**8. Classification** is a form of data analysis that extracts models describing data classes.

A classifier, or classification model, predicts categorical labels (classes).

**Numeric prediction** models continuous-valued functions. Classification and numeric prediction

are the two major types of prediction problems.

**Decision tree induction** is a top-down recursive tree induction algorithm, which uses an attribute selection measure to select the attribute tested for each non-leaf node in the tree.

**ID3**, **C4.5**, and **CART** are examples of such algorithms using different attribute selection measures. **Tree pruning** algorithms attempt to improve accuracy by removing tree branches reflecting noise in the data.

Early decision tree algorithms typically assume that the data are memory resident. Several scalable algorithms, such as **RainForest**, have been proposed for scalable tree induction.

**Na¨ıve Bayesian classification** is based on Bayes’ theorem of posterior probability. It assumes class-conditional independence—that the effect of an attribute value on a given class is independent of the values of the other attributes.

A **rule-based classifier** uses a set of IF-THEN rules for classification. Rules can be extracted from a decision tree. Rules may also be generated directly from training data using sequential covering algorithms.

A **confusion matrix** can be used to evaluate a classifier’s quality. For a two-class problem, it shows the *true positives, true negatives, false positives*, and *false negatives*.

Measures that assess a classifier’s predictive ability include **accuracy**, **sensitivity** (also known as **recall**), **specificity**, **precision**, *F*, and *F*\_. Reliance on the accuracy measure can be deceiving when the main class of interest is in the minority.

Construction and evaluation of a classifier require partitioning labeled data into a training set and a test set.

**Holdout, random sampling, cross-validation**, and **bootstrapping** are typical methods used for such partitioning.

Significance tests and ROC curves are useful tools for model selection. **Significance tests** can be used to assess whether the difference in accuracy between two classifiersis due to chance.

**ROC curves** plot the true positive rate (or sensitivity) versus thefalse positive rate (or 1􀀀*specificity*) of one or more classifiers.

**Ensemble methods** can be used to increase overall accuracy by learning and combining a series of individual (base) classifier models.

**Bagging, boosting**, and **random** **forests** are popular ensemble methods.

The **class imbalance problem** occurs when the main class of interest is represented by only a few tuples.

Strategies to address this problem include **oversampling,** **undersampling, threshold moving,** and **ensemble techniques**.