**4.3 试编程实现基于信息熵进行划分选择的决策树算法，并为表4.3中数据生成一颗决策树。**

**①问题描述：**

根据表4.3的数据基于信息熵选择最优划分属性从而递归生成一颗决策树。选择最优划分属性的依据通常有3中，分别是根据信息增益、增益率和基尼指数进行选择，不同的选择方式可能会生成不同的决策树。

**②实现过程：**

a.使用pandas数据处理库加载已导入.csv文件中的数据集。

b.针对数据集中的特殊情况进行处理，若样本全属于同一类别，则将节点标记为该类别的叶节点；若样本属性为空集或样本在属性上的取值相同，则将节点标记为叶节点，对其中所属各类别的数量进行统计，将该节点标记为统计结果最多的类别。

c.根据ID3或C4.5算法选择最优划分属性。

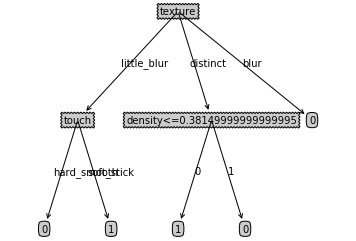
d.为最优划分属性中的每一类值生成一个分支，若样本中该属性取值为该类的样本集为空，则将该分支节点标记为叶节点，对其中所属各类别的数量进行统计，将该节点标记为统计结果最多的类别；若样本中该属性取值为该类的样本集为非空，则对该样本子集及仍未划分的属性进行递归操作，继续生成分支节点。

e.递归结束生成用字典表示的决策树。

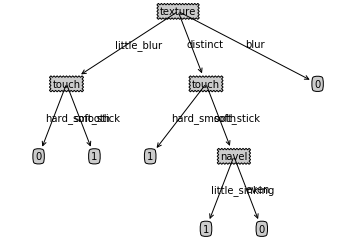
f.使用matplotlib库生成该决策树的可视化图像。

**③实现结果：**

ID3算法生成决策树（除去含糖率属性）：



C4.5算法生成决策树（除去含糖率和密度属性）：



**④结果分析：**

不同的选取最优划分属性的标准将导致生成决策树的不同，对于连续值属性的划分相较于离散值属性复杂许多，也较容易出错，C4.5算法中针对连续值属性部分代码的调试仍存在一些问题，故仅使用离散值属性进行测试。ID3算法中根据信息增益越大来选择最优划分属性，该准则对可取值数目较多的属性有所偏好，抑或是仅根据增益率越大来选择最优划分属性，该准则对可取值数目较少的属性有所偏好，C4.5算法利用启发式将二者结合，先从候选划分属性中找出信息增益高于平均水平的属性，再从中选择增益率最高的属性。