# 《机器学习与深度学习》课程

# 实 验 报 告



**姓 名： 金家耀**

**专 业：**  人工智能

**学 号： 1193210320**

**江南大学人工智能与计算机学院**

# 决策树分类

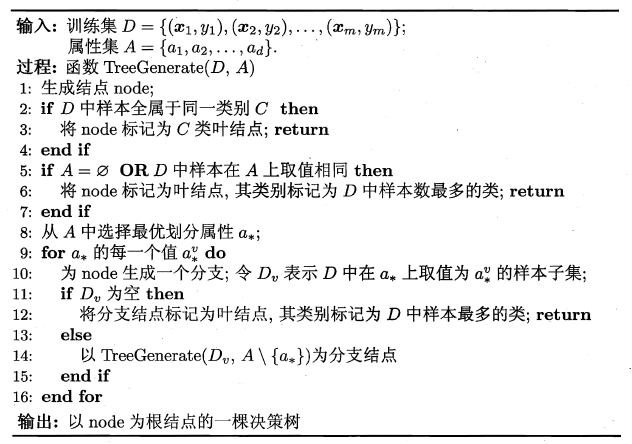
**1实验目的**

决策树是一类常用的机器学习方法，能把特征选择与分步决策相结合来完成分类或预测任务。本实验目的在于加深学生对决策树基本原理及相应算法的理解，

体会其在模式分类、识别中的作用。

**2实验原理**

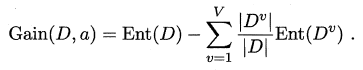
一棵决策树包含一个根结点、若干个内部结点和若干个叶结点; 叶结点对应于决策结果，其他每个结点则对应于一个属性测试;每个结点包含的样本集合根据属性测试的结果被划分到子结点中;根结点包含样本全集. 从根结点到每个叶结点的路径对应了一个判定测试序列. 决策树学习的目的是为了产生一棵泛化能力强，即处理未见示例能力强的决策树，其基本流程如下：

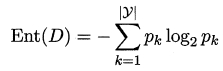


决策树学习的关键是第8 行，即如何选择最优划分属性. 一般而言，随着划分过程不断进行，我们希望决策树的分支结点所包含的样本尽可能属于同一类别，即结点的"纯度" (purity) 越来越高. 度量样本集合纯度常用的指标包括：

