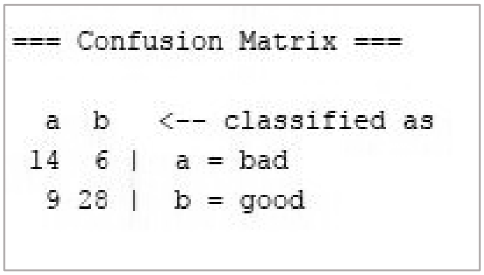
Problem 2

Part A: pruned



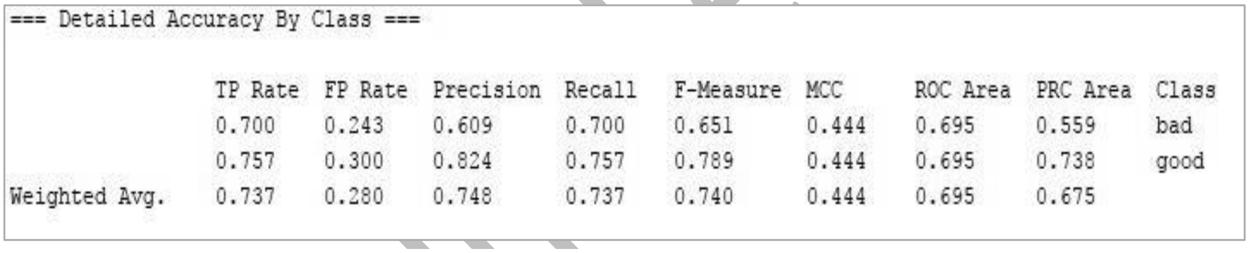
```
=== Confusion Matrix ===

  a  b   <-- classified as
 14  6 |  a = bad
  9 28 |  b = good
```

Figure 1: confusion matrix

TP = 14, TN = 28, FP = 9, FN =6

Accuracy = $\frac{TP+TN}{TP+TN+FP+FN} = \frac{14+28}{14+28+9+6} = 0.7368$



```
=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.700    0.243    0.609      0.700    0.651      0.444    0.695     0.559     bad
                 0.757    0.300    0.824      0.757    0.789      0.444    0.695     0.738     good
Weighted Avg.    0.737    0.280    0.748      0.737    0.740      0.444    0.695     0.675
```

Figure 2: Detailed accuracy by class

if we consider "bad" as class of interest, based on confusion matrix shown in figure 1, desired parameters are calculated as follows:

Recall = TPR =$\frac{TP}{TP+FN} = \frac{14}{14+6} = 0.700.$

Precision = $\frac{TP}{TP+FP} = \frac{14}{14+9} = 0.609$.

F1-Measure = $2 * \frac{Precision * Recall}{Precision+Recall} = 2 * \frac{0.609 * 0.700}{.0609 + 0.700} = 0.651.$

As you can see, the parameters obtained are equal to values shown in figure 2.

