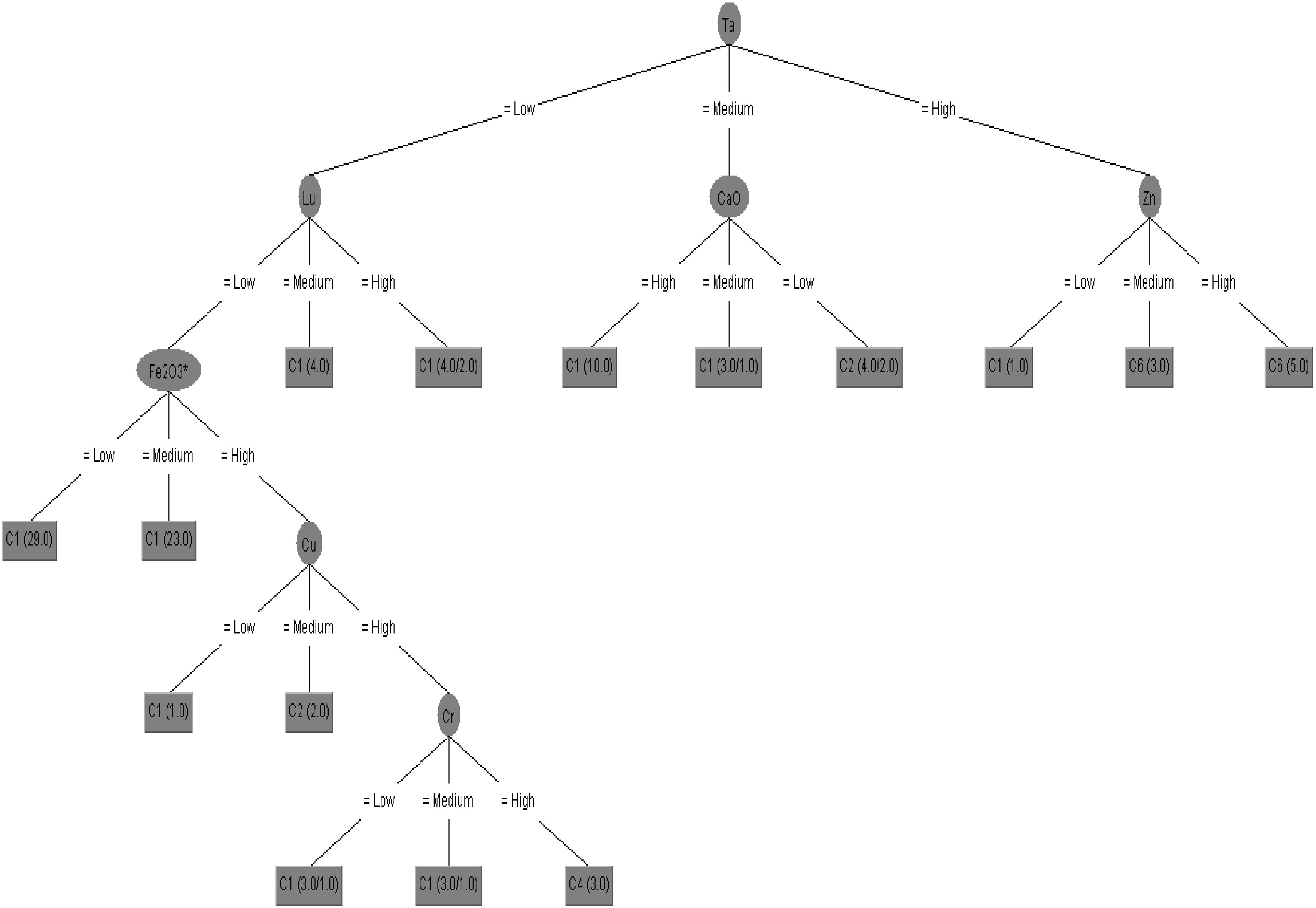
- The following experiments will be carried out on our data:
  - Full Learning: Construction of decision trees characterizing all classes.
  - Contrast Learning: Using all attributes to compare class C1 with the rest of the classes.
  - Limited Learning: Construction of decision tree using only the major attributes.

Experiment 1 - Results

- Experiment 1: Full Learning
- Decision Trees were generated using the J48 algorithm.



```
J48 unpruned tree
------------------

Ta = Low
|   Lu = Low
|   |   Fe2O3* = Low: C1 (29.0)
|   |   Fe2O3* = Medium: C1 (23.0)
|   |   Fe2O3* = High
|   |   |   Cu = Low: C1 (1.0)
|   |   |   Cu = Medium: C2 (2.0)
|   |   |   Cu = High
|   |   |   |   Cr = Low: C1 (3.0/1.0)
|   |   |   |   Cr = Medium: C1 (3.0/1.0)
|   |   |   |   Cr = High: C4 (3.0)
|   Lu = Medium: C1 (4.0)
|   Lu = High: C1 (4.0/2.0)
Ta = Medium
|   CaO = High: C1 (10.0)
|   CaO = Medium: C1 (3.0/1.0)
|   CaO = Low: C2 (4.0/2.0)
Ta = High
|   Zn = Low: C1 (1.0)
|   Zn = Medium: C6 (3.0)
|   Zn = High: C6 (5.0)

Number of Leaves  :     15

Size of the tree :      22
```


Ta
= Low
= Medium
= High
Lu
CaO
Zn
= Low
= Medium
= High
= High
= Medium
= Low
= Low
= Medium
= High
Fe2O3*
C1 (4.0)
C1 (4.0/2.0)
C1 (10.0)
C1 (3.0/1.0)
C2 (4.0/2.0)
C1 (1.0)
C6 (3.0)
C6 (5.0)
= Low
= Medium
= High
C1 (29.0)
C1 (23.0)
Cu
= Low
= Medium
= High
C1 (1.0)
C2 (2.0)
Cr
= Low
= Medium
= High
C1 (3.0/1.0)
C1 (3.0/1.0)
C4 (3.0)

Discriminant Rules

- We got the following discriminant rules.
  - IF Ta="Low" AND Lu="Low" AND Fe2O3="Low" THEN Class="C1"
  - IF Ta="Low" AND Lu="Low" AND Fe2O3="Medium" THEN Class="C1"
  - IF Ta="Low" AND Lu="Low" AND Fe2O3="High" AND Cu="Low" THEN Class="C1"
  - IF Ta="Low" AND Lu="Low" AND Fe2O3="High" AND Cu="Medium" THEN Class="C2"
  - IF Ta="Low" AND Lu="Low" AND Fe2O3="High" AND Cu="High" AND Cr="Low" THEN Class="C1"
  - IF Ta="Low" AND Lu="Low" AND Fe2O3="High" AND Cu="High" AND Cr="Medium" THEN Class="C1"
  - IF Ta="Low" AND Lu="Low" AND Fe2O3="High" AND Cu="High" AND Cr="High" THEN Class="C4"

