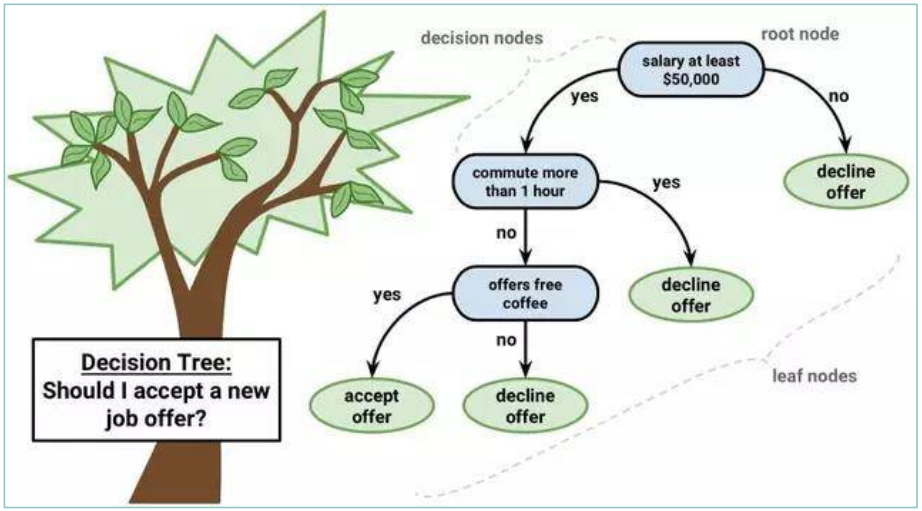
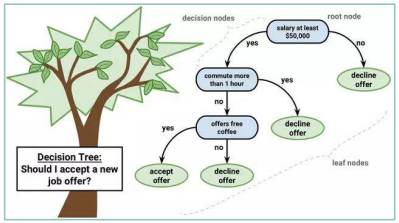
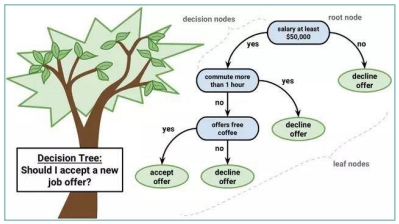
随 机 森 林

