DATA MINING

Metode supervizate vs nesupervizate

- Metode nesupervizate – variabila target nu este identificata ca atare, cautandu-se patternuri si structuri printre toate variabilele (clustering, reguli de asociere)
- Cele mai multe metode de data mining sunt supervizate
  - Exista o variabila target predefinita
  - Algoritmul are mai multe exemple in care valoarea variabilei target este cunoscuta, astfel incat algoritmul invata care valori ale variabilei target sunt asociate cu anumite valori ale variabilelor predictori
- Cele mai multe metode supervizate aplica urmatoarea metodologie pentru a construi si pentru a evolua un model:

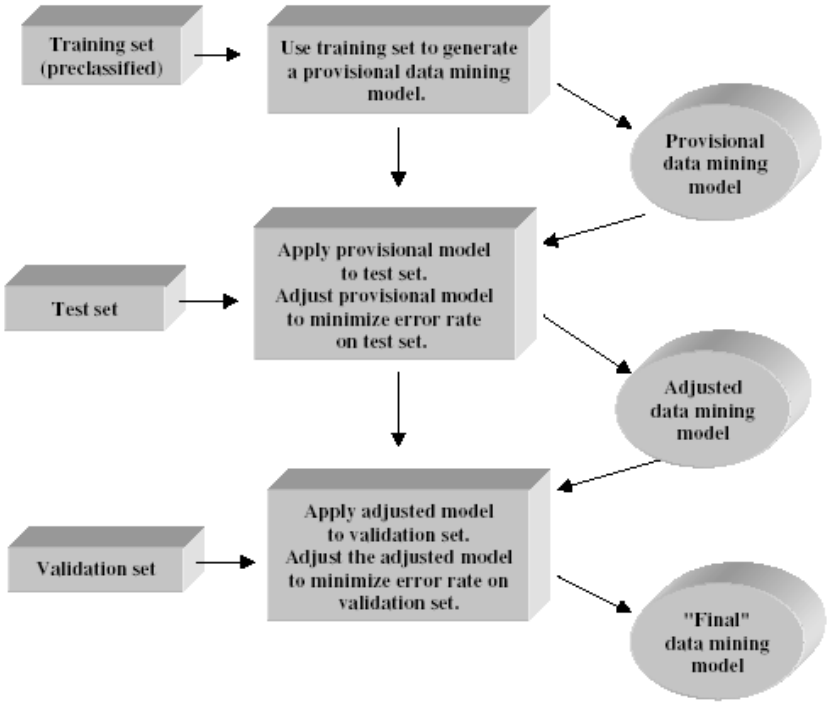

Figure 5.1 Methodology for supervised modeling.

Referinta figura: D. Larose, *Discovering Knowledge in Data. An Introduction to Data Mining*, John Wiley & Sons, 2005.

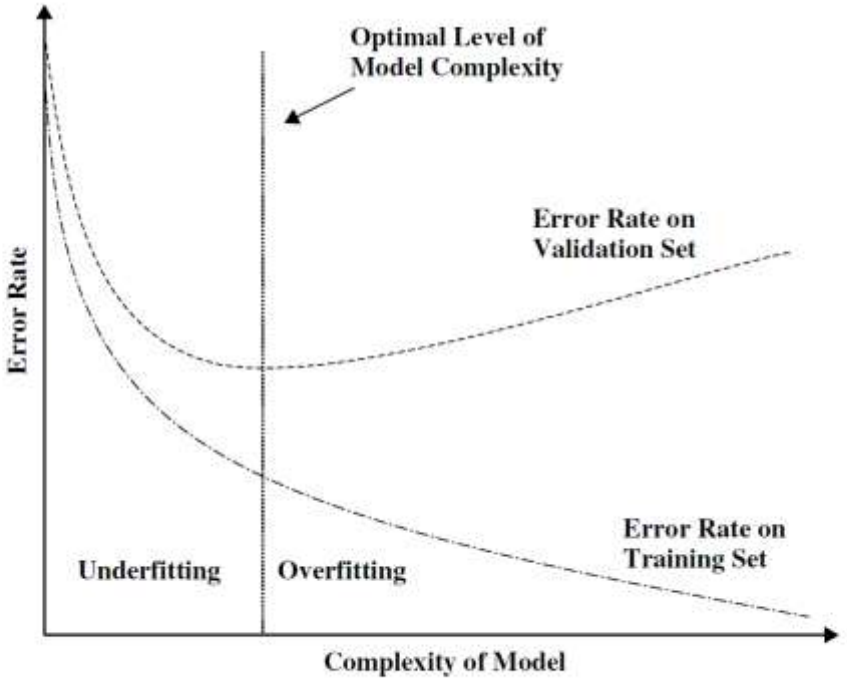

Figure 5.2 The optimal level of model complexity is at the minimum error rate on the validation set.