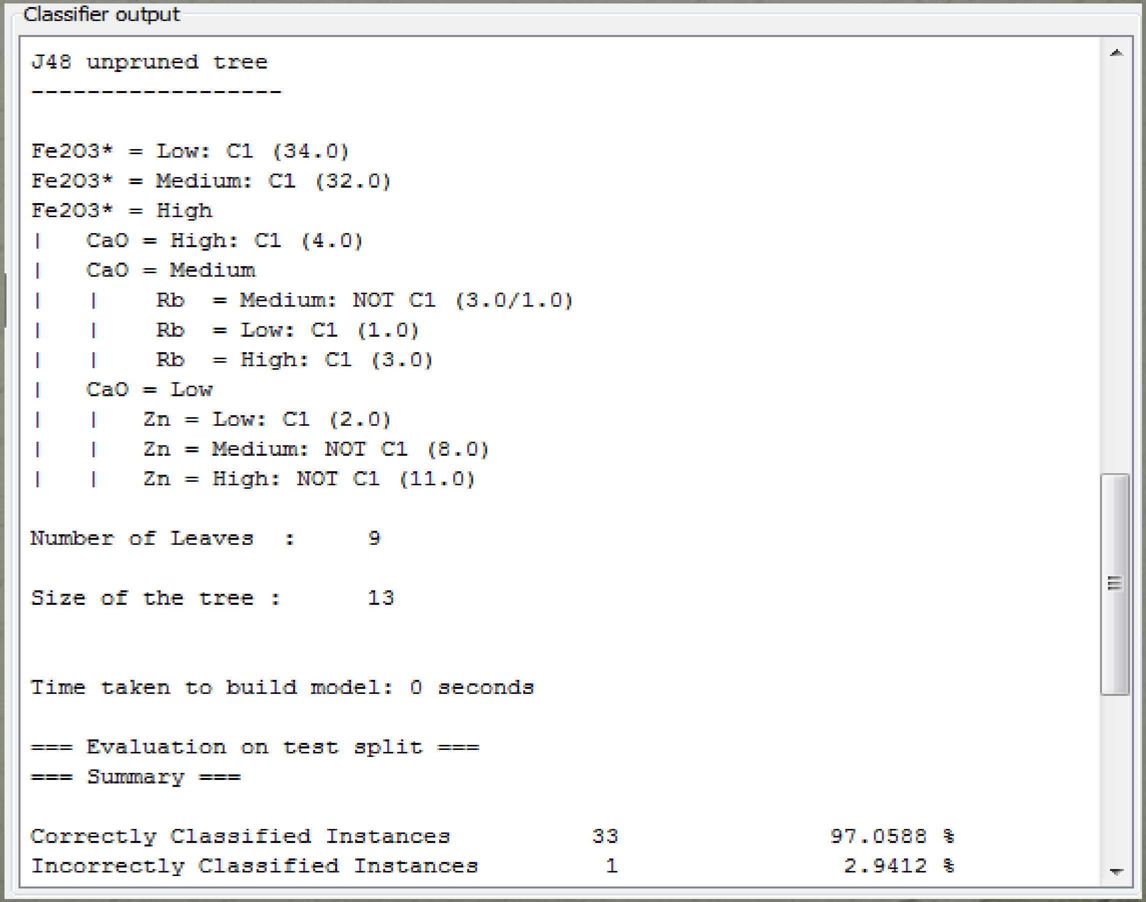
More Discriminant Rules

- We got the following discriminant rules.
  - IF Ta="Low" AND Lu="Medium" THEN Class="C1"
  - IF Ta="Low" AND Lu="High" THEN Class="C1"
  - IF Ta="Medium" AND CaO="High" THEN Class="C1"
  - IF Ta="Medium" AND CaO="Medium" THEN Class="C1"
  - IF Ta="Medium" AND CaO="Low" THEN Class="C2"
  - IF Ta="High" AND Zn="Low" THEN Class="C1"
  - IF Ta="High" AND Zn="Medium" THEN Class="C6"
  - IF Ta="High" AND Zn="High" THEN Class="C6"

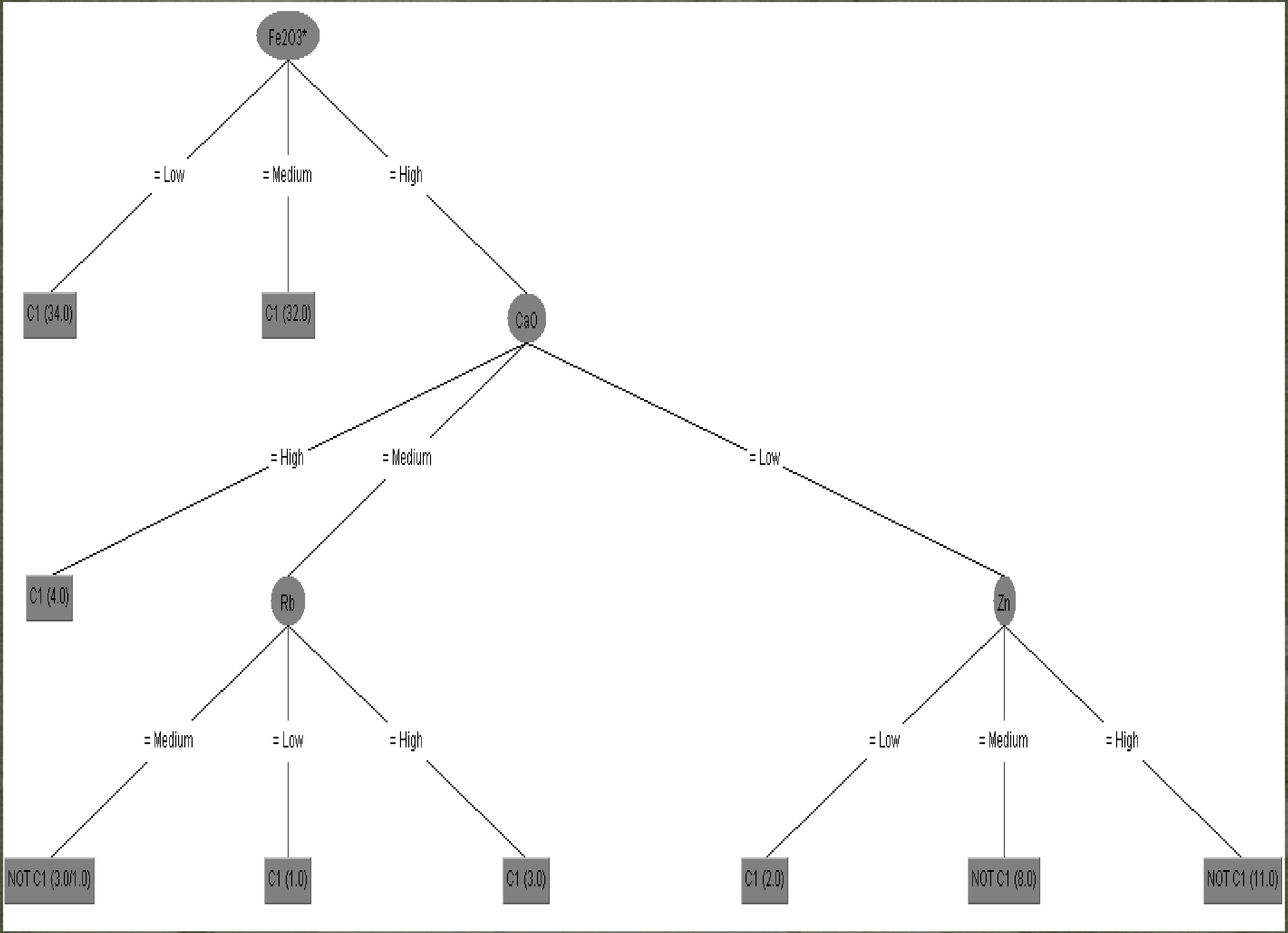
- Predictive Accuracy Determined: 70.58%

