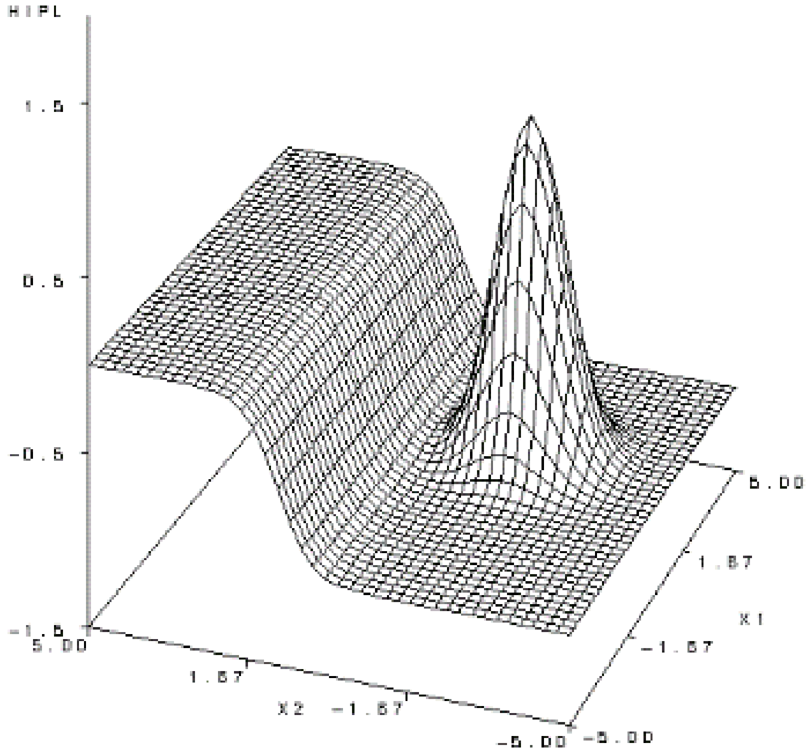
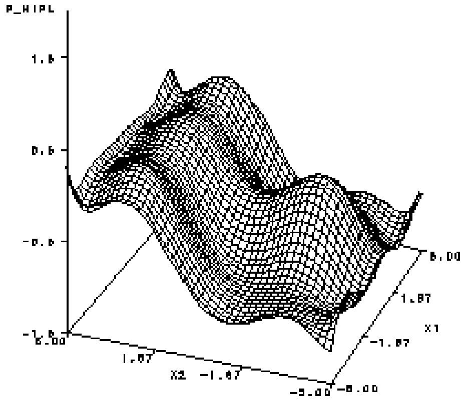
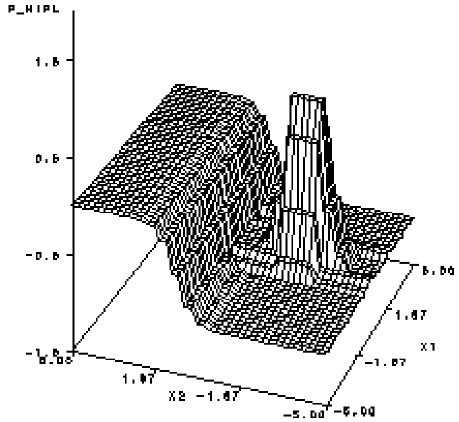
Choose the best model for your analysis



