CSC4008 Assignment3

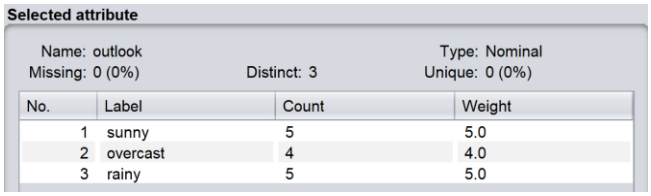
118010045 Cui Yuncong

1. **Load weather.nominal.arff, and answer the following questions. (10')**
   1) **How many instances, attributes in this dataset?**

      Instances: 14
      Attributes: 5

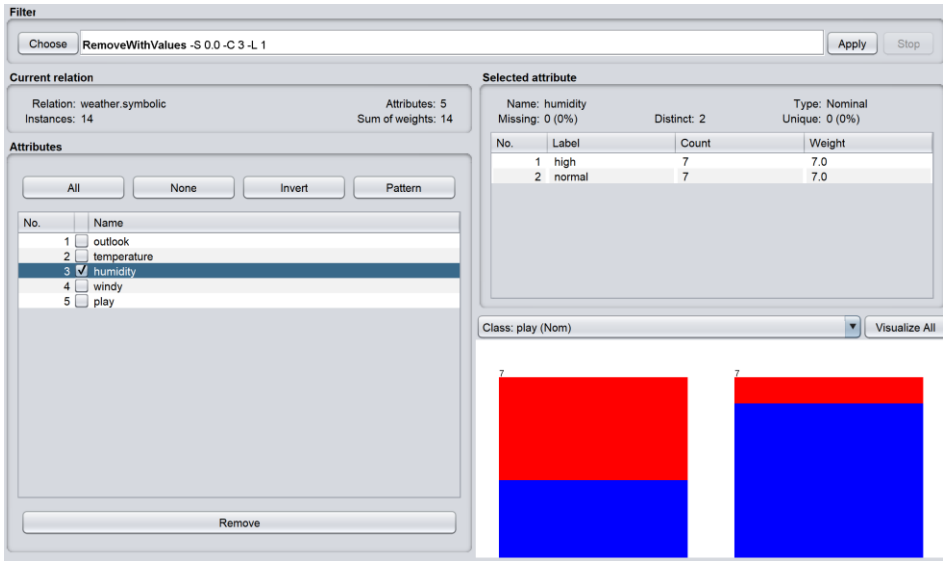
   2) **The number of the distinct label of the attribute: temperature.**