Experiment 2 - Results

- Experiment 2: Contrast Learning
- Decision Trees were generated using the J48 algorithm.



```
J48 unpruned tree
------------------


Fe2O3* = Low: C1 (34.0)
Fe2O3* = Medium: C1 (32.0)
Fe2O3* = High
|   CaO = High: C1 (4.0)
|   CaO = Medium
|   |    Rb  = Medium: NOT C1 (3.0/1.0)
|   |    Rb  = Low: C1 (1.0)
|   |    Rb  = High: C1 (3.0)
|   CaO = Low
|   |   Zn = Low: C1 (2.0)
|   |   Zn = Medium: NOT C1 (8.0)
|   |   Zn = High: NOT C1 (11.0)

Number of Leaves  :     9

Size of the tree :      13


Time taken to build model: 0 seconds

=== Evaluation on test split ===
=== Summary ===

Correctly Classified Instances          33               97.0588 %
Incorrectly Classified Instances         1                2.9412 %
```


Fe2O3*
= Low
= Medium
= High
C1 (34.0)
C1 (32.0)
CaO
= High
= Medium
= Low
C1 (4.0)
Rb
Zn
= Medium
= Low
= High
= Low
= Medium
= High
NOT C1 (3.0/1.0)
C1 (1.0)
C1 (3.0)
C1 (2.0)
NOT C1 (8.0)
NOT C1 (11.0)

Discriminant Rules

- We got the following discriminant rules.
  - IF Fe2O3=“Low” THEN Class=“C1”
  - IF Fe2O3=“Medium” THEN Class=“C1”
  - IF Fe2O3=“High” AND CaO=“High” THEN Class=“C1”
  - IF Fe2O3=“High” AND CaO=“Medium” AND Rb=“Medium” THEN Class=“NOT C1”
  - IF Fe2O3=“High” AND CaO=“Medium” AND Rb=“Low” THEN Class=“C1”
  - IF Fe2O3=“High” AND CaO=“Medium” AND Rb=“High” THEN Class=“C1”
  - IF Fe2O3=“High” AND CaO=“Low” AND Zn=“Low” THEN Class=“C1”

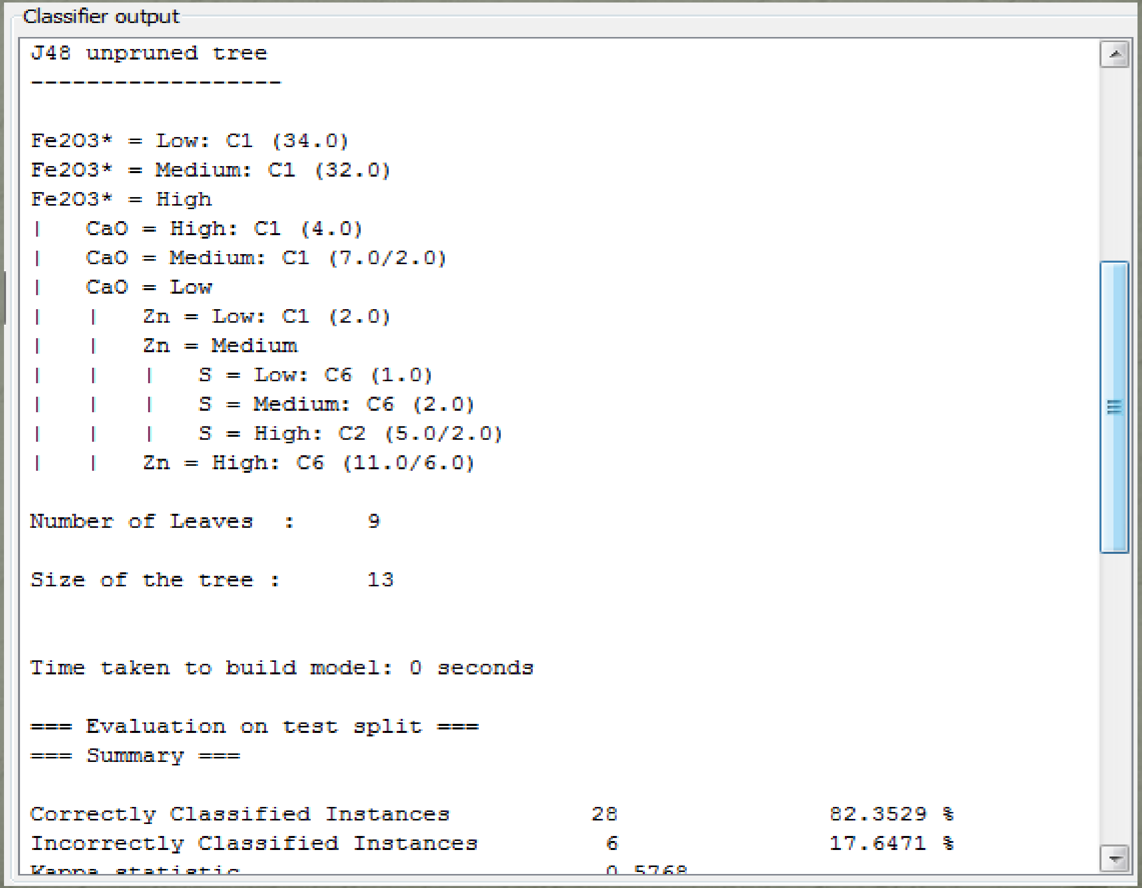
More Discriminant Rules

- We got the following discriminant rules.
  - IF Fe2O3=“High” AND CaO=“Low” AND Zn=“Medium” THEN Class=“NOT C1”
  - IF Fe2O3=“High” AND CaO=“Low” AND Zn=“High” THEN Class=“NOT C1”

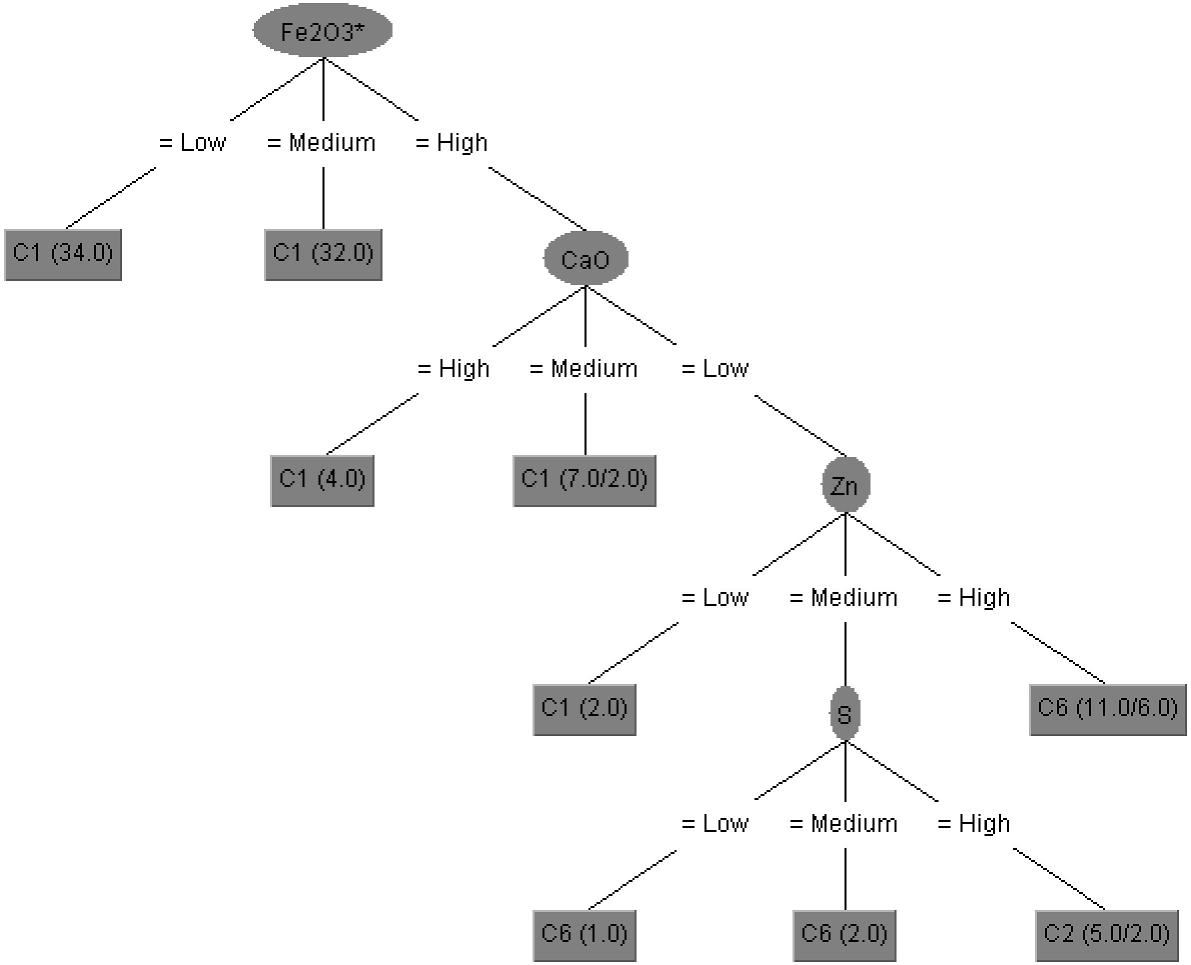
- Predictive Accuracy Determined: 97.06%