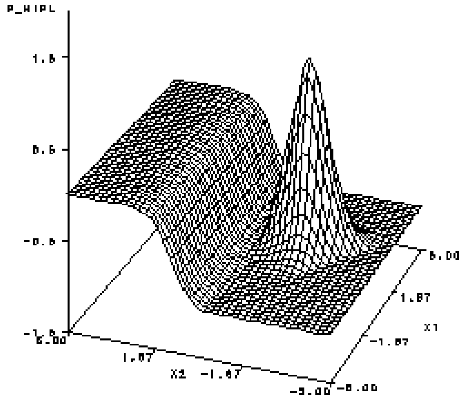
Hill and Plateau Sample Data



Traditional regression



Decision tree



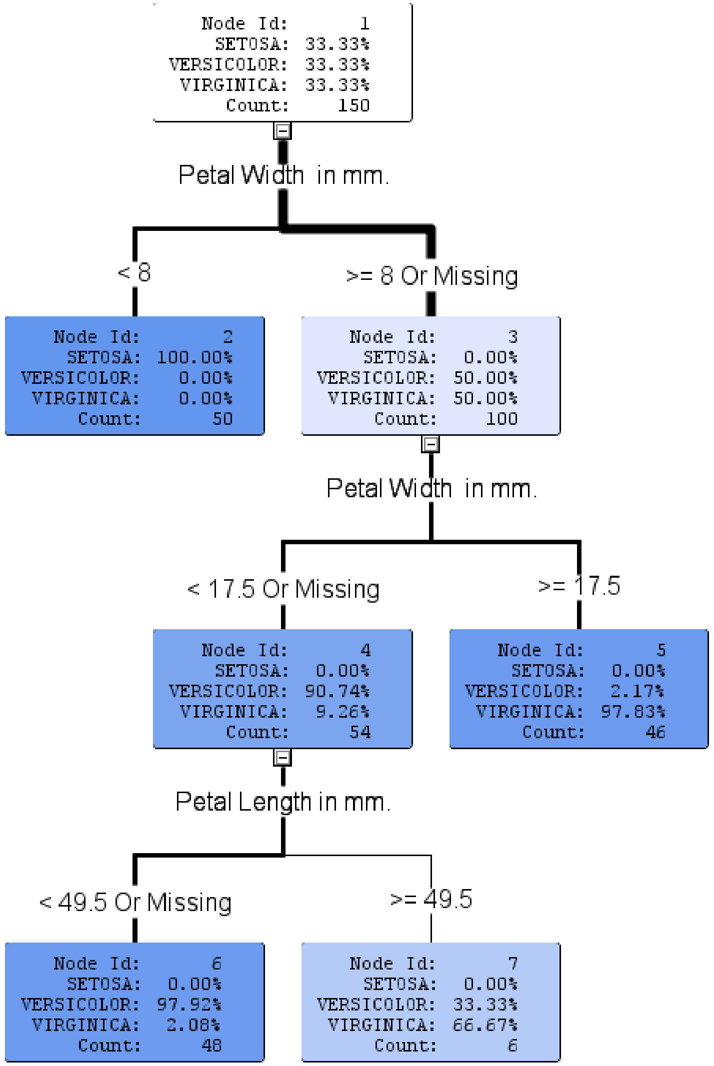
Neural network

Source: SAS Enterprise Miner Help

Decision trees

PROS	CONS
Accepted	Tendency to overfit
Understood	Many hyperparameters to tune
Interpretable	No parameters, standard errors or confidence limits
Few assumptions	Single decision trees can be unstable
Excellent for: • discontinuous, nonlinear phenomena • interactions • missing data • correlated variables • variables on different scales	Usually poor performance in pattern recognition tasks vs. neural networks