Referinta figura: D. Larose, *Discovering Knowledge in Data. An Introduction to Data Mining*, John Wiley & Sons, 2005.

BIAS–VARIANCE TRADE-OFF

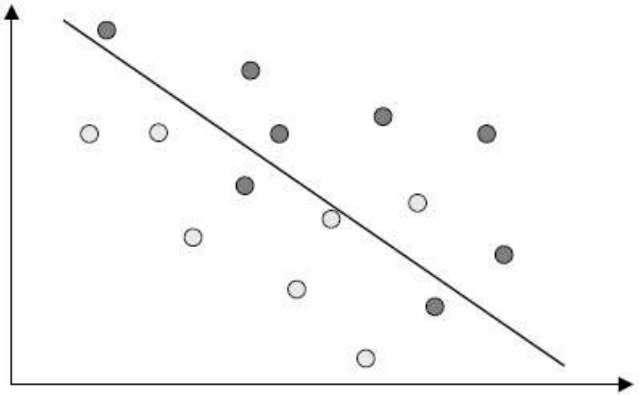

Figure 5.3 Low-complexity separator with high error rate.

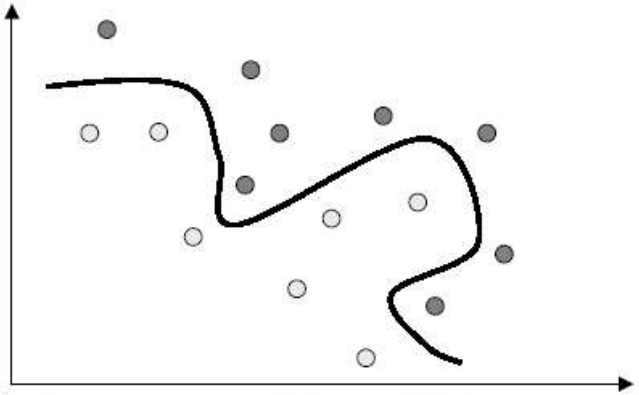

Figure 5.4 High-complexity separator with low error rate.

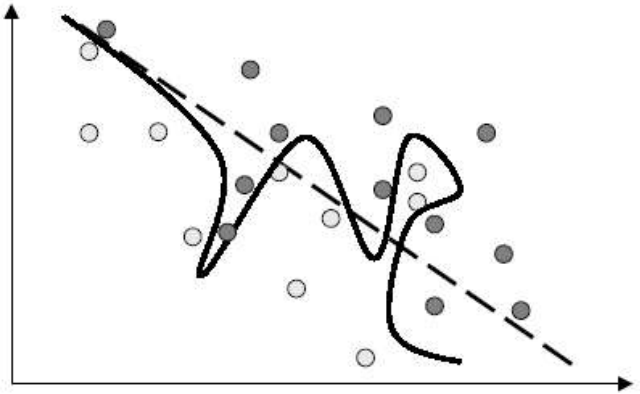

Figure 5.5 With more data: low-complexity separator need not change much; high-complexity separator needs much revision.

Referinta figura: D. Larose, *Discovering Knowledge in Data. An Introduction to Data Mining*, John Wiley & Sons, 2005.

Clasificare

- Metoda supervizata
- Exista variabila target predefinita, impartita in categorii predeterminate (ex: tipul de venit – mic, mediu, mare)
- Mai intai se examineaza setul de date care contine atat valorile predictorilor cat si ale variabilei target
- Algoritmul invata care combinatie de variabile este asociata cu o anumita valoare a variabilei target
- Metode studiate in acest curs:
  - K-nearest neighbor algorithm
  - Decision trees
  - Bayesian classification
  - Neural networks