

Consider the data in Table above (end of chapter 6 or 8). The target variable is salary. Start by discretizing salary and age as follows:

Less than $35,000 Level 1

$35,000 to less than $45,000 Level 2

$45,000 to less than $55,000 Level 3

Above $55,000 Level 4

0 – 30 <= 30

31 - 40 <= 40

Above 40 <= 50

6.1 Construct a classification and regression tree to classify salary based on the other variables only one split level.

Hint: you may want to set up the excel file like the following

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Split | PL | PR | Level | P( j |tL ) | P( j |tR) | 2PL PR | Q(s|t) | Φ(s|t) |
| 1 | 0.273 | 0.727 | L1 | 0.333 | 0.125 | 0.397 | 0.583 | 0.231 |
|  |  |  | L2 | 0.333 | 0.250 |  |  |  |
|  |  |  | L3 | 0.333 | 0.375 |  |  |  |
|  |  |  | L4 | 0.000 | 0.250 |  |  |  |
| 2 |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |

6.2 Use Excel to construct a C4.5 decision tree to classify salary based on the other variables