Example decision tree



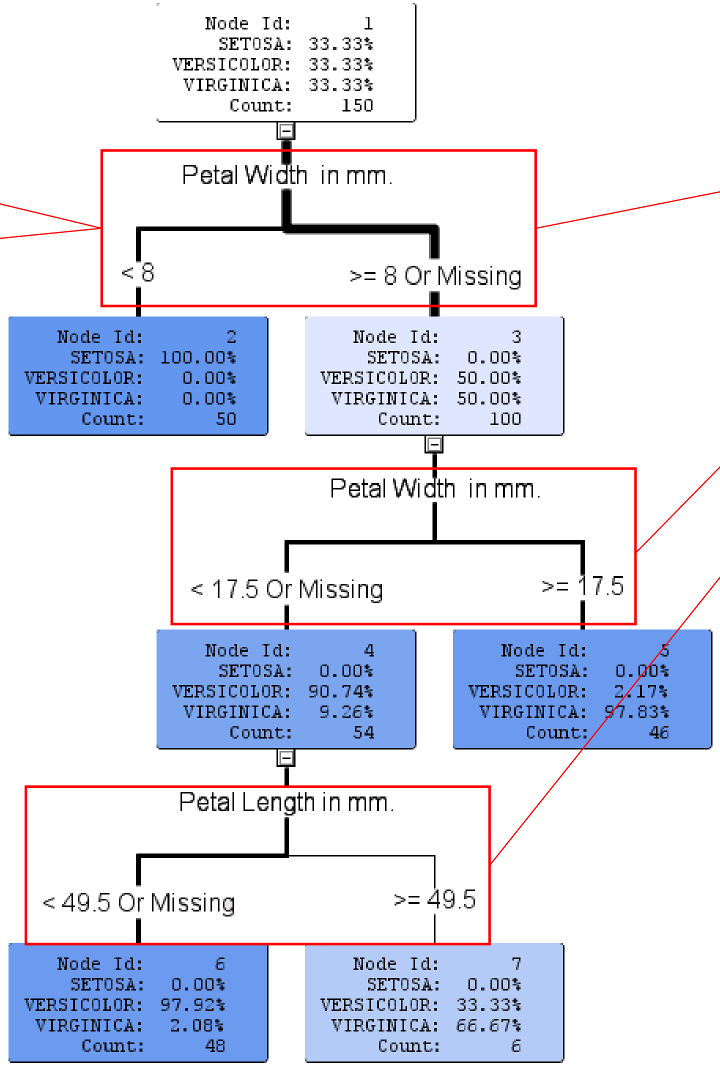
Example decision tree - basics

"If an input's petal width is less than 8 mm, then it is classified as Setosa."

This tree has binary splits, but you can split an arbitrary number of ways

Depth of 3

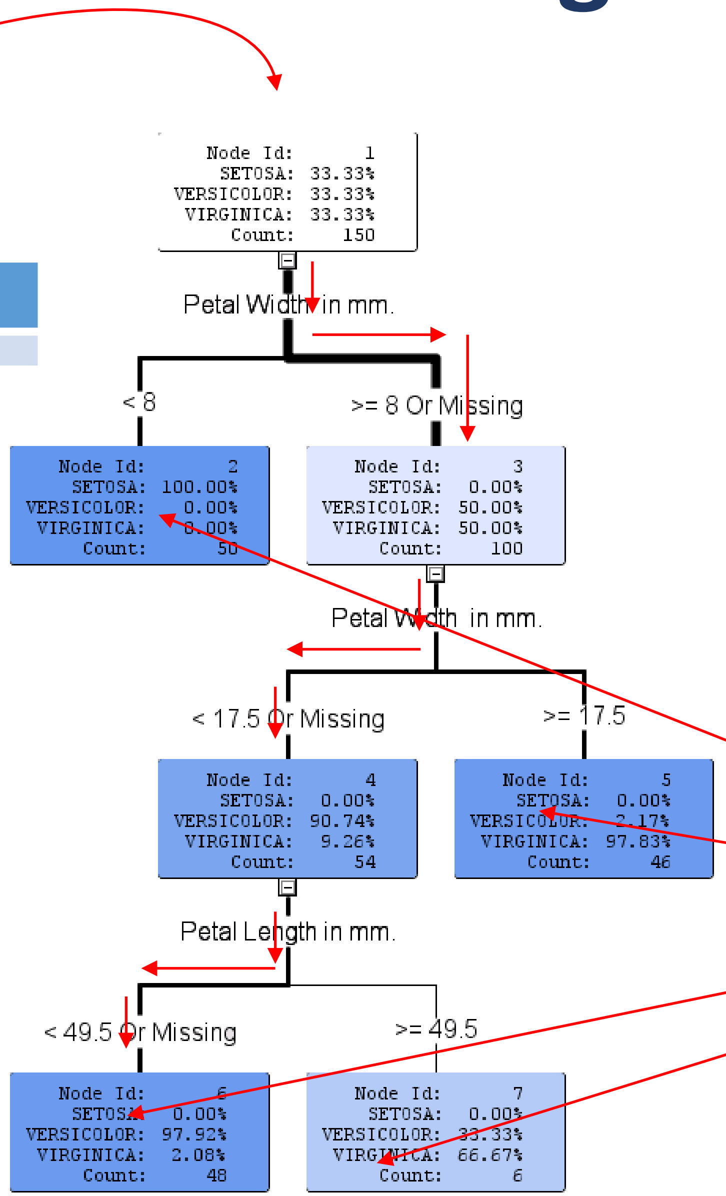
Splitting rules determined by accuracy increases or purity criterion



