## 决策树（Decision Tree）在材料学中的应用

原代码的解读及改进：

一般来说，先对数据进行了预处理

首先根据焓变进行了二分类。作者直接简单地根据金属氢化焓变进行了二分类，依据是金属氢化生成焓的分布，作者直接以40为界限，此处是可以进行调整的。

其次，将合金的组成成分由原来复杂的百分比，变为了二分类问题。

接着，勘察了所有合金中每种不同的金属出现的次数，原文中是出现次数少于Y的金属则进行剔除。

接下来，划分了测试集和数据集，参数为test\_size=0.3, random\_state=0。这是两个可以调整的参数。

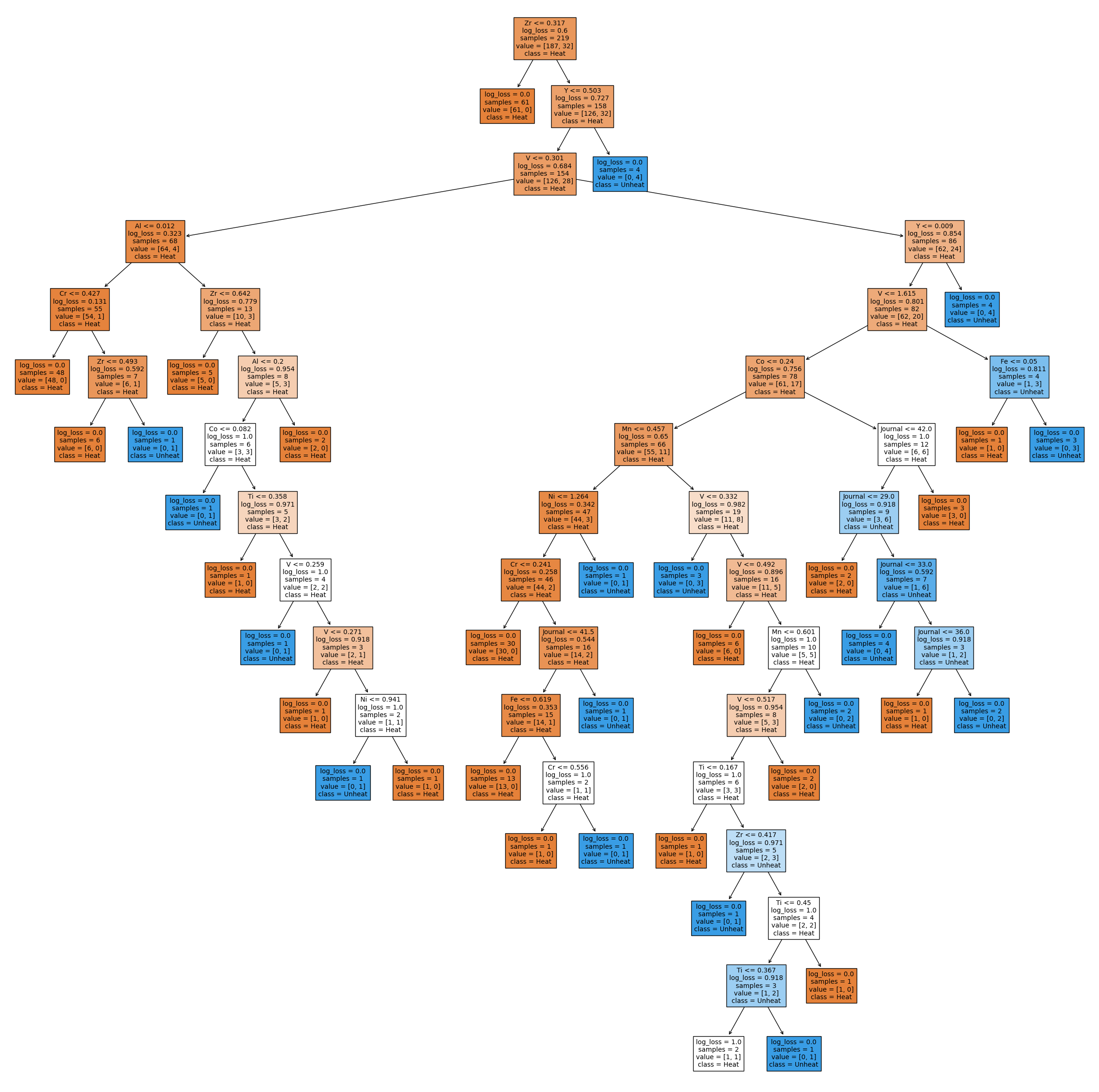