Experiment 3 - Results

- Experiment 3: Using Major Attributes
- Decision Trees were generated using the J48 algorithm.



```
J48 unpruned tree
------------------

Fe2O3* = Low: C1 (34.0)
Fe2O3* = Medium: C1 (32.0)
Fe2O3* = High
|   CaO = High: C1 (4.0)
|   CaO = Medium: C1 (7.0/2.0)
|   CaO = Low
|   |   Zn = Low: C1 (2.0)
|   |   Zn = Medium
|   |   |   S = Low: C6 (1.0)
|   |   |   S = Medium: C6 (2.0)
|   |   |   S = High: C2 (5.0/2.0)
|   |   Zn = High: C6 (11.0/6.0)

Number of Leaves  :      9

Size of the tree :      13


Time taken to build model: 0 seconds

=== Evaluation on test split ===
=== Summary ===

Correctly Classified Instances          28               82.3529 %
Incorrectly Classified Instances         6               17.6471 %
Kappa statistic                          0.5768
```


Fe2O3*
= Low
= Medium
= High
C1 (34.0)
C1 (32.0)
CaO
= High
= Medium
= Low
C1 (4.0)
C1 (7.0/2.0)
Zn
= Low
= Medium
= High
C1 (2.0)
S
C6 (11.0/6.0)
= Low
= Medium
= High
C6 (1.0)
C6 (2.0)
C2 (5.0/2.0)

Discriminant Rules

- We got the following discriminant rules.
  - IF Fe2O3=“Low” THEN Class=“C1”
  - IF Fe2O3=“Medium” THEN Class=“C1”
  - IF Fe2O3=“High” AND CaO=“High” THEN Class=“C1”
  - IF Fe2O3=“High” AND CaO=“Medium” THEN Class=“C1”
  - IF Fe2O3=“High” AND CaO=“Low” AND Zn=“Low” THEN Class=“C1”
  - IF Fe2O3=“High” AND CaO=“Low” AND Zn=“Medium” AND S=“Low” THEN Class=“C6”
  - IF Fe2O3=“High” AND CaO=“Low” AND Zn=“Medium” AND S=“Medium” THEN Class=“C6”

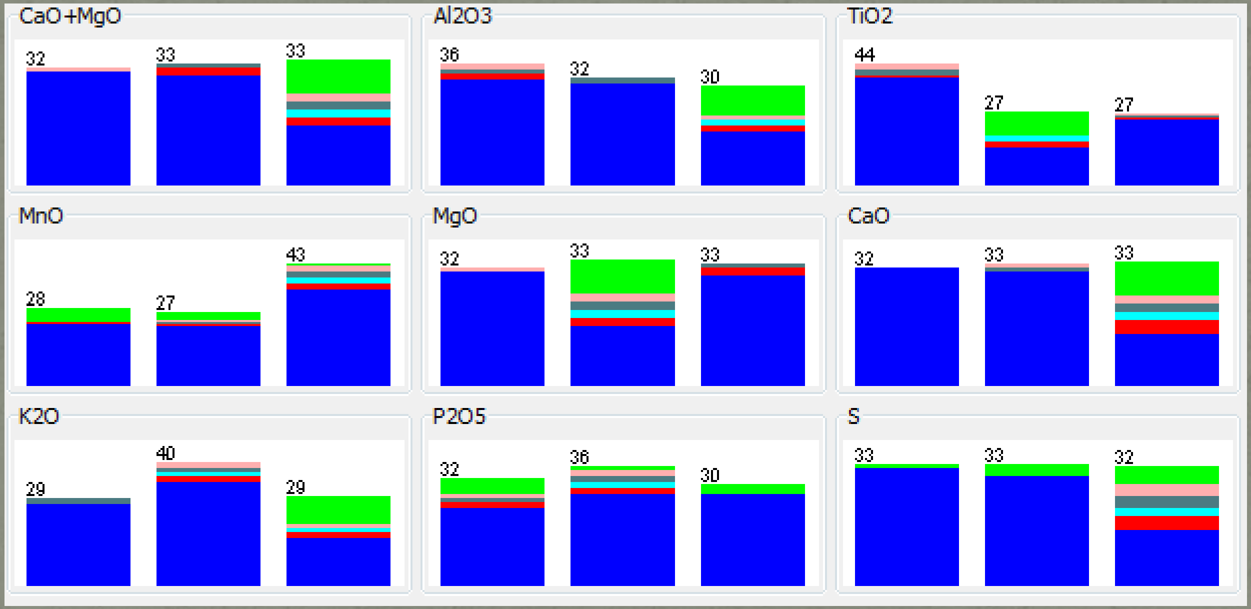
More Discriminant Rules

- We got the following discriminant rules.
  - IF Fe2O3=“High” AND CaO=“Low” AND Zn=“Medium” AND S=“High” THEN Class=“C2”
  - IF Fe2O3=“High” AND CaO=“Low” AND Zn=“High” THEN Class=“C6”

- Predictive Accuracy Determined: 82.35%

Discretization for Dataset 2

- With all the missing data filled in, the noisy data eliminated, we use another method of data discretization. (4 equal width bins)