Example decision tree – scoring a new record

New record

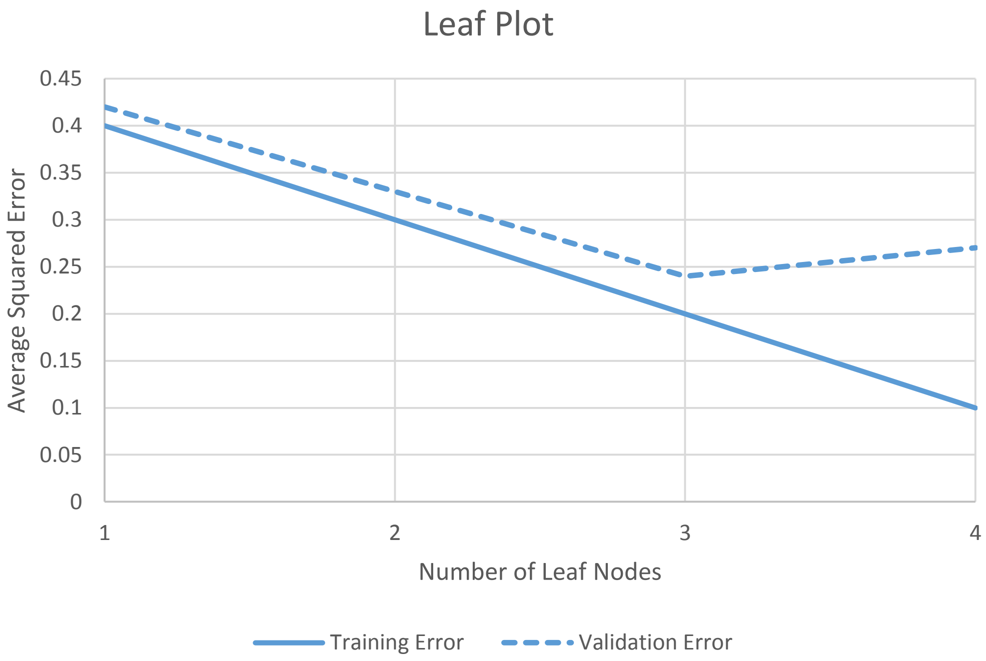
Petal Length (mm)	Petal Width (mm)	Sepal Length (mm)	Sepal Width (mm)
20	15	14	18

