什么是决策树？一种机器学习的算法，较为古老。顾名思义，最终我们应该会得到一个像树一样的结构。我们去构造一棵树，在树中，我们能进行一系列的决策，比如分类工作，当然也是可以去做回归的工作。

每一个结点都是一次划分，比如，年龄的判断，将数据进行划分一次，通过一次特征将所有的数据进行划分一次。回答为Y的，那么如果数据还没办法进行区分，那么会进行再一次判断。直到没办法划分，那么成为叶结点，每个中间结点都是一次属性的判断。

之后就是如何去确定根节点？非叶子结点的顺序？如何构造决策树？

构造的基本想法？随着树深度的增加，节点的熵迅速地降低，熵降低的速度越快越好，这样我们有望得到一颗高度最矮的决策树。也就是说不断找到影响分类最主要的元素，这样不断划分，才能使熵不断降低，越来越纯。

怎么样决定根节点？根据构造的基本想法，则可以认为哪种切分方法能够使划分后数据熵值最低，就选择哪个。可以先对所有的属性进行一次划分，划分完对于所有的划分可能计算一次熵值，找信息增益值最大的那个属性。

什么是信息增益？从原本的熵值，从一次分类之后的熵值，他们的差值就是信息增益，

信息增益有么有可能有一些其他的问题？信息增益是一个比较古老的概念，应该是ID3版本。可能出现ID这种属性的影响，都是独立分布的，熵值会特别小，信息增益会特别大，但是使用ID并没有跟最终的label没有关系。这种属性就是要解决属性分类很多而导致每个属性分类数量很少，导致信息增益很大的情况

什么是信息增益率？是为了解决信息增益而来，信息增益再去除以自身的信息增益，而自身因为分类比较多，所以这个分母是非常小的

什么是评价函数？去评价叶子结点

什么是C4.5？

如何划分为连续值？

为什么需要剪纸？这里需要了解到，决策树也是有过拟合这个说法的，当你讲所有的数据都特别好的分类，也就是说每个样本可能都去分类的时候，将会导致分的过散，而导致过拟合的情况，如果想解决这个问题，则应该考虑使用剪枝的方法，决策树不能建立的太大

剪枝的两种方式？预剪枝，在构建决策树的过程时，提前停止；后剪枝：决策树建好后，再进行剪枝。叶子结点越多，损失越大，结合到评价函数，来评价当前的决策树。

什么是随机森林？森林就是说构造了很多的决策树，每个决策树都对测试集进行测试，共同的独立进行决策。那么什么是随机的，一个是数据的随机选择性，指定一个决策树只使用部分数据，这样能够可能去掉一些异常点。也有防止过拟合的操作，类似dropout。这是第一个随机。第二重随机是指特征随机选择，特征也有可能异常的。只取部分的特征。

什么是bagging：有放回采样n个样本一共建立分类器。

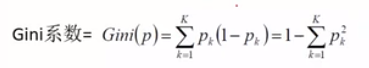
什么是有返回采样（bootstraping）？

操作流程？先根据当前的数据的label值，去计算一次熵值，直接计算label的两种概率，得到一个熵值

什么是label值？也就是标签，就是最终的分类结果

什么是熵？分子混乱程度，当前这个东西的**纯度**。从概率的角度来看，如何一个事件发生的可能性越大，则他的**不确定性**就越小。另一个角度来说，一个事件发生的可能性越小，则他的不确定性就越大（这个地方有点问题，理论上来说，应该是越不太确定的事情，这个不确定性才越大）。以上提到的两个概念，都可以映射到熵值，纯度越高，不确定性越小，熵值越低。

从公式的角度的话 ： ，可以理解为Pi，时间发生的概率越小，越不确定，则熵越大

另外一个相关指数为：，相反于熵，值越小表示越纯