1）信息增益（ID3决策树学习算法采用）

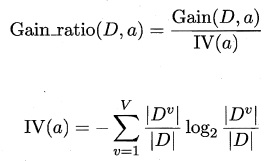
属性α对样本集D 进行划分所获得的"信息增益"为：





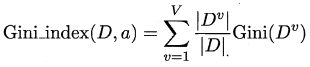
2）增益率（C4.5决策树学习算法）

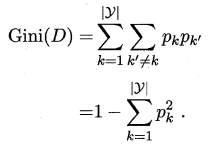
属性α对样本集D 进行划分所获得的"增益率"为：



3）基尼指数（CART决策树学习算法采用）

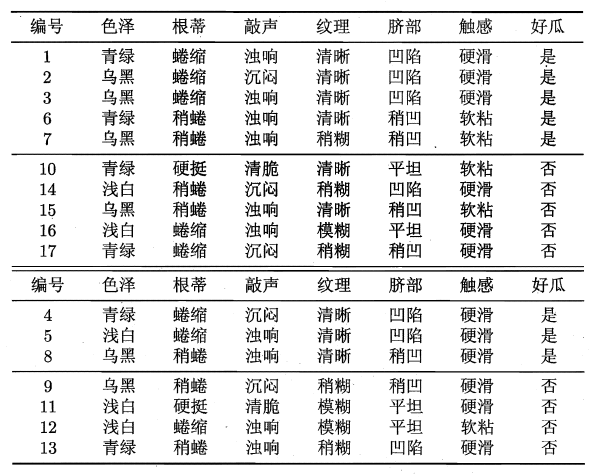
属性α的基尼指数定义为：





**3实验内容**

利用如下所示的西瓜数据集进行实验，训练生成决策树并进行测试。表中双线上部为训练集，双线下部为测试集。



**4实验要求**

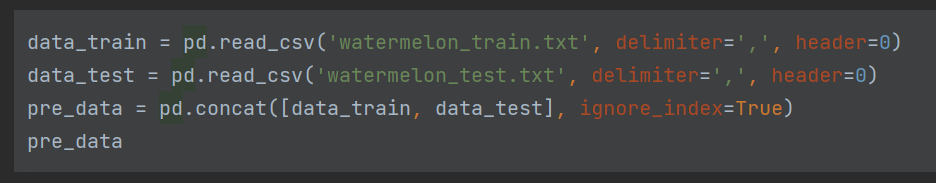
1. 根据实验原理中给出的基本流程设计实现决策树学习算法，要求度量样本集合纯度指标采用“信息增益”，画出决策树并给出相应的测试结果。
2. 将度量样本集合纯度指标分别替换为“增益率”和“基尼指数”，给出相应的决策树，并比较分类结果和分类精度，体会不同集合纯度指标对算法的影响；

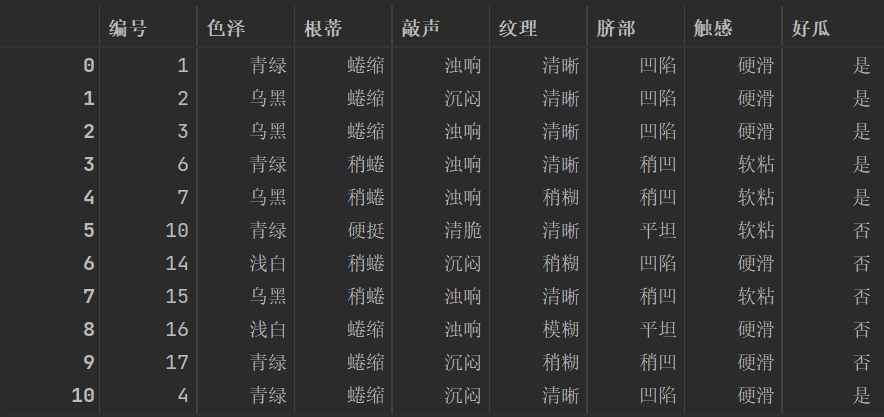
**5实验代码和结果**

**代码已上传至：**

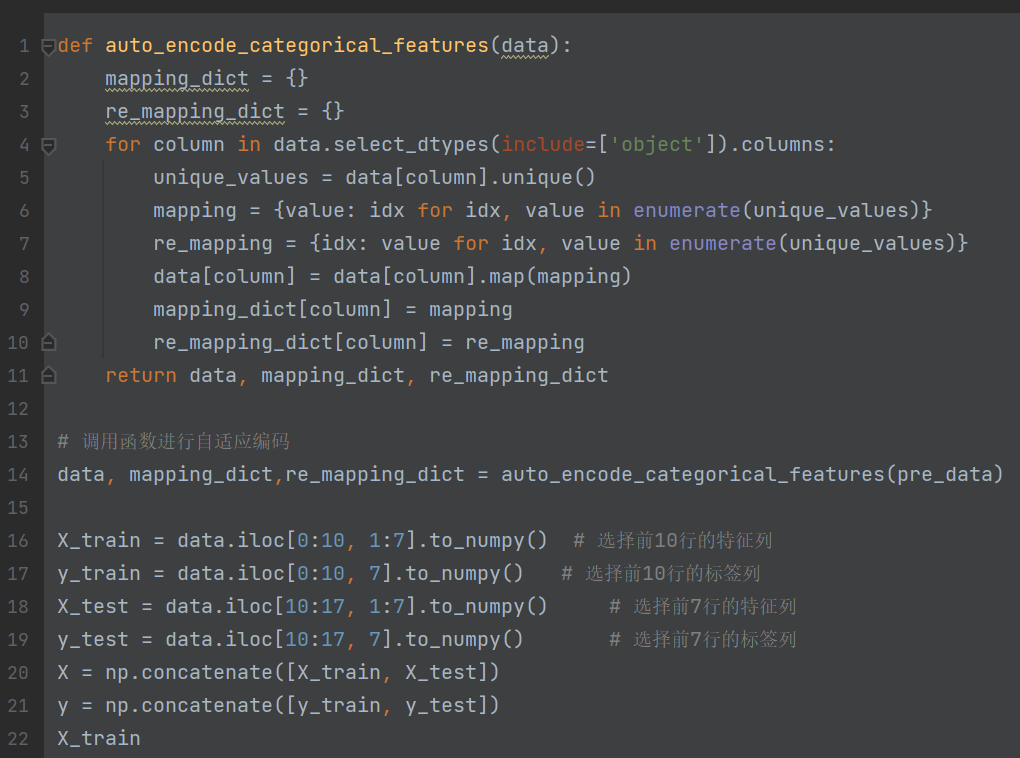
**<https://github.com/shinejjy/MachineLearningAndDeepLearning/tree/master/sy2>**