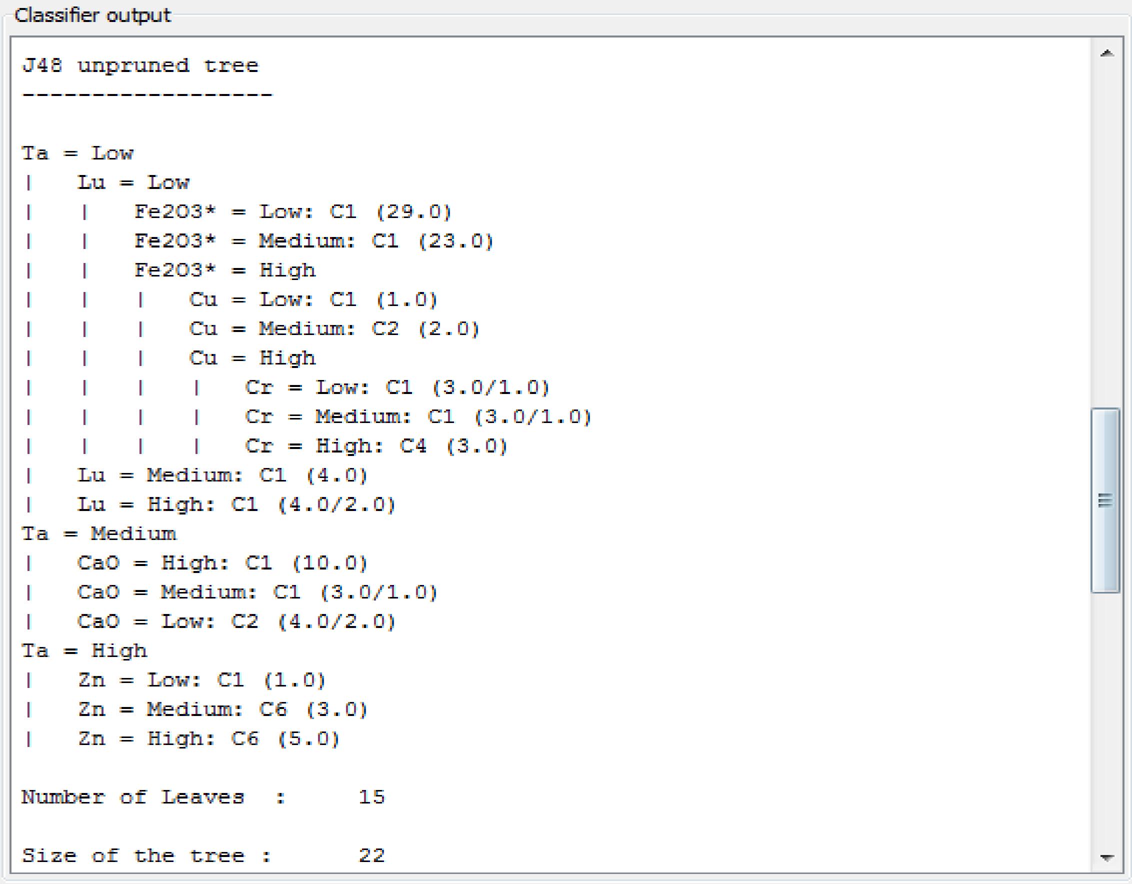


Experiments with Dataset 2

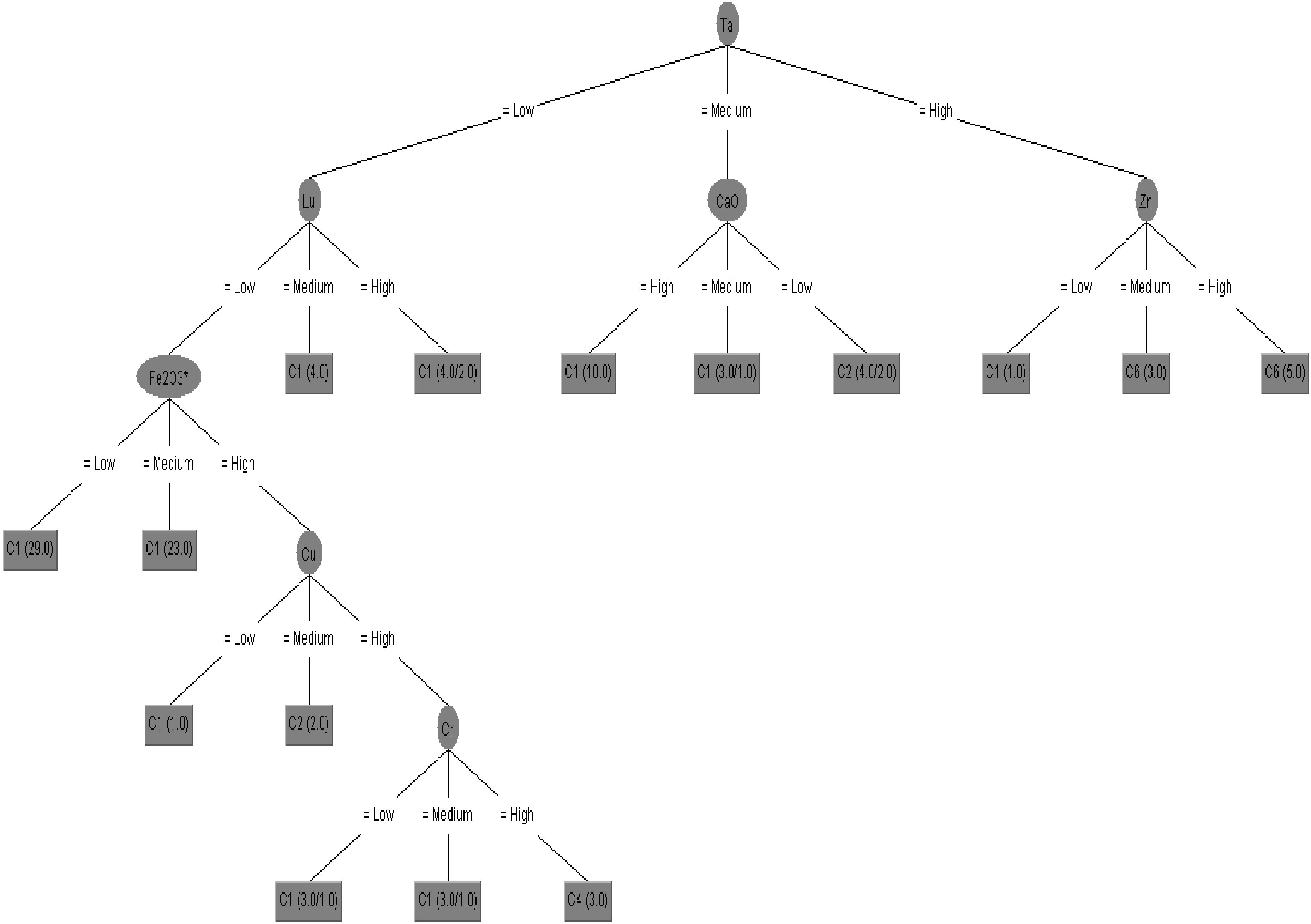
- The following experiments will be carried out on our data:
  - Full Learning: Construction of decision trees characterizing all classes.
  - Contrast Learning: Using all attributes to compare class C1 with the rest of the classes.
  - Limited Learning: Construction of decision tree using only the major attributes.

