Q1:
Generate the classification rules using OneR algorithm with the test option of 66% percentage split. Justify quantitatively how the rules were generated.

```plaintext
Classifier output

Instances: 625
Attributes: 5
    left-weight
    left-distance
    right-weight
    right-distance
    class

Test mode: split 66.0% train, remainder test

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

left-weight:
    < 2.5  -> R
    >= 2.5  -> L
(397/625 instances correct)

Time taken to build model: 0 seconds

=== Evaluation on test split ===

Correctly Classified Instances 122 57.5472 %
```
### Run Information ###

Scheme: weka.classifiers.rules.OneR -B 6  
Relation: balance-scale  
Instances: 625  
Attributes: 5  
  left-weight  
  left-distance  
  right-weight  
  right-distance  
  class  
Test mode: split 66.0% train, remainder test  

### Classifier model (full training set) ###

left-weight:  
  < 2.5  -> R  
  >= 2.5  -> L  
(397/625 instances correct)  

Time taken to build model: 0 seconds  

### Detailed Accuracy By Class ###

<table>
<thead>
<tr>
<th>Class</th>
<th>TP Rate</th>
<th>FP Rate</th>
<th>Precision</th>
<th>Recall</th>
<th>F-Measure</th>
<th>ROC Area</th>
</tr>
</thead>
<tbody>
<tr>
<td>L</td>
<td>0.505</td>
<td>0.278</td>
<td>0.605</td>
<td>0.505</td>
<td>0.551</td>
<td>0.613</td>
</tr>
<tr>
<td>B</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.5</td>
</tr>
<tr>
<td>R</td>
<td>0.76</td>
<td>0.5</td>
<td>0.557</td>
<td>0.76</td>
<td>0.643</td>
<td>0.63</td>
</tr>
</tbody>
</table>

Weighted Avg.  

### Confusion Matrix ###

a  b  c  <-- classified as  
49 0 48  | a = L  
9 0 10   | b = B  
23 0 73  | c = R

(following the left figure)
The procedure of the quantitative justification (referring Figure 4.1):
Weka Explorer

Filter
Choose Discretize -v -R last

Current relation
Relation: balance-scale
Instances: 625
Attributes: 5

Attributes

About
An instance filter that discretizes a range of numeric attributes in the dataset into nominal attributes.

attributeIndices: last
invertSelection: True
makeBinary: False
useBetterEncoding: False
useKononenko: False
Weka Explorer

Preprocess | Classify | Cluster | Associate | Select attributes | Visualize |

Open file... | Open URL... | Open DB... | Generate... | Undo | Edit... | Save...

Filter

Choose Discretize -v -r last

Current relation

Relation: balance-scale-weka.filters.supervised.attribute.Discretize-V...
Instances: 625
Attributes: 5

Selected attribute

Name: left-weight
Missing: 0 (0%)
Distinct: 2
Type: Nominal
Unique: 0 (0%)

No. | Label | Count
--- | --- | ---
1 | (-inf-2.5] | 250
2 | (2.5-inf] | 375

Attributes

No. | Name
--- | ---
1 | left-weight
2 | left-distance
3 | right-weight
4 | right-distance
5 | class
Choose **RemoveWithValues** -S 0.0 -C last -L first -V