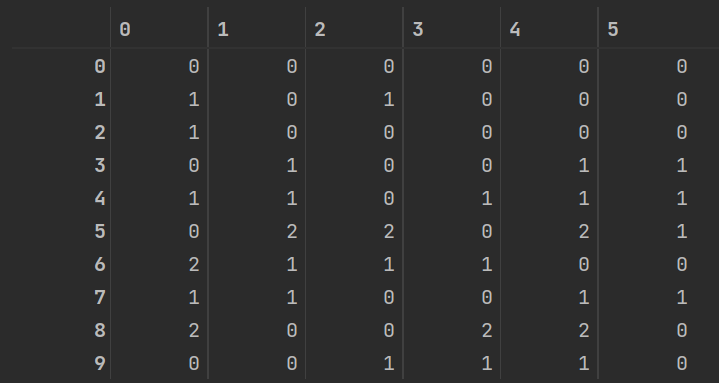
**5.1 读取西瓜数据集并作自适应数字化映射**





利用pandas库读取西瓜的训练数据集和测试数据集，为了方便后续处理，先将他们合并。如上图所示，为读取的部分数据集标签。观察到这些特征都由中文表示，不方便后续决策树算法的应用。

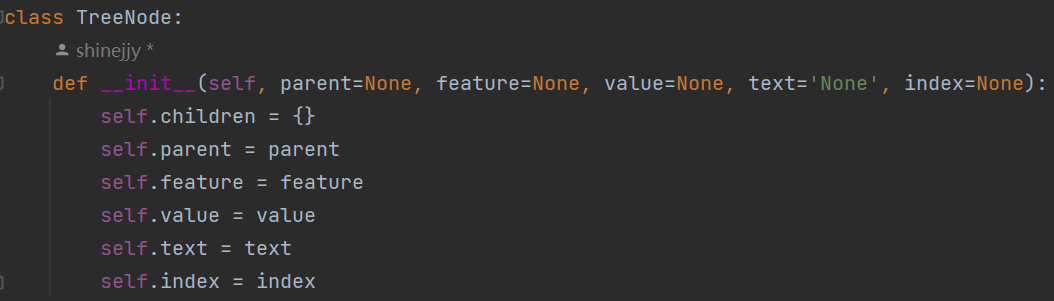




于是我设计了一个自动编码转换器的函数，能够将不同特征中的不同表示独一无二的表示成一个数字。如上图所示，为数字化之后的数据集编码，这里我把标题也同样转化成了数字嵌入表示。