Experiment 1 Dataset 2- Results

- Experiment 1: Full Learning
- Decision Trees were generated using the J48 algorithm.



```
J48 unpruned tree
------------------


Ta = Low
|   Lu = Low
|   |   Fe2O3* = Low: C1 (29.0)
|   |   Fe2O3* = Medium: C1 (23.0)
|   |   Fe2O3* = High
|   |   |   Cu = Low: C1 (1.0)
|   |   |   Cu = Medium: C2 (2.0)
|   |   |   Cu = High
|   |   |   |   Cr = Low: C1 (3.0/1.0)
|   |   |   |   Cr = Medium: C1 (3.0/1.0)
|   |   |   |   Cr = High: C4 (3.0)
|   Lu = Medium: C1 (4.0)
|   Lu = High: C1 (4.0/2.0)
Ta = Medium
|   CaO = High: C1 (10.0)
|   CaO = Medium: C1 (3.0/1.0)
|   CaO = Low: C2 (4.0/2.0)
Ta = High
|   Zn = Low: C1 (1.0)
|   Zn = Medium: C6 (3.0)
|   Zn = High: C6 (5.0)

Number of Leaves  :     15

Size of the tree :      22
```


Ta
= Low
= Medium
= High
Lu
CaO
Zn
= Low
= Medium
= High
= High
= Medium
= Low
= Low
= Medium
= High
Fe2O3*
C1 (4.0)
C1 (4.0/2.0)
C1 (10.0)
C1 (3.0/1.0)
C2 (4.0/2.0)
C1 (1.0)
C6 (3.0)
C6 (5.0)
= Low
= Medium
= High
C1 (29.0)
C1 (23.0)
Cu
= Low
= Medium
= High
C1 (1.0)
C2 (2.0)
Cr
= Low
= Medium
= High
C1 (3.0/1.0)
C1 (3.0/1.0)
C4 (3.0)

Discriminant Rules

- We got the following discriminant rules.
  - IF Ta="Low" AND Lu="Low" AND Fe2O3="Low" THEN Class="C1"
  - IF Ta="Low" AND Lu="Low" AND Fe2O3="Medium" THEN Class="C1"
  - IF Ta="Low" AND Lu="Low" AND Fe2O3="High" AND Cu="Low" THEN Class="C1"
  - IF Ta="Low" AND Lu="Low" AND Fe2O3="High" AND Cu="Medium" THEN Class="C2"
  - IF Ta="Low" AND Lu="Low" AND Fe2O3="High" AND Cu="High" AND Cr="Low" THEN Class="C1"
  - IF Ta="Low" AND Lu="Low" AND Fe2O3="High" AND Cu="High" AND Cr="Medium" THEN Class="C1"
  - IF Ta="Low" AND Lu="Low" AND Fe2O3="High" AND Cu="High" AND Cr="High" THEN Class="C4"

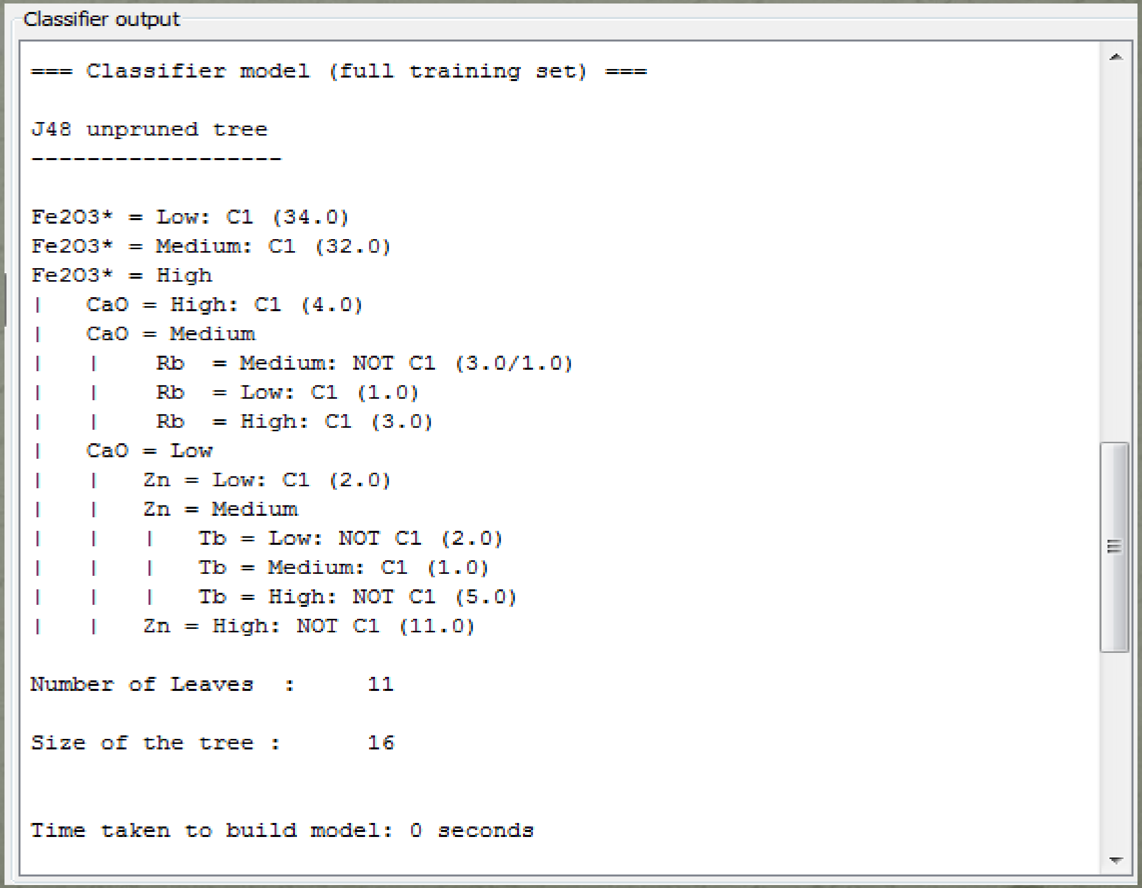
More Discriminant Rules

- We got the following discriminant rules.
  - IF Ta="Low" AND Lu="Medium" THEN Class="C1"
  - IF Ta="Low" AND Lu="High" THEN Class="C1"
  - IF Ta="Medium" AND CaO="High" THEN Class="C1"
  - IF Ta="Medium" AND CaO="Medium" THEN Class="C1"
  - IF Ta="Medium" AND CaO="Low" THEN Class="C2"
  - IF Ta="High" AND Zn="Low" THEN Class="C1"
  - IF Ta="High" AND Zn="Medium" THEN Class="C6"
  - IF Ta="High" AND Zn="High" THEN Class="C6"