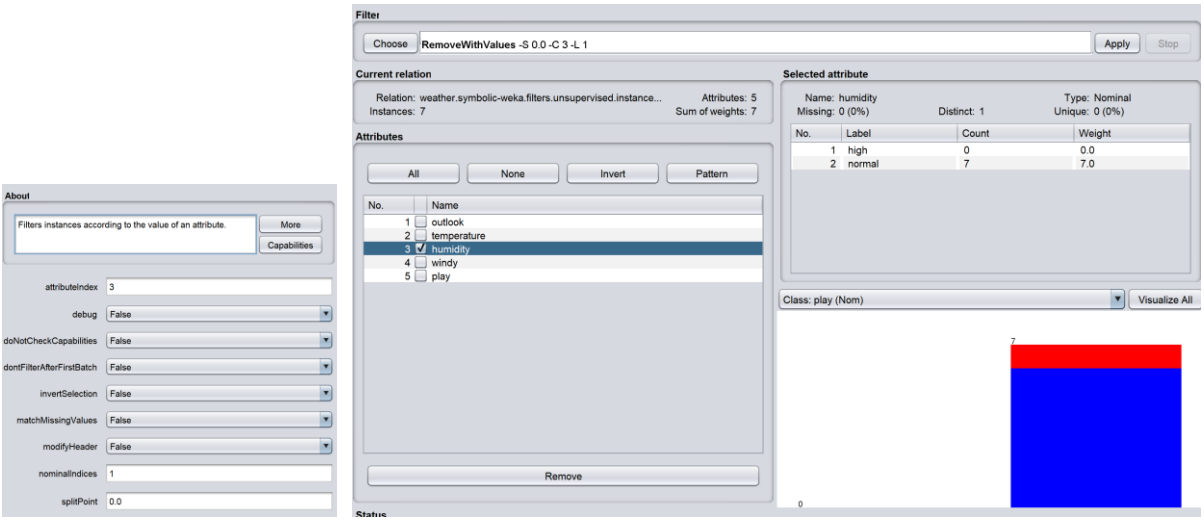
      Number: 3



   3) **Use the filter to remove the value: high in the humidity attribute from the dataset.**

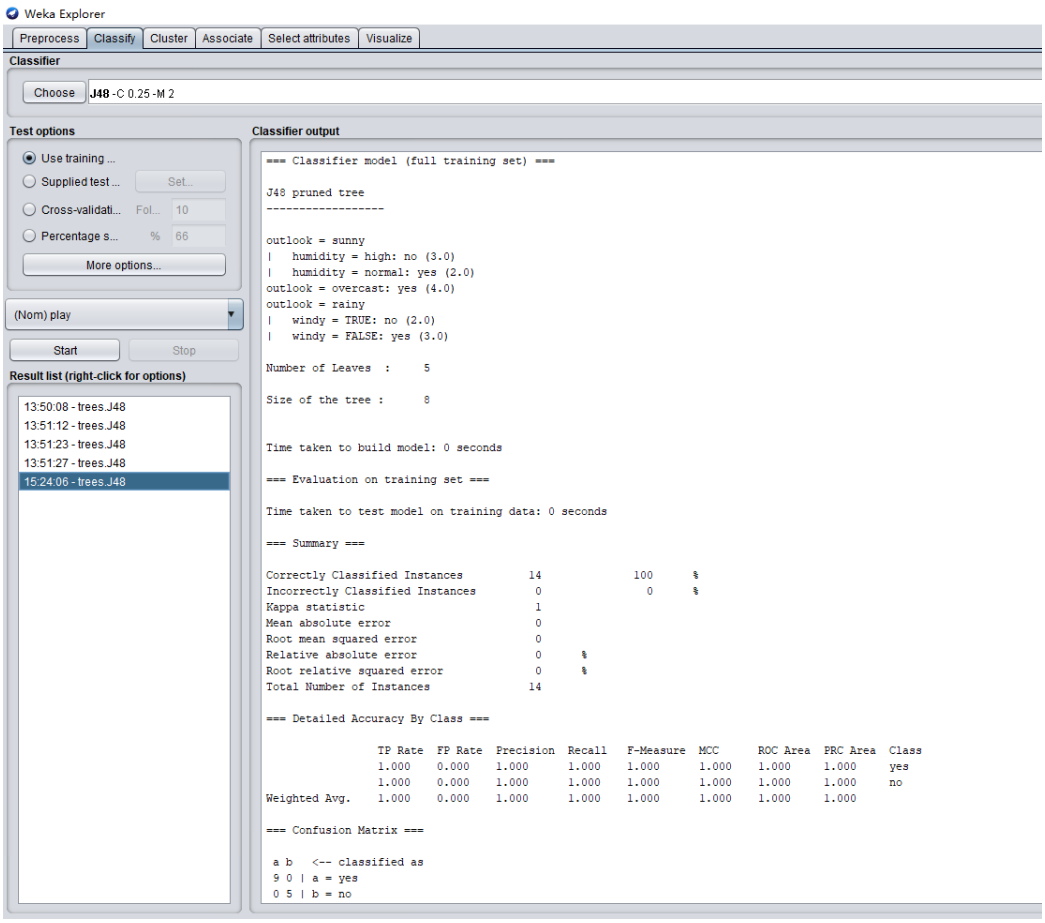
      Choose the RemoveWithValues filter.
      Fill in the attributeIndex with 3, fill in nomialIndex with 1.
      Apply.