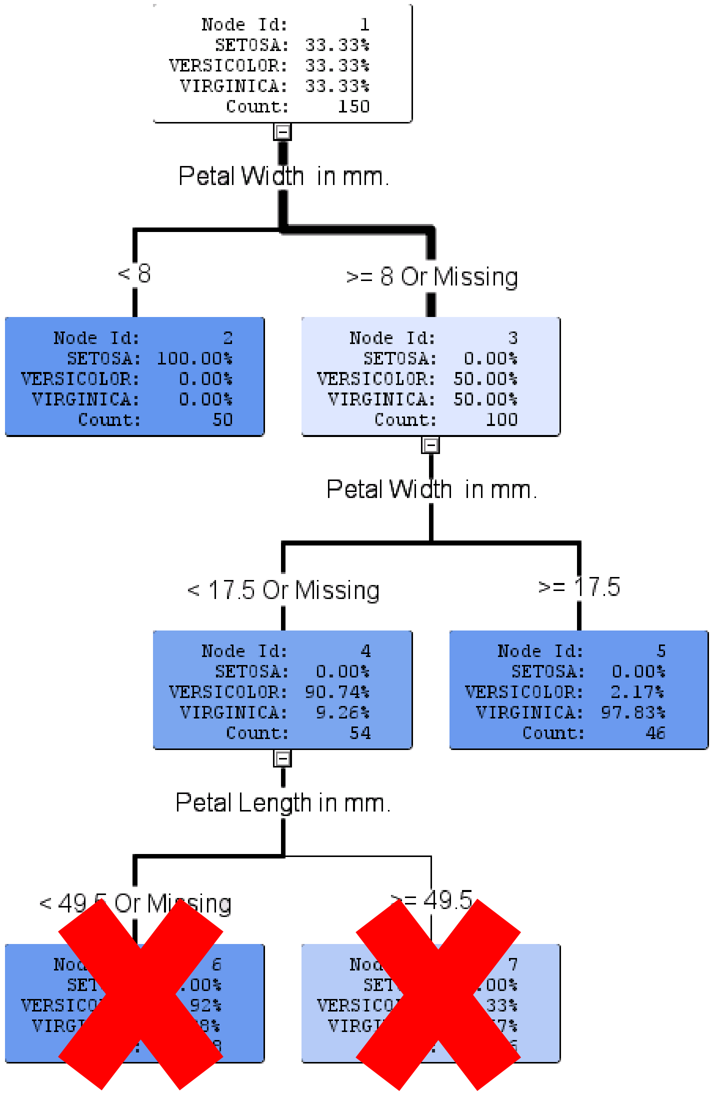


4 leaf nodes – these define the predicted values

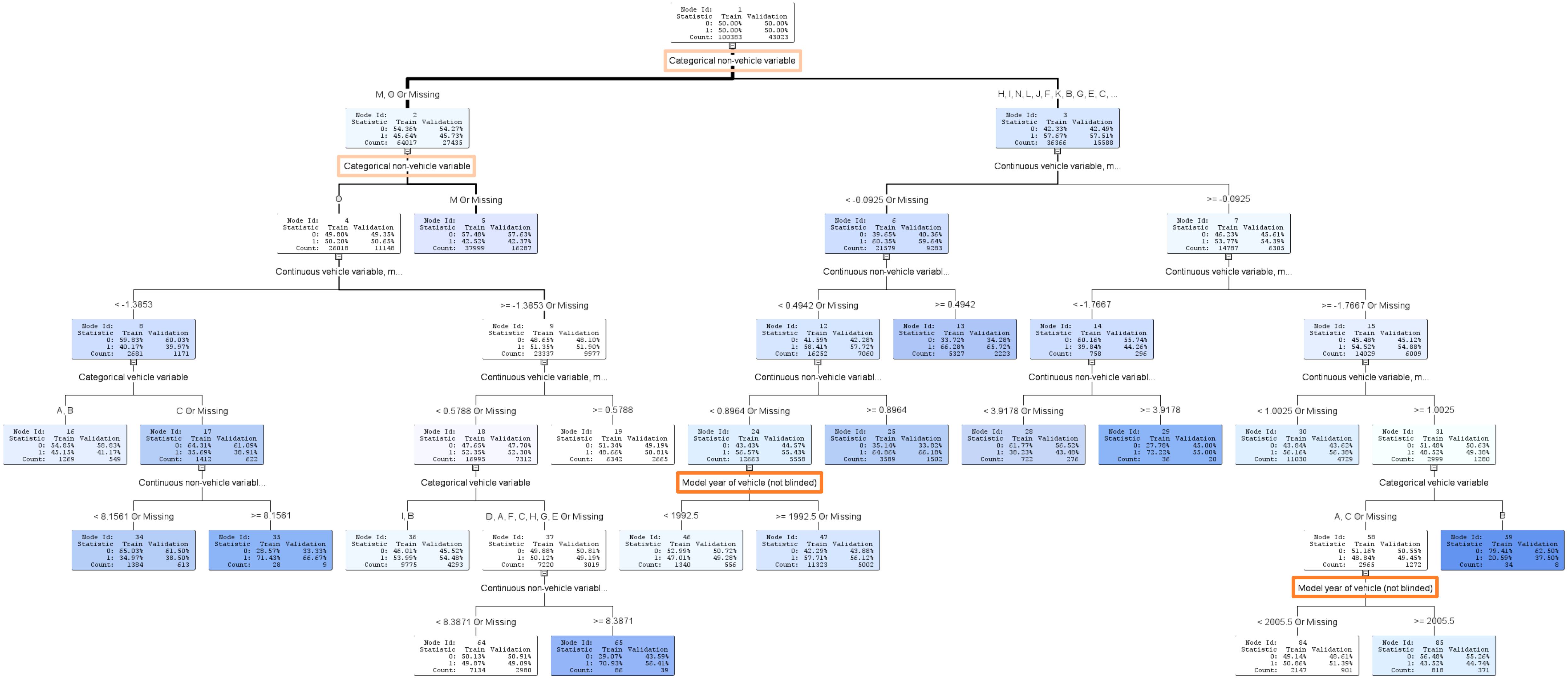
Versicolor 97.92% and Virginica 2.08%

Example decision tree – pruning based on validation data





Variable importance in decision trees



Variable importance in decision trees

Variable Name	Label	Number of Splitting Rules	Importance
NVCat	Categorical non-vehicle variable	2	1
Var8	Continuous vehicle variable, mean 0 stdev 1	4	0.392026454
Var3	Continuous vehicle variable, mean 0 stdev 1	1	0.298358498
NVVar3	Continuous non-vehicle variable, mean 0 stdev 1	4	0.267762691
NVVar2	Continuous non-vehicle variable, mean 0 stdev 1	1	0.241597405
Model_Year	Model year of vehicle (not blinded)	2	0.198911935