**5.2 构建决策树**

**5.2.1 构建树节点类**

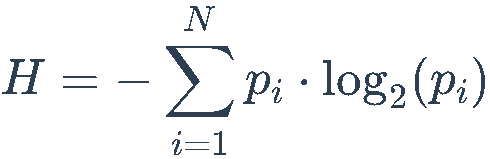


决策树也同样属于树，满足树的数据结构特征。这里我定义了一个TreeNode类，其中children表示该节点的子节点，parent表示该节点的父亲结点，feature表示非叶子结点时选取的最佳决策特征，value表示叶子节点时最后的决策判断值，text表示父亲结点指向该节点边上的文字，index表示该结点的索引（用于画最终的决策树图）。

**5.2.2 选择最佳决策特征的方法**

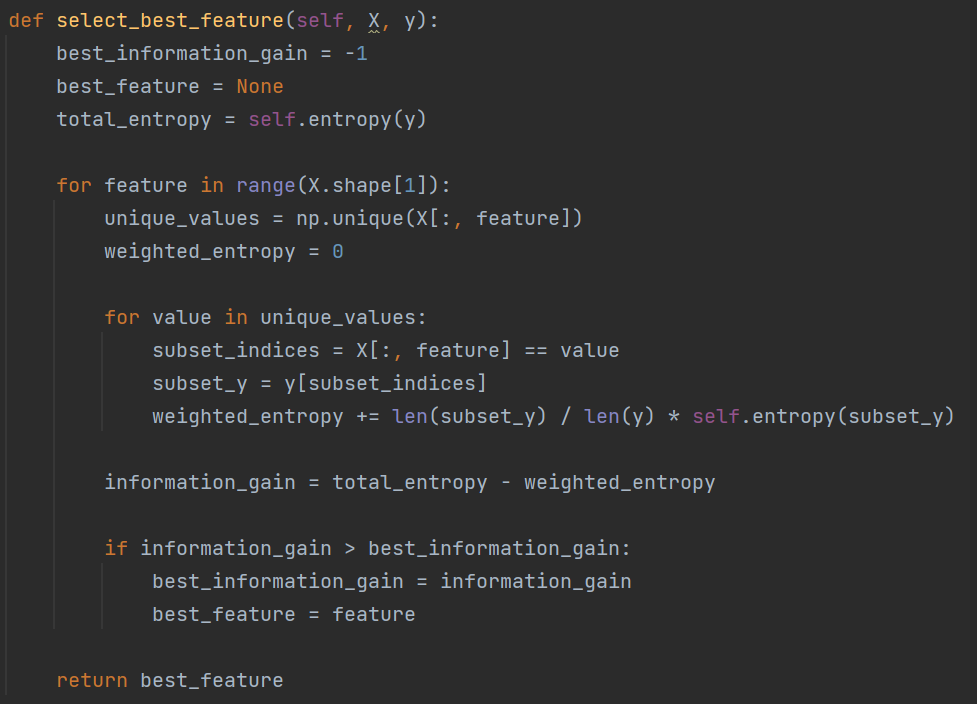
在信息理论中，熵是对不确定性的度量。在机器学习和数据科学中，熵常被用于衡量数据的混乱程度或无序程度。在决策树和其他机器学习算法中，熵是一种重要的概念，用于决定如何分割数据以获得更好的分类结果。

对于一个给定的系统，假设有N个不同的状态，每个状态i的概率为p\_i。系统的熵（entropy）可以用下面的公式来表示：



这里的log\_2通常以2为底的对数。但概率p\_i 越小时，p\_i·log\_2(p\_i)趋于0。熵的计算结果越高，表示系统的混乱程度越大。

**①信息增益**



信息增益是在决策树算法中用来选择最佳特征以进行数据拆分的度量。它基于信息熵的变化来衡量使用某个特征进行数据划分后的纯度提高程度。

在决策树的构建过程中，信息增益被用来确定哪个特征是最好的拆分特征，即哪个特征能够将数据分割成更“纯”的子集。

具体计算方法：

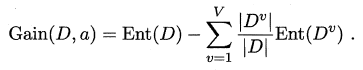
计算初始数据集的熵（总体不确定性）：首先，计算整个数据集的熵。熵越高，数据越不确定或混乱。