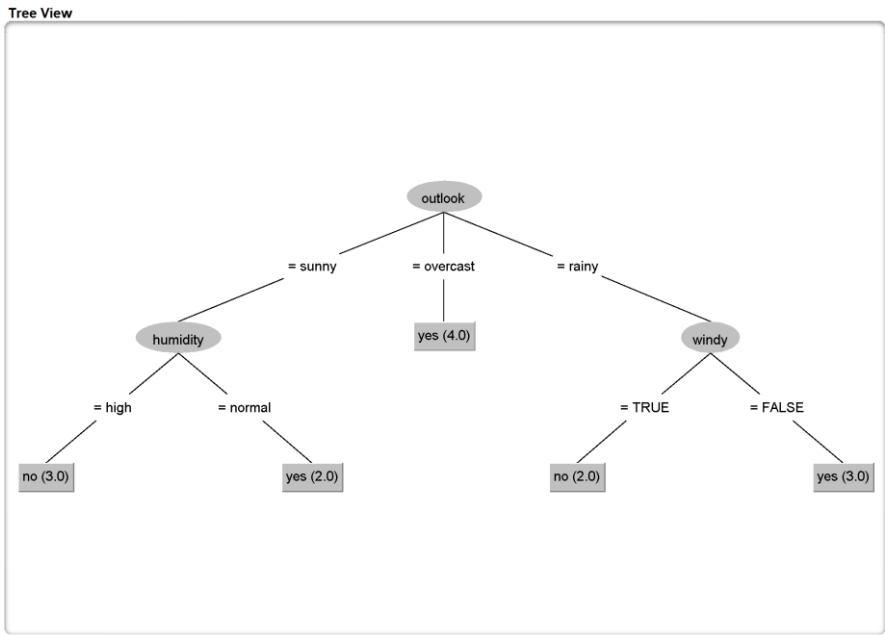




4) Please build a J48 decision tree to classify the dataset and visualize the tree in a figure.

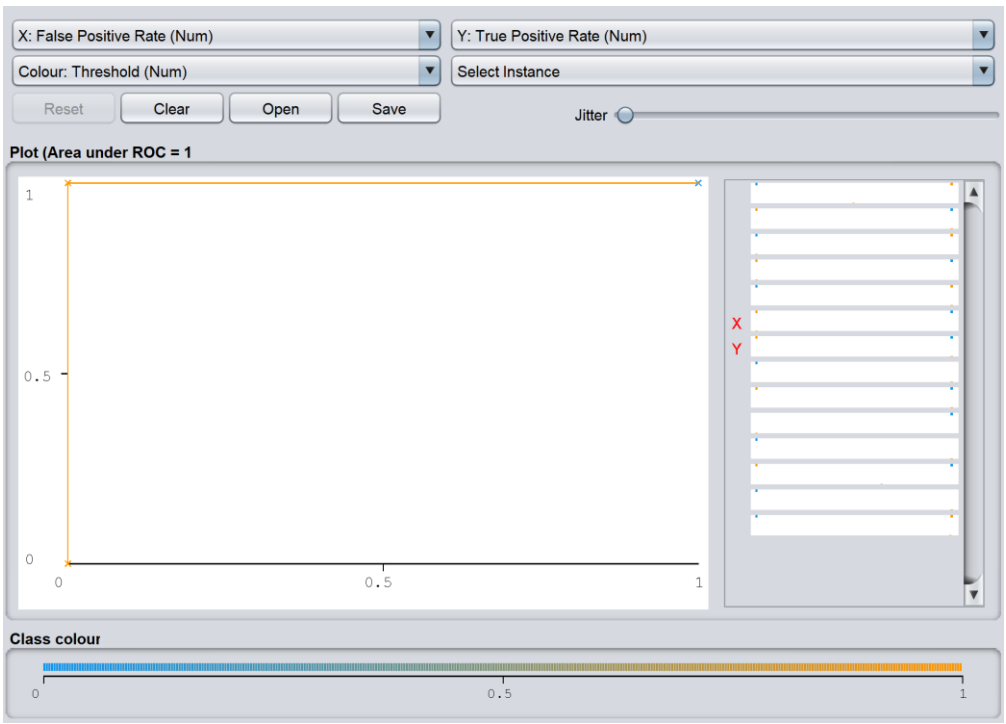
Choose the J48 decision tree.