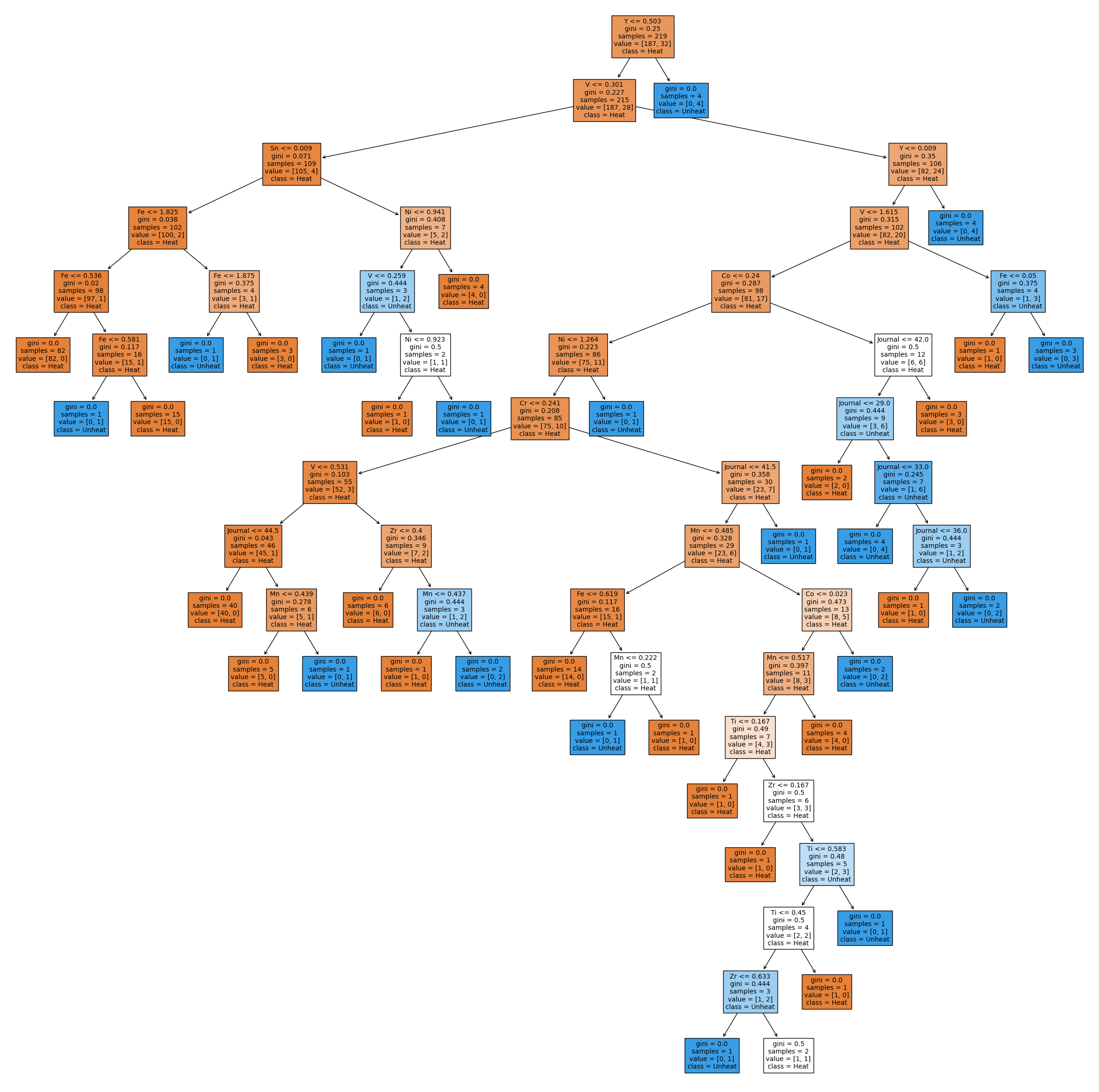
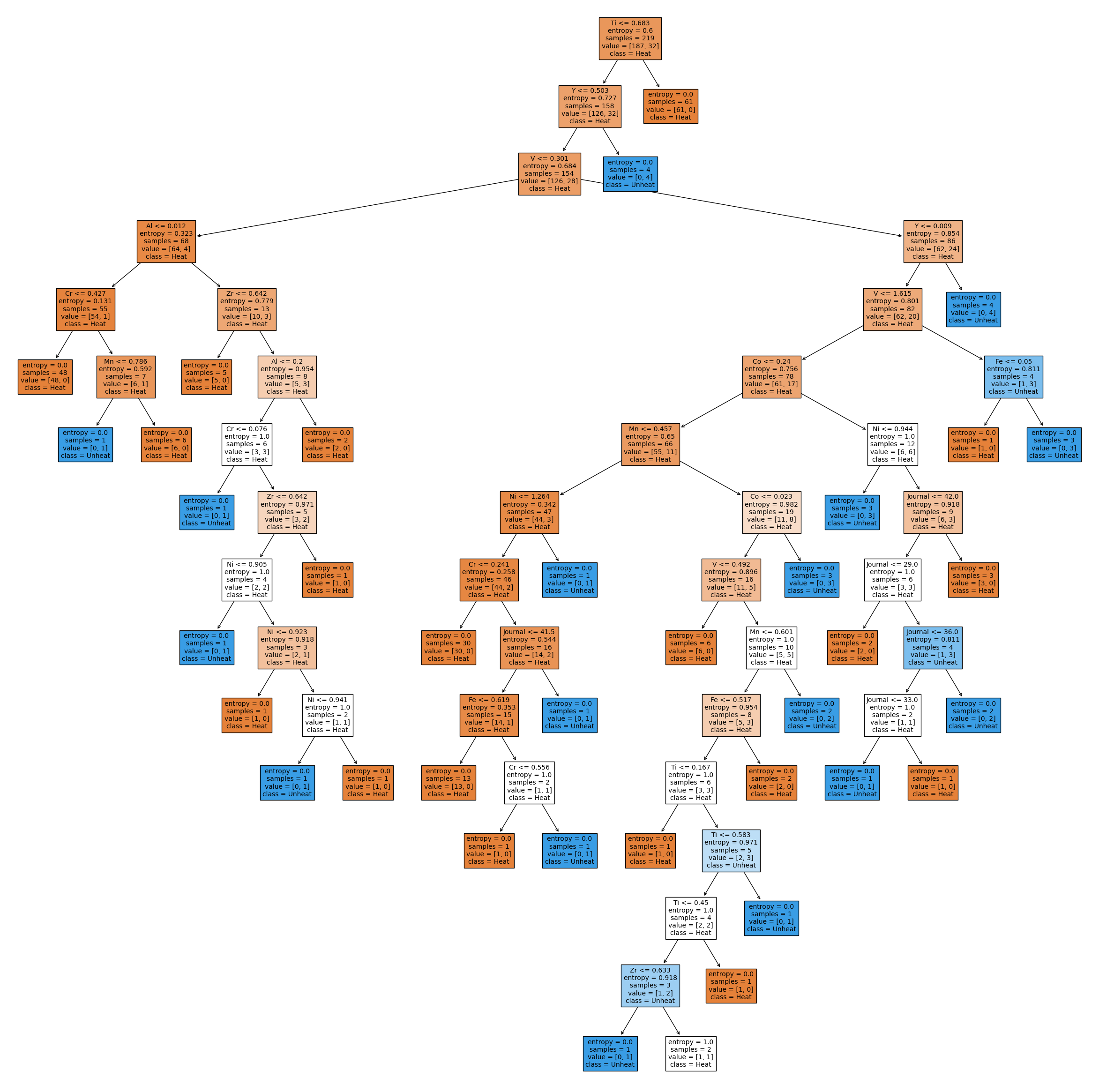
然后建立的模型，**在这里我们测试了三种不同的判断依据**，分别是信息熵，基尼系数，对数损失函数。

