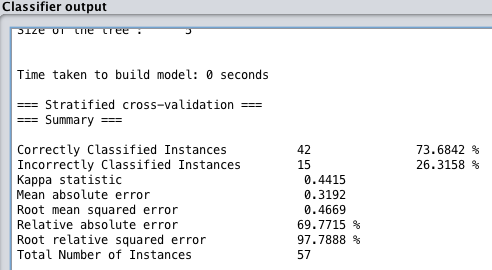
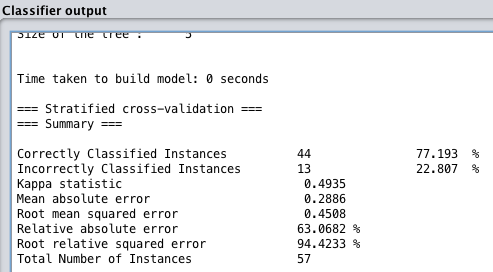
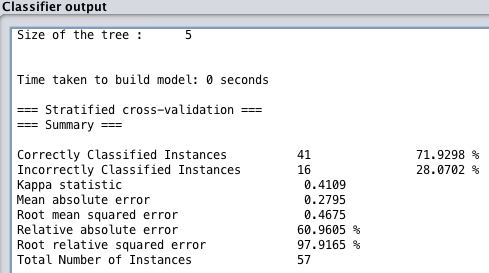
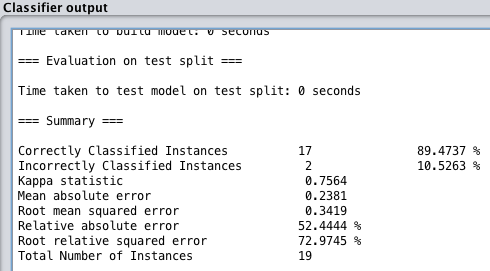
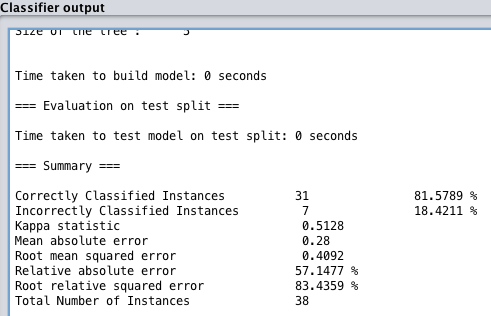
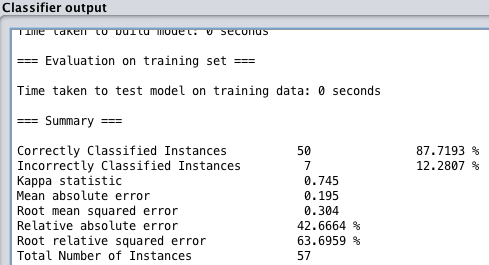
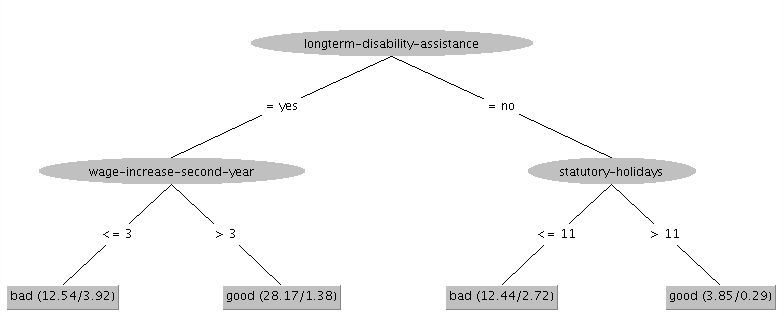
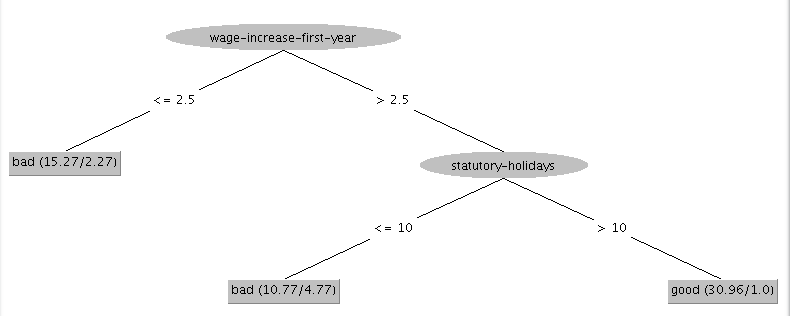
Lab 2: Exploring the Explorer

In this lab, you will practice using the *Weka Explorer*.