5) Analyze the performance of J48 (including but not limited to TPR, FPR, ROC, AUC).

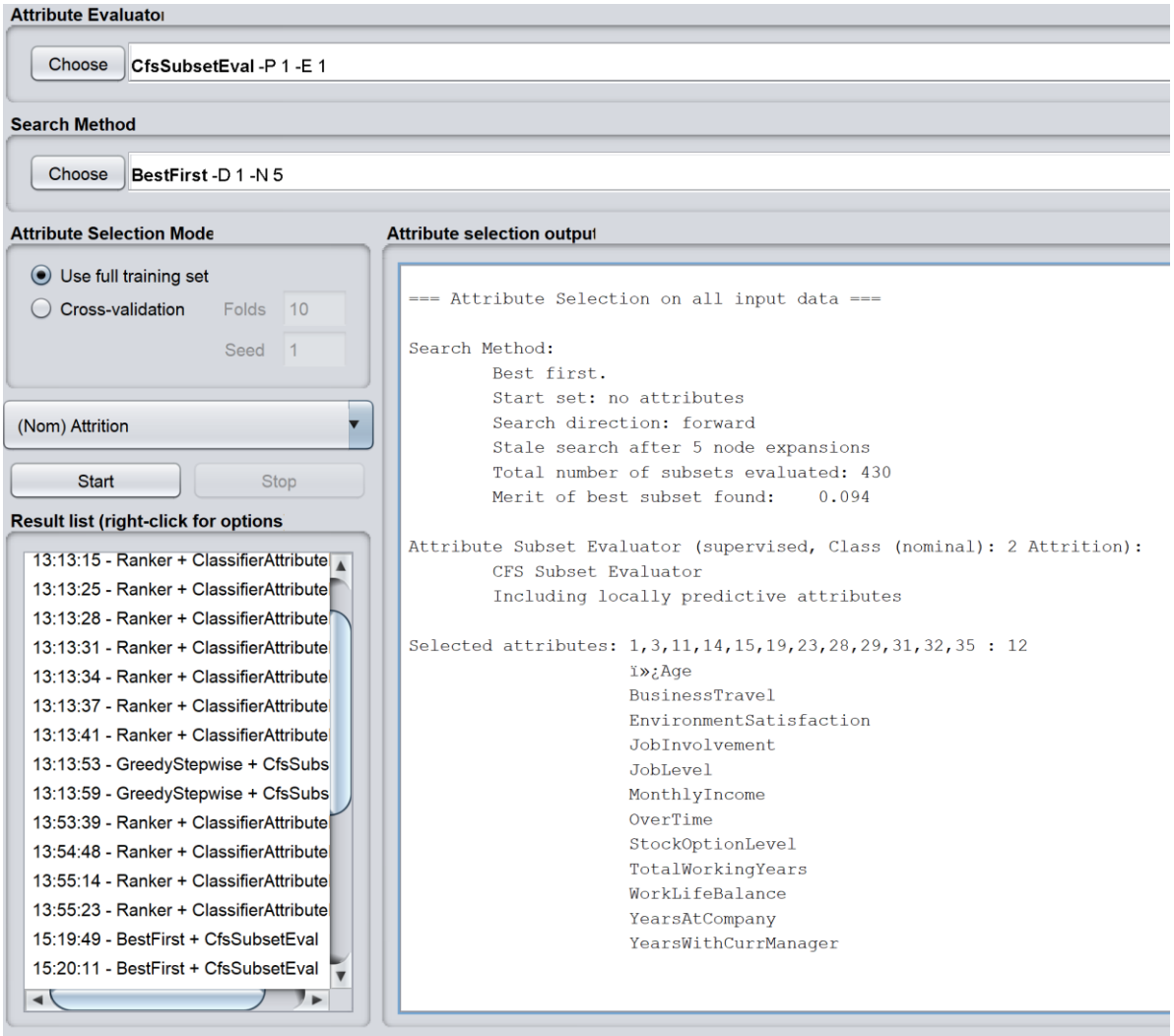
N = 14 TP = 9 FN = 0 FP = 0 TN = 5
TPR = TP/(TP+FN) = 1 FPR = FP/(FP+TN) = 0
AUC = 1
The higher the TPR is, the better the classifier. The lower the FPR is, the better the classifier. TPR = 1 and FPR = 0 means J48 is the perfect classifier for this dataset. The AUC value is the area covered by the ROC curve. Obviously, the larger the AUC, the better the classifier. AUC = 1 means J48 is the perfect classifier for this dataset.



