什么是随机森林

森林里有很多树，随机森林里有很多决策树。

随机森林是决策树的升级版，随机指的是树的生长过程。世上没有两片相同的树叶，随机森林中的树也各不相同。在构建决策树时，我们会从训练数据中有放回的随机选取一部分样本，同样的，我们也不会使用数据的全部特征，而是随机选取部分特征进行训练，每棵树使用的样本和特征各不相同，训练的结果自然也不同。

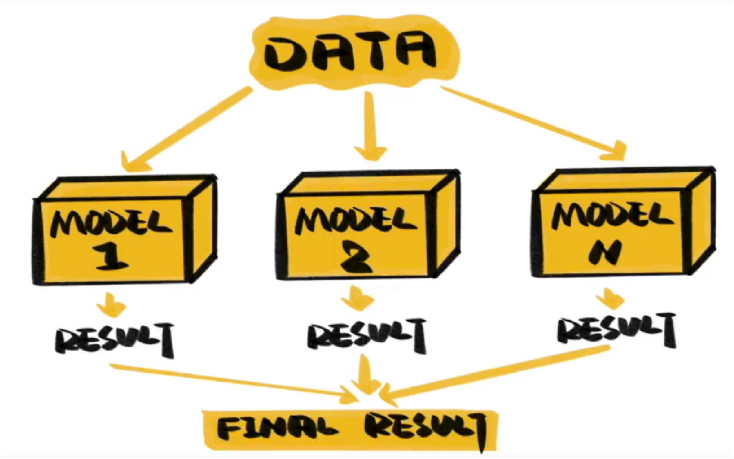
为什么要这么做?

在训练的最初，我们并不知道哪些是异常样本，也不知道哪些特征对分类结果影响更大，随机的过程降低了两者对分类结果的影响。

随机森林的输出结果由投票决定，如果大部分决策树认为测试数据是好苹果，那我们就认为它是个好苹果。这很像是人类的民主决策，虽然每个人拥有的信息，推理过程和结论各不相同。但当每个人都拥有投票权时，往往能做出较优的决策。因为树与树之间的独立，它们可以同时训练，不需要花费太长时间。随机的过程让它不容易过拟合。

能处理特征较多的高维数据，也不需要做特征选择，合理训练后准确性很高。不知道使用什么分类方法时，先试试随机森林准没错。

在机器学习中，随机森林属于集成学习，也就是将多个模型组合起来解决问题，这些模型会独立学习、预测、再投票出结果。准确度往往比单独的模型高很多。



除了决策树，还可以使用神经网络等其他模型。

同样的，集成学习内部不必是同样的模型，神经网络和决策树可以共存于一个系统中。