**ensemble methods:**

1. create multiple dataset

2. build multiple classifiers

3. combine classifiers

bagging (bootstrap aggregation)

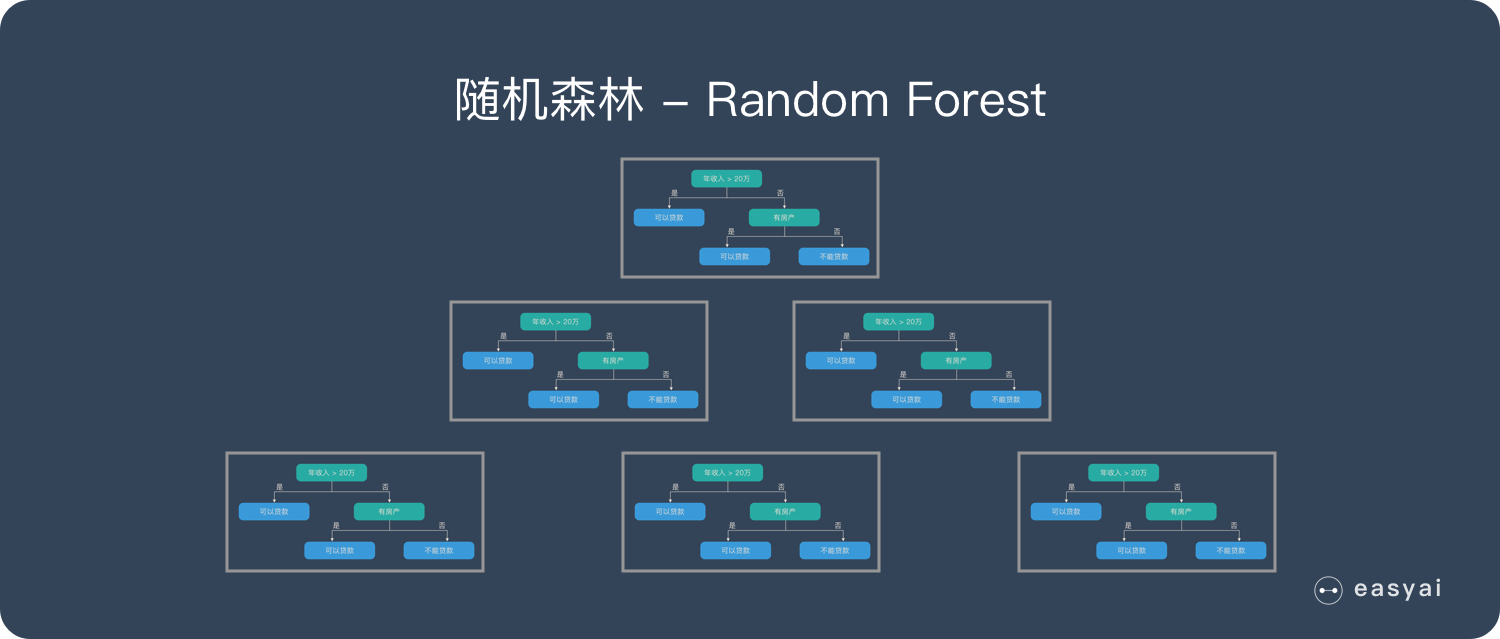
* making multiple replicates of original data by sampling with replacement from the training set (replicates same size as original)
* construct a single classifier for each replicate
* combine classifiers by taking a majority vote to produce the final decision



random forest:

a refinement of bagged decision trees

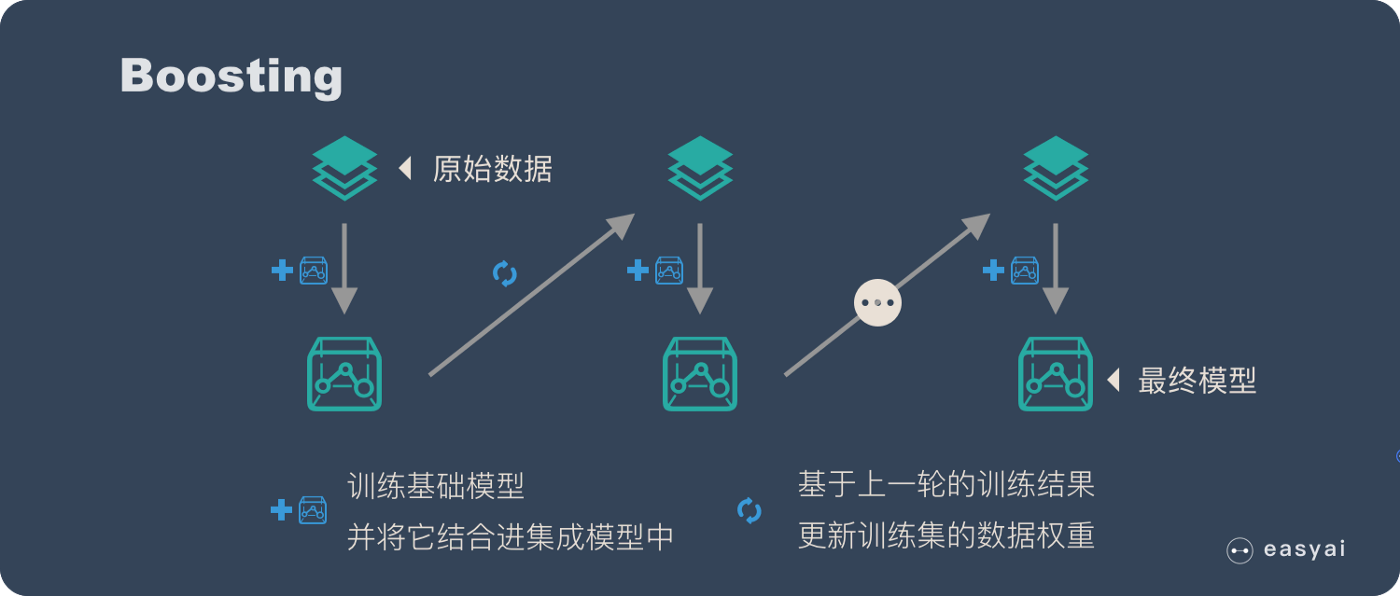
* create multiple data sets from original training set using subsets of data points and **subsets of attributes**
* build a decision tree classifier for each data set
* combine classifiers by taking a majority vote to produce the final decision



boosting:

for each tree:

* assign equal weights to each point in training set, fit basic tree
* repeat n iterations:
  + update weights of misclassified items and normalise
  + update tree
  + build on current tree
* output the final classifier as weighted sum of votes from each tree



Bagging VS Boosting

样本选择上：

Bagging：训练集是在原始集中有放回选取的，从原始集中选出的各轮训练集之间是独立的。

Boosting：每一轮的训练集不变，只是训练集中每个样例在分类器中的权重发生变化。而权值是根据上一轮的分类结果进行调整。

样例权重：

Bagging：使用均匀取样，每个样例的权重相等

Boosting：根据错误率不断调整样例的权值，错误率越大则权重越大。

预测函数：

Bagging：所有预测函数的权重相等。

Boosting：每个弱分类器都有相应的权重，对于分类误差小的分类器会有更大的权重。

并行计算：

Bagging：各个预测函数可以并行生成

Boosting：各个预测函数只能顺序生成，因为后一个模型参数需要前一轮模型的结果。