2. **Load HR-Employee-Attrition.csv dataset. (10')**
   1) **Implement attribute selection.**

Evaluator: CfsSubsetEval
Search Method: BestFirst
N = 12

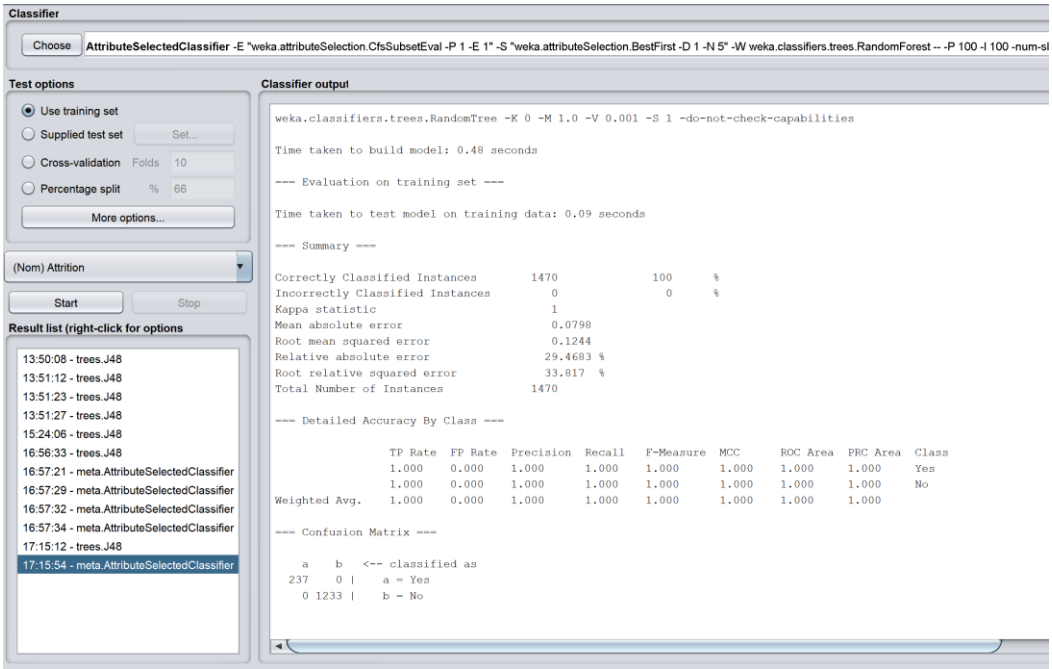
These attributes are Age, BusinessTravel, EnvironmentSatisfaction, JobInvolvement, JobLevel, MonthlyIncome, OverTime, StockOptionLevel, TotalWorkingYears, WorkLifeBalance, YearsAtCompany, YearsWithCurrManager