对每个特征计算条件熵：针对每个特征，计算当数据集根据该特征进行划分后的加权熵。条件熵衡量的是在已知特征的情况下，对数据集的不确定性。

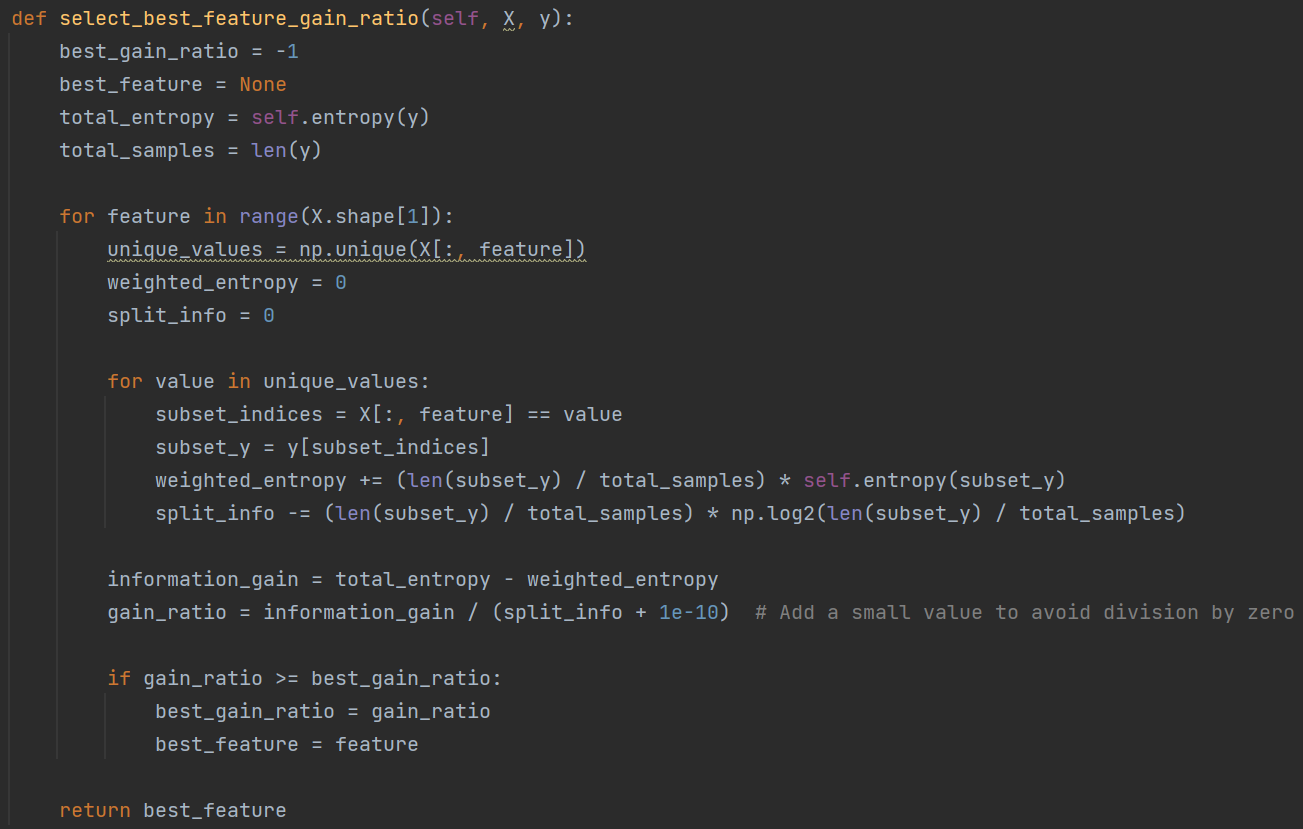
计算信息增益：信息增益是总体熵与特征划分后的条件熵之差。具体来说，信息增益等于总体熵减去特征划分后的条件熵。