Cat1	Categorical vehicle variable	1	0.120725455

Ensemble models

Ensemble models combine the results of many other models, often called **base learners**

Ensembles are often **more accurate than single models**

There are several common approaches to ensembles:

- Bootstrap aggregation (Bagging)
- Boosting
- Stacking (Super learner)

Ensemble models: *intuition*

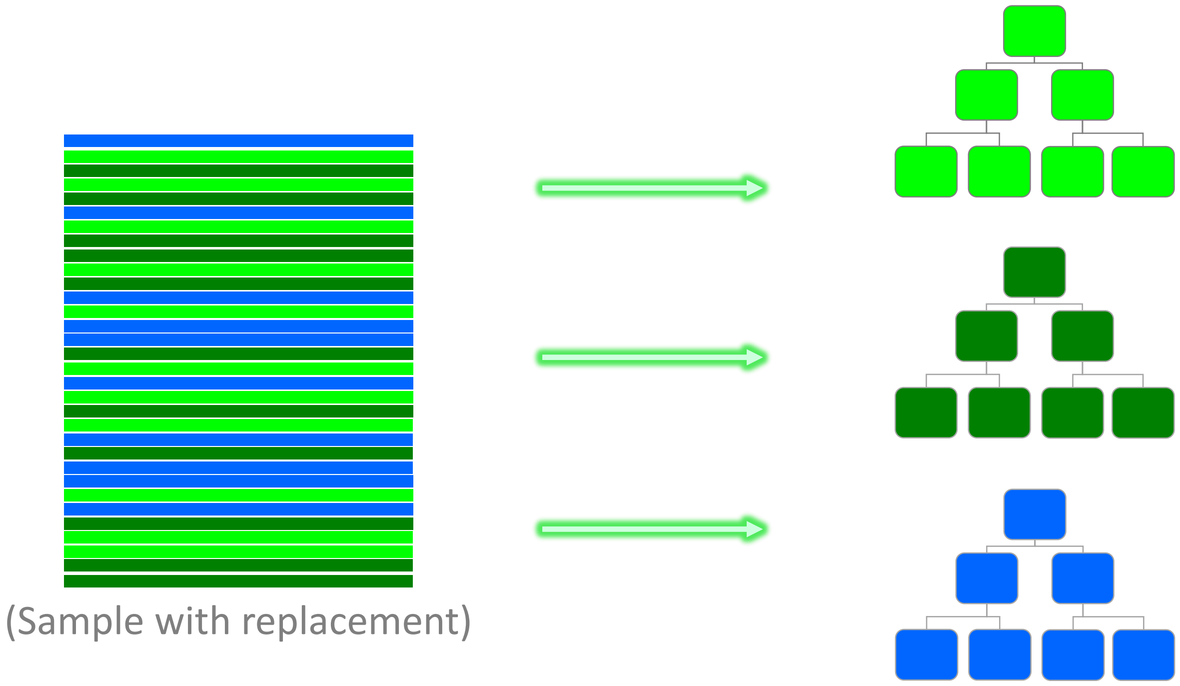
Variable hiding – important variables are often correlated and can hide one-another (only the single most important variable from a group of important correlated variables will be used in many models); in different samples, many different important variables can shine through