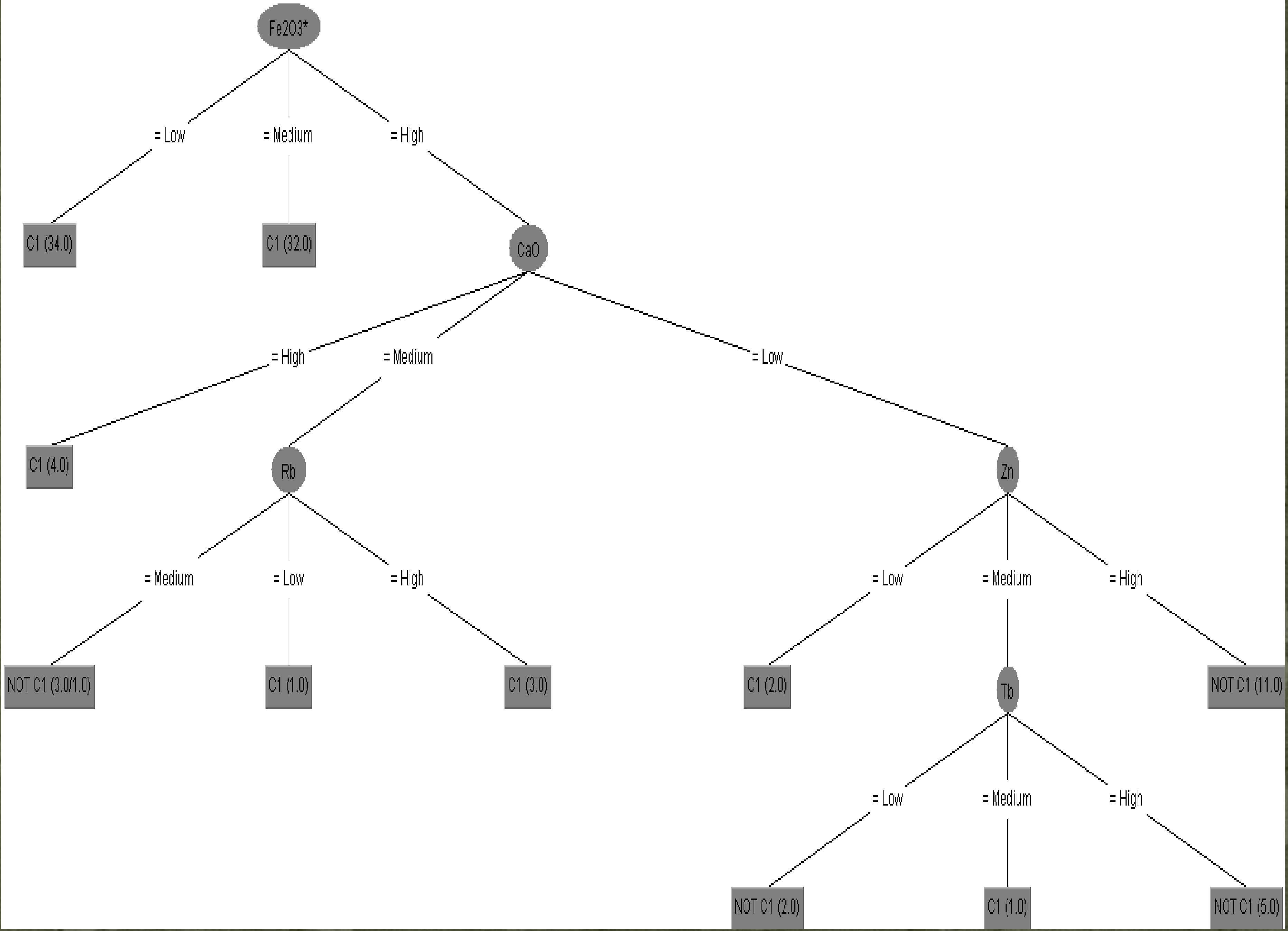
- Predictive Accuracy Determined: 79.59%

Experiment 2 Dataset 2 - Results

- Experiment 2: Contrast Learning
- Decision Trees were generated using the J48 algorithm.



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

J48 unpruned tree
------------------

Fe2O3* = Low: C1 (34.0)
Fe2O3* = Medium: C1 (32.0)
Fe2O3* = High
|   CaO = High: C1 (4.0)
|   CaO = Medium
|   |     Rb   = Medium: NOT C1 (3.0/1.0)
|   |     Rb   = Low: C1 (1.0)
|   |     Rb   = High: C1 (3.0)
|   CaO = Low
|   |    Zn = Low: C1 (2.0)
|   |    Zn = Medium
|   |   |   Tb = Low: NOT C1 (2.0)
|   |   |   Tb = Medium: C1 (1.0)
|   |   |   Tb = High: NOT C1 (5.0)
|   |    Zn = High: NOT C1 (11.0)