选择信息增益最大的特征来进行数据划分，能够帮助决策树更有效地学习数据并找出最相关的特征，从而建立更准确的预测模型。

计算公式如下：



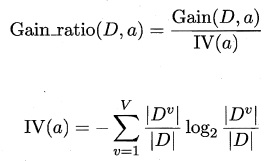
**②信息增益率**



信息增益率是对信息增益的一种扩展，用来解决信息增益存在的偏向于选择取值较多的特征的问题。

信息增益率考虑了特征本身取值的数量对信息增益的影响。信息增益倾向于选择取值更多的特征，因为这些特征可能有更多的可能性来进行划分，从而获得更高的信息增益。但这可能会导致过度拟合的问题。

计算公式如下：



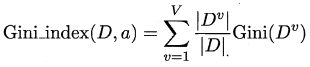
**③基尼指数**

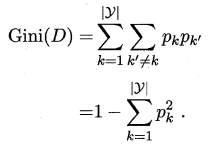


基尼指数被用来衡量从一个数据集中随机选取两个样本，其类别标签不一致的概率。基尼指数越低，表示数据集的纯度越高。

在决策树的构建过程中，当需要选择节点的拆分特征时，算法会计算每个可能特征的基尼指数，然后选择基尼指数最小的特征作为最佳拆分特征。这意味着选择能够使子节点更加纯净的特征来进行数据拆分。

计算公式如下：





**5.2.3 递归构建决策树并保存节点信息**

决策树算法流程：

**① 开始：**

·选择最佳的特征作为根节点。

**② 构建节点：**

·对于当前节点，根据选定的特征将数据集分割成子集。

·选择分割数据的特征以最大程度减少不纯度。

**③ 递归构建：**

·对每个子集，递归执行步骤 2 直至达到停止条件：

·如果达到最大深度或者子集中的标签唯一，停止分裂，将该节点标记为叶节点。

**④ 停止条件：**

·达到预设的最大深度。

·子集中的标签唯一。

**⑤ 输出：**

·返回构建好的决策树。



如上图所示，为我写的代码片段，其中：

**① 参数：**

·X 是特征数据。

·y 是标签数据。

·parent 是树的父节点。

·depth 是当前节点的深度。

·text 是当前节点表示的文本信息。

·criterion 是用于特征选择的衡量标准，可以是'entropy'、'gini' 或 'gain\_ratio'。

**② 终止条件：**

·如果达到最大深度（由 self.max\_depth 控制）或者所有标签相同，就停止树的分裂， 创建一个叶节点，该节点以出现次数最多的标签作为预测结果。

**③ 特征选择：**

·根据给定的衡量标准，选择最佳的特征来分割数据。根据不同的标准调用不同的选择 特征的方法：

·select\_best\_feature（信息熵）。

·select\_best\_feature\_gini（基尼指数）。

·select\_best\_feature\_gain\_ratio（信息增益率）。

**④ 构建节点：**

·创建当前节点 tree，记录选择的最佳特征，以及该节点对应的索引和文本信息。

**⑤ 特征分裂：**

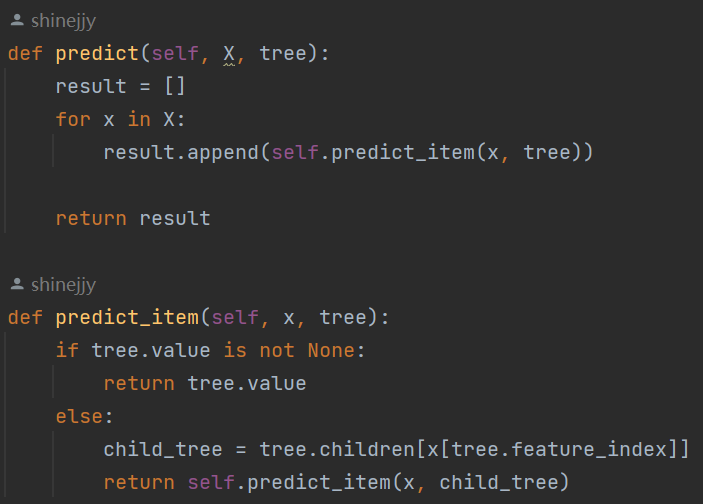
·对于选定的最佳特征，根据其不同取值对数据集进行分割，创建子节点。每个取值对 应一个分支，每个分支递归调用 fit 方法构建决策树。