Representative samples – some samples can be highly representative of new data

Stability - the predictions of ensemble models are stable w.r.t. minor perturbations of training data

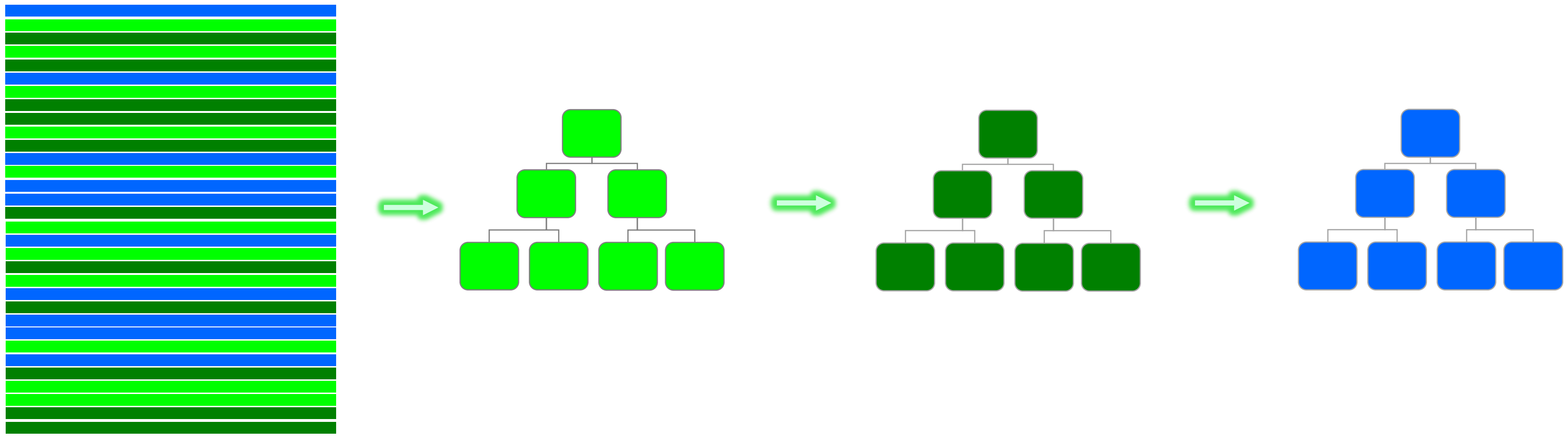
Decision Tree Bagging: Random Forest

Bagging is essentially a **parallel** process where the results of base learners are combined



Decision Tree Boosting: GBM

Boosting is essentially a **sequential** process where each subsequent base learner attempts to improve on past results



(Sample with replacement)