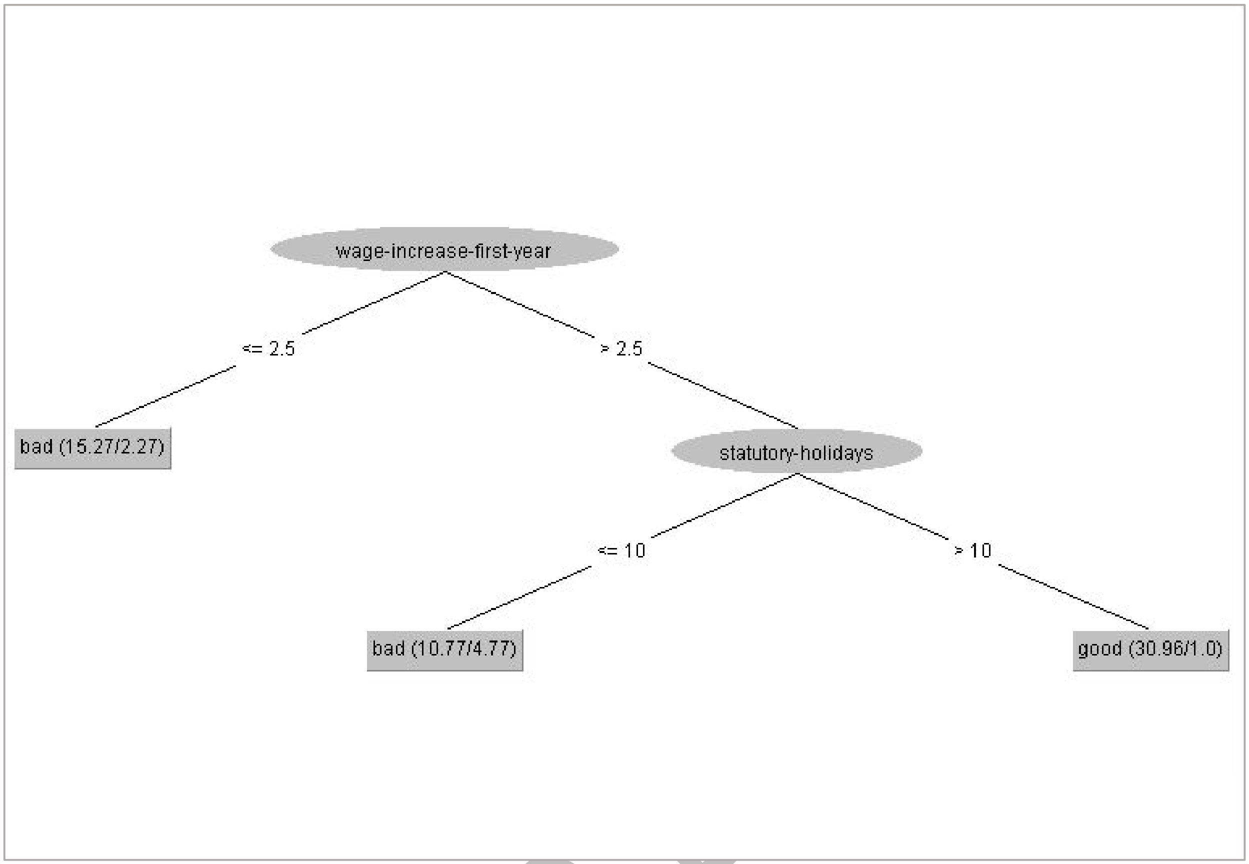
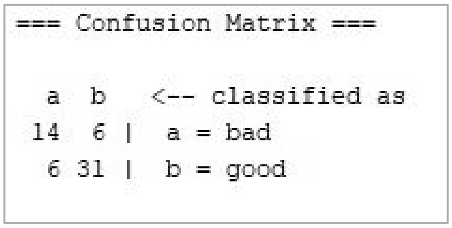


Figure 3: Decision tree

Based on decision tree shown in figure 3, first we consider value of “wage-increase-first-year”. Because its value is greater than 3, we go to the right of the node. After that, we examine value of “statutory-holiday”. Because its value is greater than 12, we go to the right of the node. Finally, we conclude this data in belongs to **good** class.

Part B: unpruned



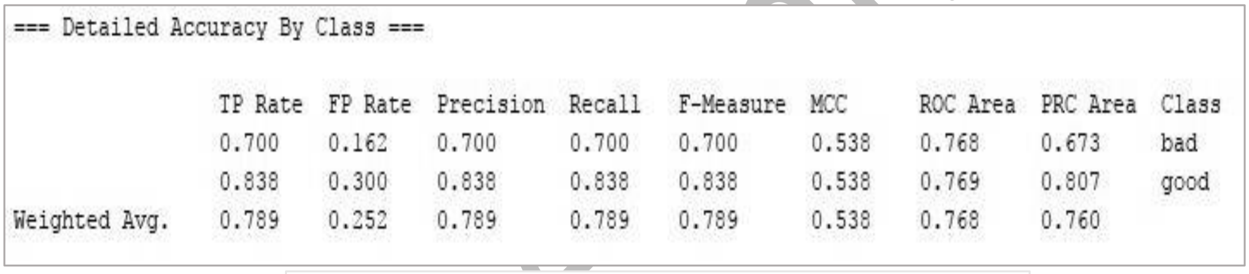
```
=== Confusion Matrix ===

  a  b   <-- classified as
 14  6 |  a = bad
  6 31 |  b = good
```

Figure 4: confusion matrix

TP = 14, TN = 31, FP = 6, FN =6

Accuracy = $\frac{TP+TN}{TP+TN+FP+FN} = \frac{14+31}{14+31+6+16} = 0.7895$



```
=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.700    0.162    0.700      0.700    0.700      0.538    0.768     0.673     bad
                 0.838    0.300    0.838      0.838    0.838      0.538    0.769     0.807     good
Weighted Avg.    0.789    0.252    0.789      0.789    0.789      0.538    0.768     0.760
```

Figure 4: Detailed accuracy by class

if we consider “bad” as class of interest, based on confusion matrix shown in figure 3, desired parameters are calculated as follows:

Recall = TPR =$\frac{TP}{TP+FN} = \frac{14}{14+6} = 0.700.$

Precision = $\frac{TP}{TP+FP} = \frac{14}{14+6} = 0.700$.