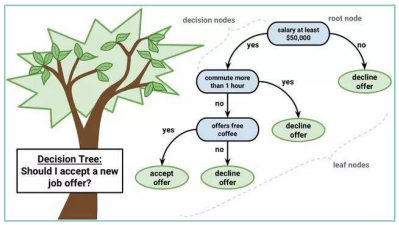
2 决　策　树

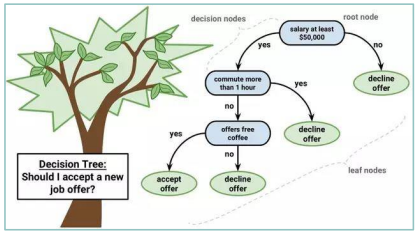


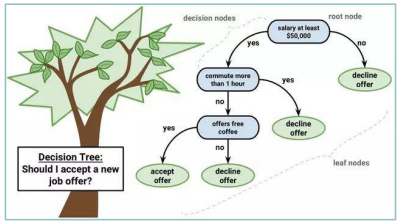
决策树是单个模型进行决策

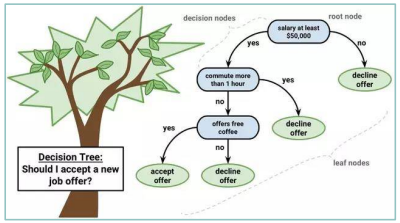








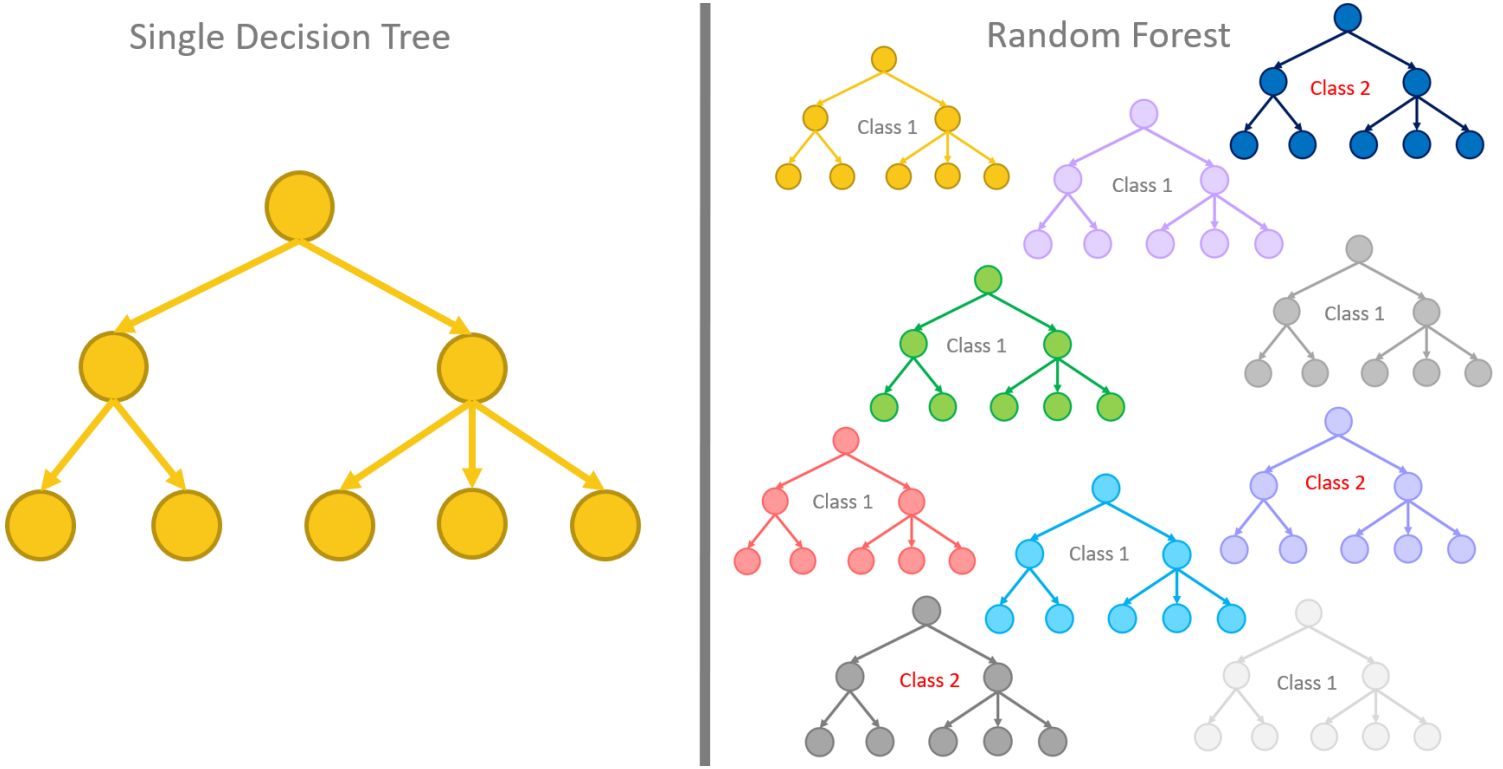




3 随机森林

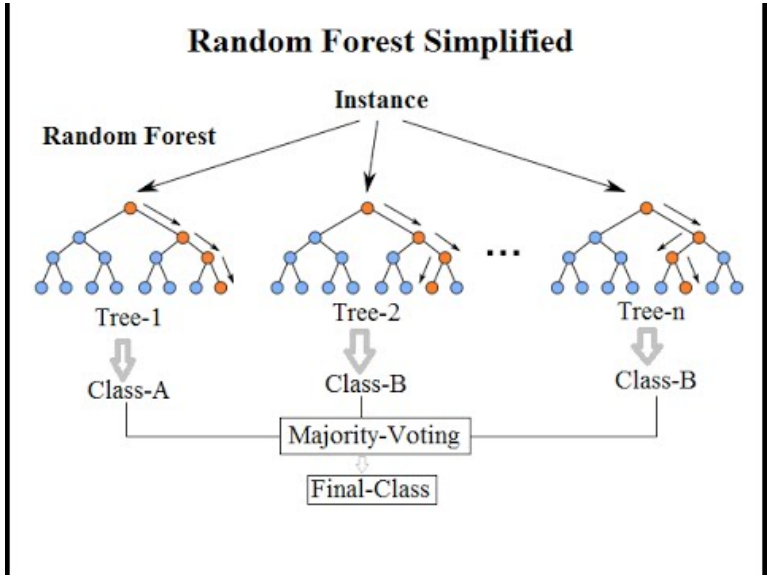
群体智慧：随机森林是多个决策树投票进行决策

4 随 机 森 林

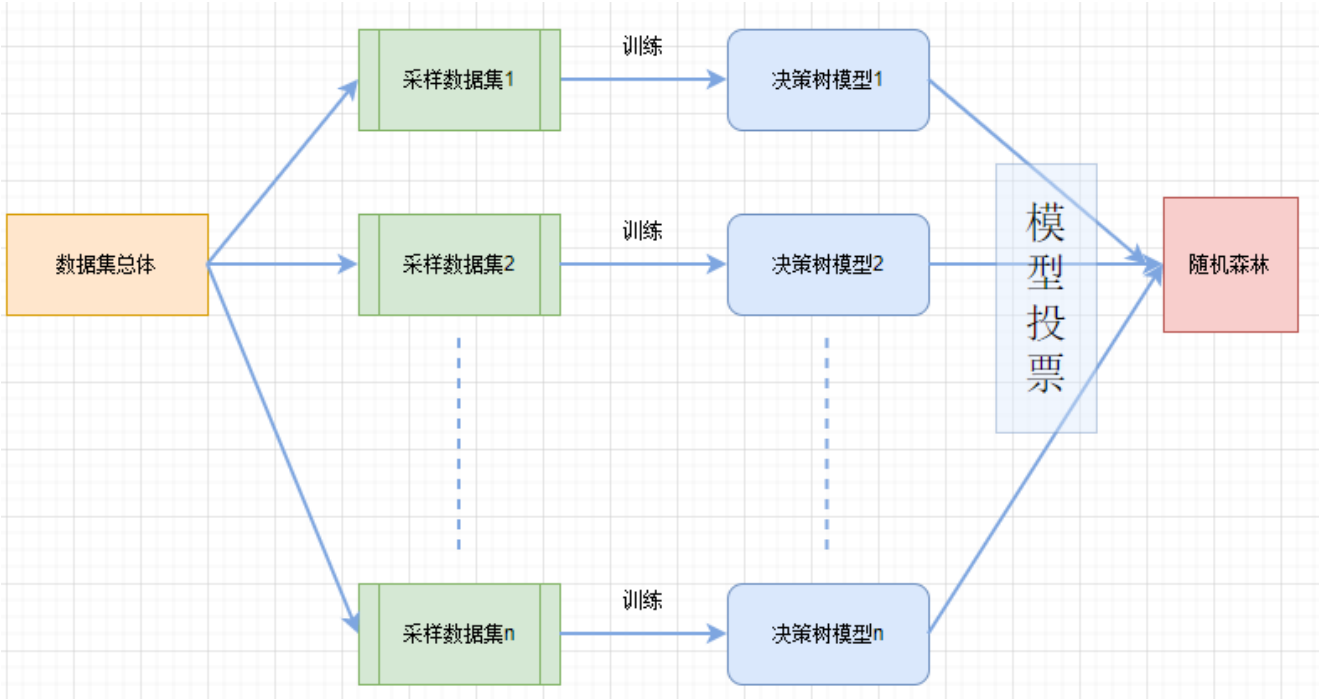

Single Decision Tree
Random Forest
Class 1
Class 2
Class 1
Class 1
Class 1
Class 1
Class 2
Class 1
Class 2
Class 1

5 随机森林定义

随机森林是用于分类，回归和其他任务的集成学习方法，其通过在训练时构建多个决策树并输出作为类的模式（分类）或平均预测（回归）的类来操作。随机决策森林纠正决策树过度拟合其训练集的习惯。



6 随机森林原理


数据集总体
采样数据集1
采样数据集2
采样数据集n
训练
训练
训练
决策树模型1
决策树模型2
决策树模型n
模型投票
随机森林

7 随机森林算法

Step 1 随机抽样

对具有M个样本的数据集进行随机有放回的抽样，构建出m个子集样本数据。比如每次只抽取70%的数据