信息熵的准确度为：0.995

基尼系数的准确度为：0.8

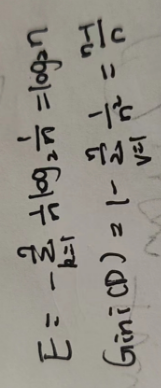
对数损失函数的准确度为：0.81

如下图左中右分别是信息熵，基尼系数和对数损失函数生成的结果，可以看出，在固定数据集的情况下，信息熵最为准确，且三种树的划分并不相同。

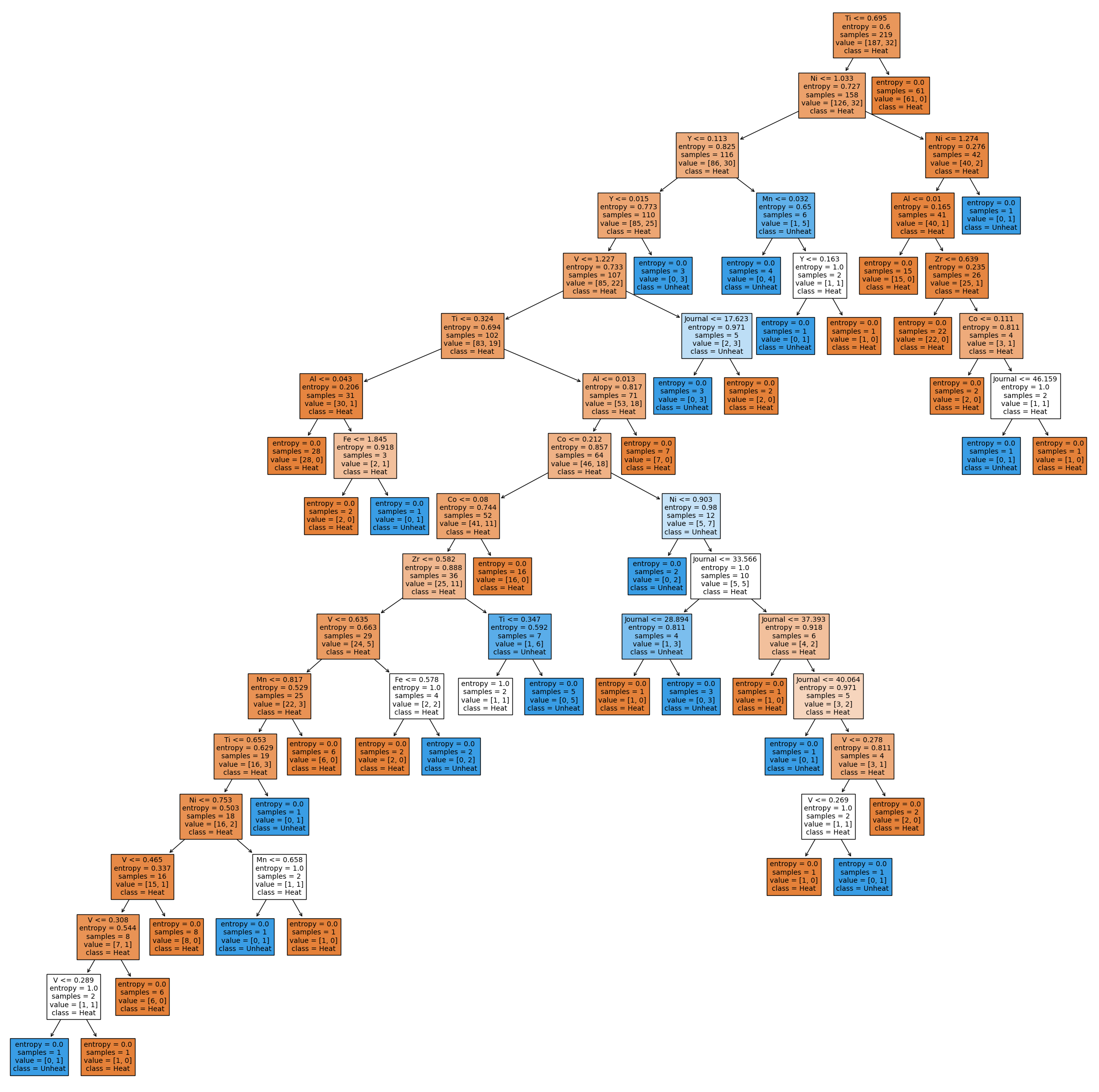


**我们自己从数学角度来看**：信息增益对较混乱的集合有很好的表现力，但是基尼指数有所欠缺。另一方面，这也说明较纯的集合，基尼指数可能会区分得更清楚。

其实可以从它们的极限来看这件事，显然在极端混乱的情况下E分的更清楚。



然而使用信息熵，容易出现过拟合的情况，尤其是对于混乱的模型，因此可以引入random\_state和改变splitter。可以发现在一定的范围内random\_state越大，测试集的准确度越低：当random\_state = 0, 1, 5, 10, 30 时，准确率分别为：0.800, 0.810, 0.810, 0.821, 0.810，绘制出的树也与前面的不尽相同。不过重复性有待进一步考证。



此时我引入了PCA降维（**文献中并没有引入**），**这是我们完全新引入的**，开始时为15维，然后分别取降到不同维度测试，其训练集误差均为0.995，但测试集误差分别为：

3维：0.758 4维：0.852 5维：0.842 6维：0.853 7维：0.821 10维：0.821 15维：0.810

可以发现，PCA降维确实能够在一定范围内提高决策树模型的泛化能力。但如果降维过多，会丢失重要特征。

还有一些其它的剪枝策略：比如限制树的深度，限制节点的子节点的sample个数，限制特征数和限制信息增益的大小。

然后可以通过优化曲线来确认最终的参数。当然也有其它的方法来做超参数优化，比如高斯过程。但限于时间关系，鄙人暂未实现相关代码。

**然后我们自己考虑将**PCA和max\_depth 融合，在不变参数的情况下，测试集准确度反而降低了，仅为0.832。说明不同的模型需要有不同的超参数。

引入目标权重，相当于引入了先验概率，可以进行人为干涉。在此数据集中可以根据期刊类型来赋予不同的权重。

唐浩程，李昭阳：负责优化代码，debug，引入新的代码进行测试，数学推导

钟圣，沈略号：阅读文献写文档