## Q1:

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



## The procedure of the quantitative justification (referring Figure 4.1):








You can repeat the same procedure to generate the other statistics.

| attribute | category | num | L | B | R |
| :---: | :---: | :---: | :---: | :---: | :---: |
| left-weight | $(-\infty, 2.5]$ | 250 | $60(24 \%)$ | $21(8.4 \%)$ | $169(67.6 \%)$ |
|  | $(2.5, \infty)$ | 375 | $228(60.8 \%)$ | $28(7.5 \%)$ | $119(31.7 \%)$ |
| left-distance | $(-\infty, 2.5]$ | 250 | $60(24 \%)$ | $21(8.4 \%)$ | $169(67.6 \%)$ |
|  | $(2.5, \infty)$ | 375 | $228(60.8 \%)$ | $28(7.5 \%)$ | $119(31.7 \%)$ |
| right-weight | $(-\infty, 2.5]$ | 250 | $169(67.6 \%)$ | $21(8.4 \%)$ | $60(24 \%)$ |
|  | $(2.5, \infty)$ | 375 | $119(31.7 \%)$ | $28(7.5 \%)$ | $228(60.8 \%)$ |
| right-distance | $(-\infty, 2.5]$ | 250 | $169(67.6 \%)$ | $21(8.4 \%)$ | $60(24 \%)$ |
|  | $(2.5, \infty)$ | 375 | $119(31.7 \%)$ | $28(7.5 \%)$ | $228(60.8 \%)$ |

$\begin{array}{lll}\text { 1. left-weight } & (-\infty, 2.5]=>R & (2.5, \infty)=>L \\ \text { 2. left-distance } & (-\infty, 2.5]=>R & (2.5, \infty)=>L \\ \text { 3. right-weight } & (-\infty, 2.5]=>L & (2.5, \infty)=>R \\ \text { 4. right-distance } & (-\infty, 2.5]=>L & (2.5, \infty)=>R\end{array}$

## 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).




## 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.


```
info(60,21,169) = 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 $(228 / 375,28 / 375,119 / 375)$

$$
\begin{aligned}
& =-(228 / 375) \log _{2}(228 / 375)-(28 / 375) \log _{2}(28 / 375)-(119 / 375) \log _{2}(119 / 375) \\
& =0.437+0.280+0.525 \\
& =1.242
\end{aligned}
$$

gain $($ left-weight $)=$ info $(288,49,288)-$ info $([60,21,169],[228,28,119])$

$$
\begin{aligned}
& =(0.5151+0.288+0.5151)-(250 / 625) \star 1.176-(375 / 625) \star 1.242 \\
& =1.318-0.4704-0.7452 \\
& =0.103
\end{aligned}
$$

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