**weka.gui.GenericObjectEditor**

weka.filters.unsupervised.instance.RemoveWithValues

**About**
Filters instances according to the value of an attribute.

- **attributeIndex**: last
- **invertSelection**: True
- **matchMissingValues**: False
- **modifyHeader**: False
- **nominalIndices**: first
- **splitPoint**: 0.0
You can repeat the same procedure to generate the other statistics.
<table>
<thead>
<tr>
<th>attribute</th>
<th>category</th>
<th>num</th>
<th>L</th>
<th>B</th>
<th>R</th>
</tr>
</thead>
<tbody>
<tr>
<td>left-weight</td>
<td>(-∞, 2.5]</td>
<td>250</td>
<td>60 (24%)</td>
<td>21 (8.4%)</td>
<td>169 (67.6%)</td>
</tr>
<tr>
<td></td>
<td>(2.5, ∞)</td>
<td>375</td>
<td>228 (60.8%)</td>
<td>28 (7.5%)</td>
<td>119 (31.7%)</td>
</tr>
<tr>
<td>left-distance</td>
<td>(-∞, 2.5]</td>
<td>250</td>
<td>60 (24%)</td>
<td>21 (8.4%)</td>
<td>169 (67.6%)</td>
</tr>
<tr>
<td></td>
<td>(2.5, ∞)</td>
<td>375</td>
<td>228 (60.8%)</td>
<td>28 (7.5%)</td>
<td>119 (31.7%)</td>
</tr>
<tr>
<td>right-weight</td>
<td>(-∞, 2.5]</td>
<td>250</td>
<td>169 (67.6%)</td>
<td>21 (8.4%)</td>
<td>60 (24%)</td>
</tr>
<tr>
<td></td>
<td>(2.5, ∞)</td>
<td>375</td>
<td>119 (31.7%)</td>
<td>28 (7.5%)</td>
<td>228 (60.8%)</td>
</tr>
<tr>
<td>right-distance</td>
<td>(-∞, 2.5]</td>
<td>250</td>
<td>169 (67.6%)</td>
<td>21 (8.4%)</td>
<td>60 (24%)</td>
</tr>
<tr>
<td></td>
<td>(2.5, ∞)</td>
<td>375</td>
<td>119 (31.7%)</td>
<td>28 (7.5%)</td>
<td>228 (60.8%)</td>
</tr>
</tbody>
</table>

1. left-weight   (-∞, 2.5] => R    (2.5, ∞) => L
2. left-distance (-∞, 2.5] => R    (2.5, ∞) => L
3. right-weight  (-∞, 2.5] => L    (2.5, ∞) => R
4. right-distance(-∞, 2.5] => L    (2.5, ∞) => R
Q2:
Generate a decision tree using the J48 algorithm with the test option of 66% percentage split. Was the root attribute selected based on the notion of information gain? Justify quantitatively your answer. For this purpose, assume that each numerical attribute is discretized into two ranges, the first consists of all values less than or equal to 2 and the second consists of all the values larger than 2.

A2:
The root attribute is “left-weight” shown as below (partly).
--- Run information ---

Evaluator: weka.attributeSelection.InfoGainAttributeEval
Search: weka.attributeSelection.Ranker -T -1.7976931348623157E308 -N -1
Relation: balance-scale
Instances: 62f
Attributes: 5
  left-weight
  left-distance
  right-weight
  right-distance
  class
Evaluation mode: evaluate on all training data

--- Attribute Selection on all input data ---

Search Method:
  Attribute ranking.

Attribute Evaluator: (supervised, Class (nominal): 5 class):
  Information Gain Ranking Filter

Ranked attributes:
  0.121 2 left-distance
  0.121 1 left-weight
  0.121 4 right-distance
  0.121 3 right-weight

Selected attributes: 2,1,4,3 : 4
The procedure of the quantitative justification:

Note that since all numeric values are integers, you will get the same statistics whether you split at 2.5 or 2.0. So, you can use the previous results of question 1.
\[ \text{info}(60,21,169) = \text{entropy}(60/250, 21/250, 169/250) \]
\[ = -(60/250) \log_2(60/250) - (21/250) \log_2(21/250) - (169/250) \log_2(169/250) \]
\[ = 0.494 + 0.3 + 0.382 \]
\[ = 1.176 \]
info(228,28,119) = entropy(228/375, 28/375, 119/375)

= -(228/375)\log_2(228/375) – (28/375)\log_2(28/375) – (119/375)\log_2(119/375)

= 0.437 + 0.280 + 0.525

= 1.242
gain (left-weight) = info (288, 49, 288) – info ([60, 21, 169], [228, 28, 119])
    = (0.5151 + 0.288 + 0.5151) – (250/625)*1.176 – (375/625)*1.242
    = 1.318 – 0.4704 – 0.7452
    = 0.103

You can repeat the same procedure to get the information gain corresponding to each of the remaining attributes.