F1-Measure = $2 * \frac{Precision * Recall}{Precision + Recall} = 2 * \frac{0.700 * 0.700}{0.700 + 0.700} = 0.7000.$

As you can see, the parameters obtained are equal to values shown in figure 4.

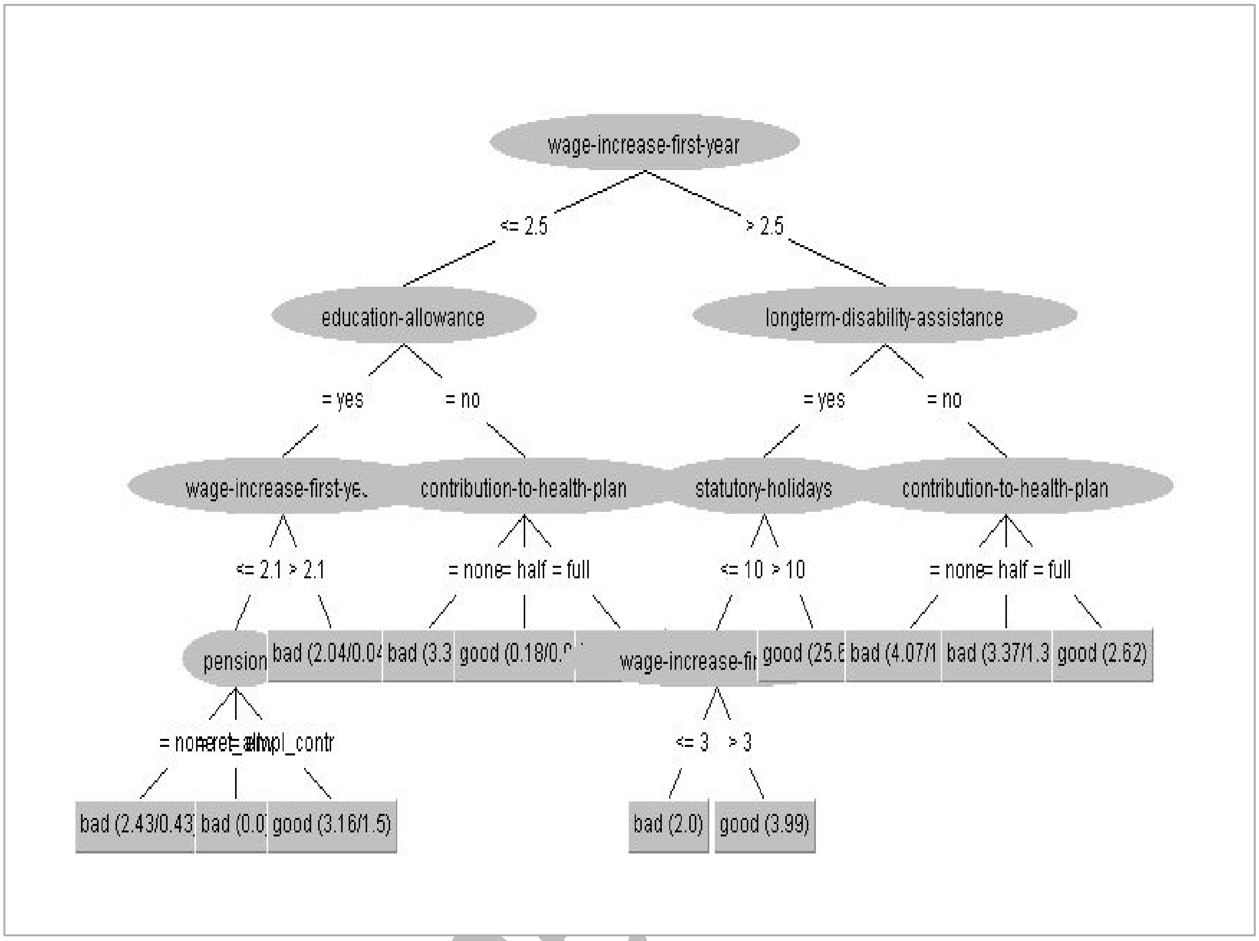


Figure 6: Decision tree

Based on decision tree shown in figure 3, first we consider value of “wage-increase-first-year”. Because its value is greater than 3, we go to the right of the node. After that, we examine value of “longterm-disability-assistance”. Because its yes, we go to the left of the node. After that, we examine value of “statutory-holiday”. Because its value is greater than 12, we go to the right of the node. Finally, we conclude this data in belongs to **good** class.

The difference between the tree taught in this section and the previous section is obtained due to pruning. In the Part B, the depth of tree is larger than Part A and model in Part B prone to overfitting because the pruning has not been done.