1. Start the *Weka Explorer and load the* ***labour negotiations*** *data (see lab 1).*
2. Run the J48 classifier several times, once for each of the following test options, and for each record the percentage of correctly classified instances. The purpose of this is to give you an initial understanding of the test output data. **Cross-validation** with n folds means that the training data is divided in to n portions, and the algorithm is trained with n-1 parts as the training data, while one part is used to test the resulting model. This process is done in the n ways possible and the results are averaged. Theory and practice suggest that n = 10 folds is best in general. Testing by **using the training set** gives the most optimistic accuracy evaluation, since the same data is used both for training and testing. **Percentage split** 66% means that two thirds of the data are used for training, and one third is held back for testing. In your report for this question, show the resulting accuracy evaluations, and comment on the range of values.
   1. **Cross-validation folds 10**Correctly Classified Instances 42 73.6842 %  
      
   2. Cross-validation folds 5  
      Correctly Classified Instances 44 77.193 %  
      
   3. Cross-validation folds 2  
      Correctly Classified Instances 41 71.9298 %  
      
   4. Percentage split 66%  
      Correctly Classified Instances 17 89.4737 %  
      
   5. Percentage split 33%  
      Correctly Classified Instances 31 81.5789 %  
      
   6. Use training set  
      Correctly Classified Instances 50 87.7193 %  
      
3. Go back to the **preprocess** tab, and under **Filter**, click on **Choose**. As an example of using a filter, choose the unsupervised attribute filter **Remove**. Click the Remove line that appears next to Choose to get an object editor for this filter. Click on **More** to read how to use it. Remove all even-numbered attributes, and run the J48 classifier again with cross-validation folds 10. Compare the tree and accuracy with the tree and accuracy you had when all attributes were considered.   
   REMOVED  
   Correctly Classified Instances 47 82.4561 %  
     
   ORIGINAL  
   Correctly Classified Instances 42 73.6842 %  
   
4. Load the iris data using the Preprocess panel. Evaluate C4.5 algorithm (J48) on this data using (a) the training set and (b) cross-validation. What is the estimated percentage of correct classifications for (a) and (b)? Which estimate is more realistic?  
   (a) training set Correctly Classified Instances 147 98 %  
   (b) cross-validation Correctly Classified Instances 144 96 %  
     
   cross-validation looks more realistic since it validates the data using a subset that was not used to train the model.
5. Right click on the trees J48 entry in the result list and choose Visualize classifier errors. The little crosses indicate instances correctly classified, and the squares represent the instances incorrectly classified. What can you say about the location of the errors?  
   The squares are relatively close to the crosses of the same colour but maybe a bit further to the edge.
6. Load the **CPU performance dataset** (with vendor). The attributes are vendor, cycle time, memory min and max, cache, channels min and max, and class, where class is a numeric measurement of performance. Since the data is primarily numeric, linear regression seems to be the best approach to fitting the data. Run the M5P tree classifier with 10-fold cross validation. You will see that it makes a decision stump where it divides the data based on MMAX into two groups LM1 and LM2. Then the two groups are each given a linear regression analysis. The “vendor=a,b,c” term has the value 1 if vendor is a or b or c, and 0 otherwise. Note the correlation coefficient. Run the classifier simple Linear Regression, listed under Functions, and compare its accuracy with that of M5'. Why do you suppose one has less error than the other?  
   a M5P tree classifier went through an decision tree using MMAX before linear regression to train the model.
7. Hand in a simple text file with your answers to the above questions.