The rightmost node resulting from this split contains 167 instances associated with good credit risk. Because most of the instances have the same value of the dependent variable (creditability), this node is termed pure and will not be split further. The leftmost node in the first split contains 244 instances. This node is then further split based on the predictor value of savings or stock, resulting in two more nodes and so on.

The order of the splits—balance of current account, value of savings or stocks, and then payment of previous credits—is determined by an induction algorithm.

A tree that has only pure leaf nodes is called a pure tree, a condition that is not only unnecessary but is usually undesirable. Most trees are impure; that is, their leaf nodes contain cases with more than one outcome to avoid overfitting.

The rules for the leaf nodes in Figure E.4 are generated by following a path down the branches until a leaf node is encountered. For example,

**IF Balance of current account = no running account, no balance**
**AND Value of Savings or Stocks = no savings, less than 100 DM**
**AND Payment of previous credits = hesitant payment of previous credits**
**THEN Creditability = bad**

**Classification Matrix: CHAID Model**

The classification matrix can be computed for old cases as well as new cases. Only the classification of new cases (testing data set) allows us to assess the predictive validity of the model; the classification of old cases only provides a useful diagnostic tool to identify outliers or areas where the model seems to be less adequate.

The program computes the matrix of predicted and observed classification frequencies for the testing data set, which are displayed in a results spreadsheet, as well as a bivariate histogram, as shown in Figure E.5.

The classification matrix shows the number of cases that were correctly classified (on the diagonal of the matrix) and those that were misclassified as the other category.

In this case, the overall model could correctly predict whether the customer’s credit standing was good or bad with 62.83% accuracy. Our main goal is to reduce the proportion of bad credits. The percent of correct predictions for the bad category is 66.30%. In other words, if there are 100 bad customers, our model will correctly classify approximately 66 as bad (which is far better than the law of chance).

**COMPARATIVE ASSESSMENT OF THE MODELS (EVALUATION)**

It is good practice to experiment with a number of different methods when modeling or mining data rather than relying on a single model for final deployment. Different techniques may shed new light on a problem or confirm previous conclusions.

The gains chart provides a visual summary of the usefulness of the information provided by one or more statistical models for predicting categorical dependent variables.