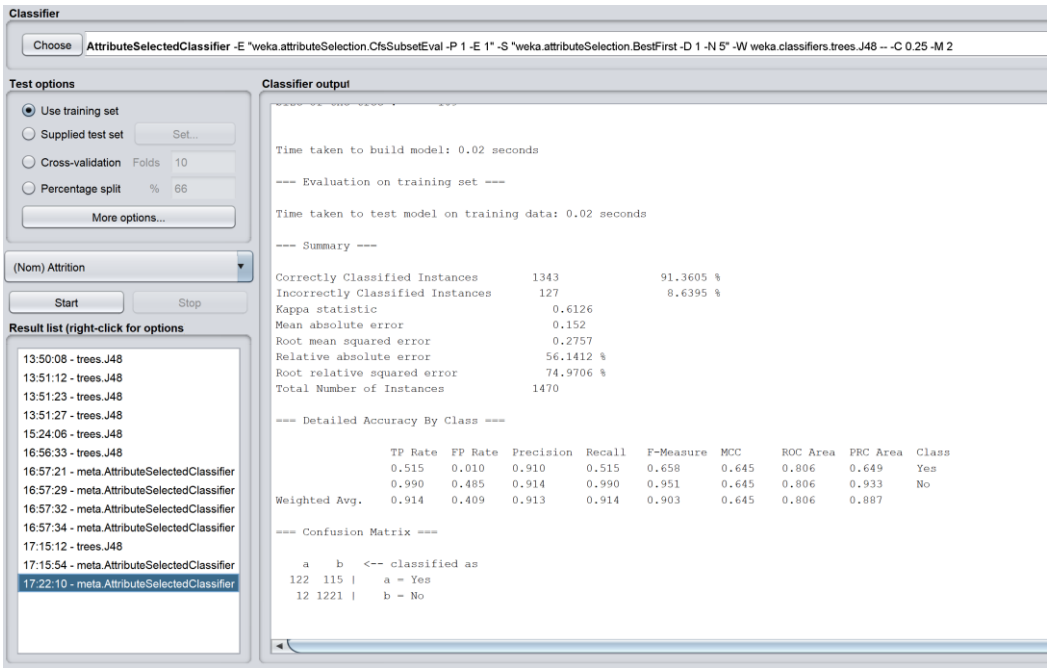
2) Using these N attributes to build a classifier.

The RandomForest is chosen among these 3 classifiers (RandomForest, J48, NaiveBayes). Because its TPR = 1 and FPR = 0, which means the accuracy is 100%.