Number of Leaves  :      11

Size of the tree :       16


Time taken to build model: 0 seconds
```


Fe2O3*
= Low
= Medium
= High
C1 (34.0)
C1 (32.0)
CaO
= High
= Medium
= Low
C1 (4.0)
Rb
Zn
= Medium
= Low
= High
= Low
= Medium
= High
NOT C1 (3.0/1.0)
C1 (1.0)
C1 (3.0)
C1 (2.0)
Tb
NOT C1 (11.0)
= Low
= Medium
= High
NOT C1 (2.0)
C1 (1.0)
NOT C1 (5.0)

Discriminant Rules

- We got the following discriminant rules.
  - IF Fe2O3="Low" THEN Class="C1"
  - IF Fe2O3="Medium" THEN Class="C1"
  - IF Fe2O3="High" AND CaO="High" THEN Class="C1"
  - IF Fe2O3="High" AND CaO="Medium" AND Rb="Medium" THEN Class="NOT C1"
  - IF Fe2O3="High" AND CaO="Medium" AND Rb="Low" THEN Class="C1"
  - IF Fe2O3="High" AND CaO="Medium" AND Rb="High" THEN Class="C1"
  - IF Fe2O3="High" AND CaO="Low" AND Zn="Low" THEN Class="C1"

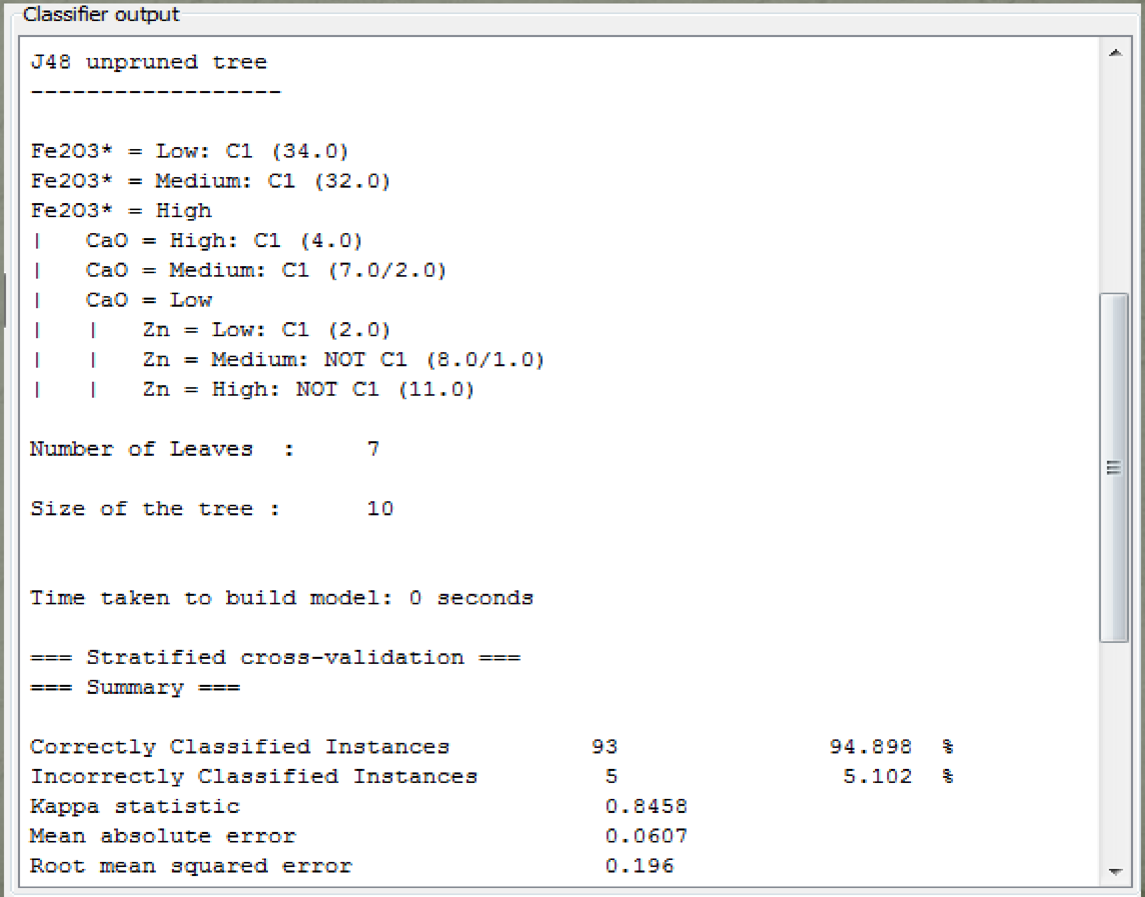
More Discriminant Rules

- We got the following discriminant rules.
  - IF Fe2O3="High" AND CaO="Low" AND Zn="Medium" AND Tb="Low" THEN Class="NOT C1"
  - IF Fe2O3="High" AND CaO="Low" AND Zn="Medium" AND Tb="Medium" THEN Class="C1"
  - IF Fe2O3="High" AND CaO="Low" AND Zn="Medium" AND Tb="High" THEN Class="NOT C1"
  - IF Fe2O3="High" AND CaO="Low" AND Zn="High" THEN Class="NOT C1"

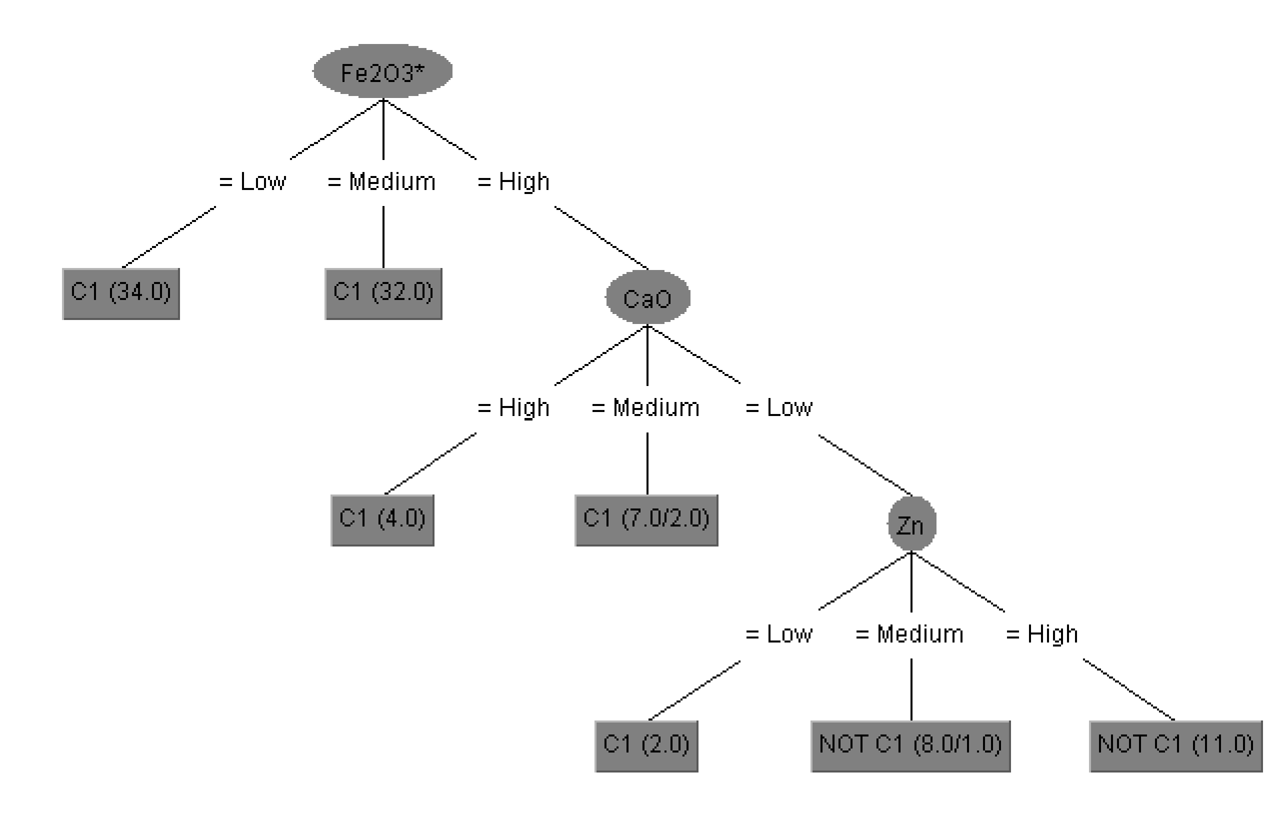
- Predictive Accuracy Determined: 89.79%

Experiment 3 Dataset 2- Results

- Experiment 3: Using Major Attributes
- Decision Trees were generated using the J48 algorithm.



```
J48 unpruned tree
------------------

Fe2O3* = Low: C1 (34.0)
Fe2O3* = Medium: C1 (32.0)
Fe2O3* = High
|   CaO = High: C1 (4.0)
|   CaO = Medium: C1 (7.0/2.0)
|   CaO = Low
|   |   Zn = Low: C1 (2.0)
|   |   Zn = Medium: NOT C1 (8.0/1.0)
|   |   Zn = High: NOT C1 (11.0)

Number of Leaves  :     7

Size of the tree :      10


Time taken to build model: 0 seconds

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

Correctly Classified Instances          93               94.898  %
Incorrectly Classified Instances         5                5.102  %
Kappa statistic                          0.8458
Mean absolute error                      0.0607
Root mean squared error                  0.196
```


Fe2O3*
= Low
= Medium
= High
C1 (34.0)
C1 (32.0)
CaO
= High
= Medium
= Low
C1 (4.0)
C1 (7.0/2.0)
Zn
= Low
= Medium
= High
C1 (2.0)
NOT C1 (8.0/1.0)
NOT C1 (11.0)

Discriminant Rules

- We got the following discriminant rules.
  - IF Fe2O3=“Low” THEN Class=“C1”
  - IF Fe2O3=“Medium” THEN Class=“C1”
  - IF Fe2O3=“High” AND CaO=“High” THEN Class=“C1”
  - IF Fe2O3=“High” AND CaO=“Medium” THEN Class=“C1”
  - IF Fe2O3=“High” AND CaO=“Low” AND Zn=“Low” THEN Class=“C1”
  - IF Fe2O3=“High” AND CaO=“Low” AND Zn=“Medium” THEN Class=“NOT C1”
  - IF Fe2O3=“High” AND CaO=“Low” AND Zn=“High” THEN Class=“NOT C1”

More Discriminant Rules

- Predictive Accuracy Determined: 94.90%

Accuracy Analysis for both Datasets

- Here is a comparison of the accuracy achieved with each Dataset.

	Dataset #1	Dataset #2
Experiment 1	70.58%	79.59%
Experiment 2	97.06%	89.79%
Experiment 3	82.35%	94.90%

- Dataset 1 was carried out using 3 bins – equal frequency discretization.

- Dataset 2 was carried out using 4 bins – equal width discretization.

Conclusion & Thoughts

- High accuracy for a particular value can sometimes be misleading since there is a lot of data (77 records) for C1 as compared to data (21 records) for other classes.
- WEKA produces different rules depending on the techniques used for data preparation.
- Dataset 2 generally had better accuracy. Thus, we can conclude that the 4-bin equal width method was slightly more accurate than the 3-bin equal frequency method.
- Comparing classes with each other gave the best overall accuracy. (i.e. comparing C1 with all other classes)