Step 2 随机选取分裂点

样本有N个特征，在每个子集当中随机抽取n个特征用来构建决策树。比如每次抽取80%的特征

Step 3 重复构建多棵树

重复Step 2直到满足收敛条件。比如决策树不再可分或者达到最大的子树的数量为止

Step 4 投票组合形成森林

构建好的 T 颗数就形成了森林，每颗数都可以对样本进行打分或者投票，最终结果取平均值。

8 随机森林-OOB

OOB：out of bag；决策树每次约有1/3的样本不会出现在抽样的集合中，它可以被用来进行泛化误差的估计。

样本	g_1	g_2	g_3	……	g_n
(x_1, y_1)	D_1	*	D_3	*	*
(x_2, y_2)	D_1	*	*	*	D_n
(x_3, y_3)	*	D_2	*	*	D_n
…….	*	*	D_3	$D_{...}$	*
(x_m, y_m)	D_1	*	*	*	*

样本(x_m, y_m)没有被模型g_2，g_3，g_n抽样选取到，因此可以用这几个模型对该样本进行预测和评估，按照模型投票或者平均原则。同理，其它样本也可以这样被进行计算

9 随机森林-单颗树构建

单颗树构建过程

1. 用M来表示训练用例（样本）的个数，N表示特征数目。
2. 输入特征数目n，用于确定决策树上一个节点的决策结果；其中n应远小于N。
3. 从M个训练用例（样本）中以有放回抽样的方式，取样m次，形成一个训练集（即bootstrap取样），并用未抽到的用例（样本）作预测，评估其误差。
4. 对于每一个节点，随机选择n个特征，决策树上每个节点的决定都是基于这些特征确定的。根据这n个特征，计算其最佳的分裂方式。
5. 每棵树都会完整成长而不会剪枝