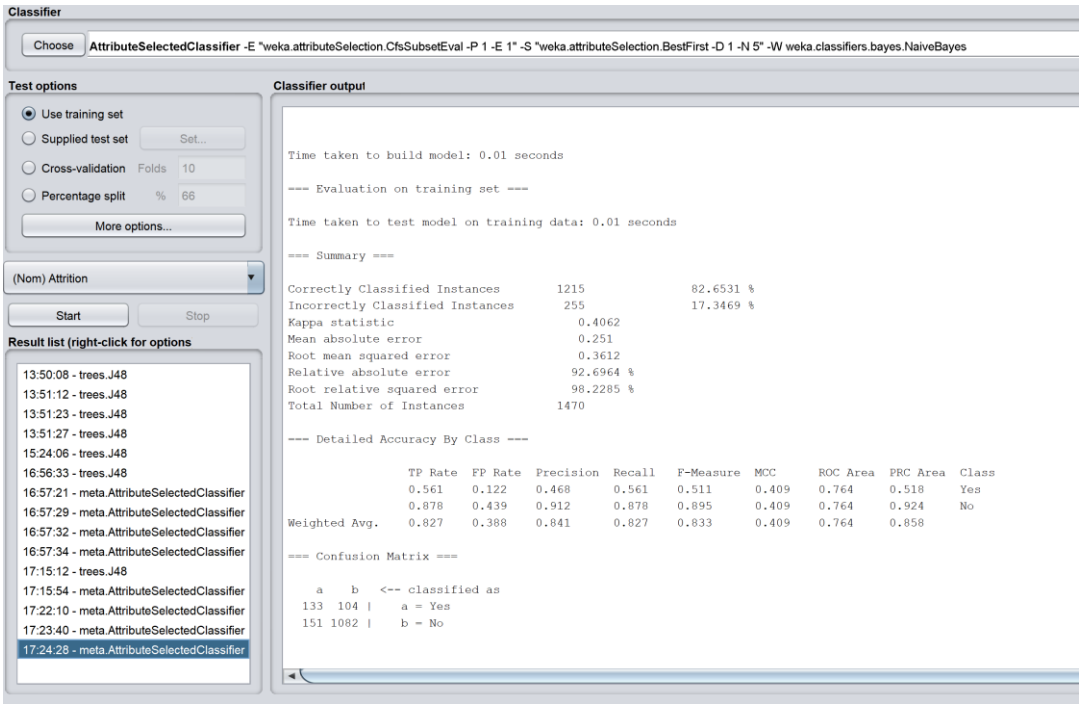
Classifier: RandomForest



Classifier: J48



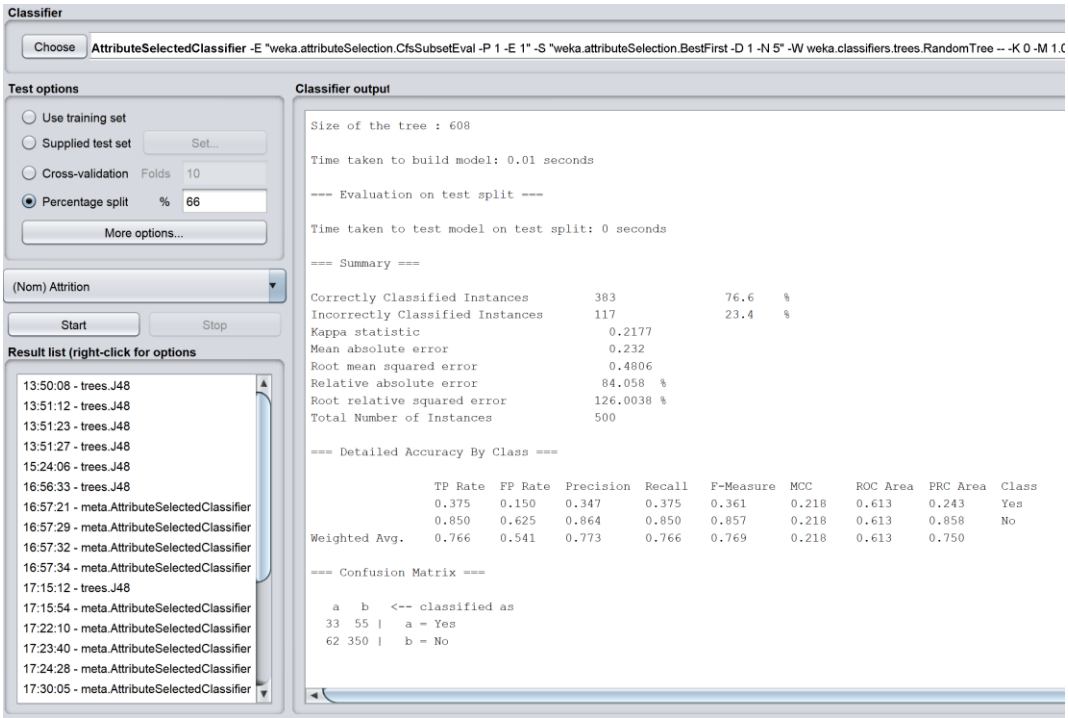
Classifier: NaiveBayes



3) Compare different test options (Cross-validation and Percentage split), and explain which one is the best.

The cross-validation is the best. The method, cross validation, takes full advantage of all the samples. The percentage split only uses part of the dataset. As a result, the accuracy of cross-validation is better than the percentage split.

Percentage split (66%):



Cross-validation (10 folds):