**⑥ 递归构建：**

·每个子节点会根据相应的特征取值创建新的子集，直到满足终止条件。

**5.2.4 预测结点结果**

前者讲述了如何通过训练样本构建一颗决策树，下面就是如何通过一个决策树来预测样本。具体操作依据我个人理解与树的遍历类似，首先从根节点开始，读取根节点的特征类别，按照该样本的特征值继续往下搜索，当寻找到叶子节点时，即返回，或未到叶子节点但是其特征值分支不存在，我定义为“无法判断”。



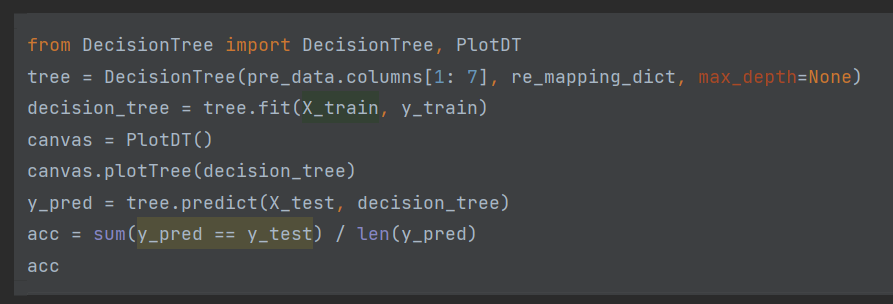
**5.3 绘制决策树图（自己手搓的代码）**

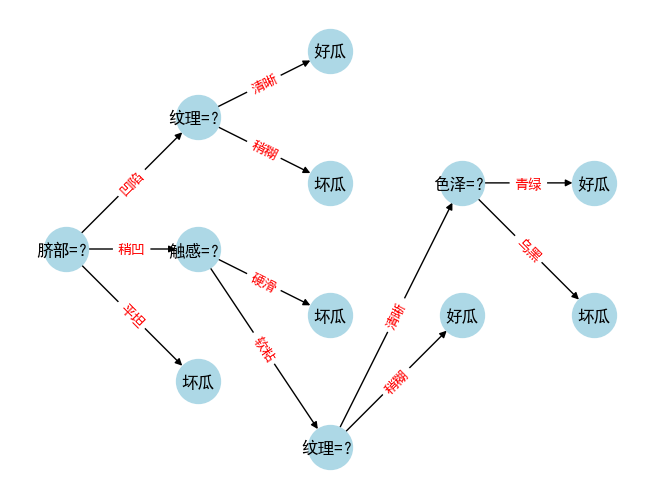


之前接触过GNN、GCN的一些算法，其中有些可视化代码中会使用networkx库来可视化图信。而决策树也属于图，这里，本人利用networkx库来实现决策树的可视化（**当然不只是决策树，它还可以被迁移到各种可视化树的场景中**）

**5.5 构建与绘制决策树并计算准确率**

**①信息增益**

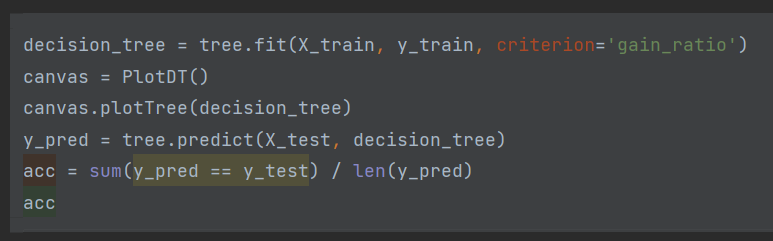


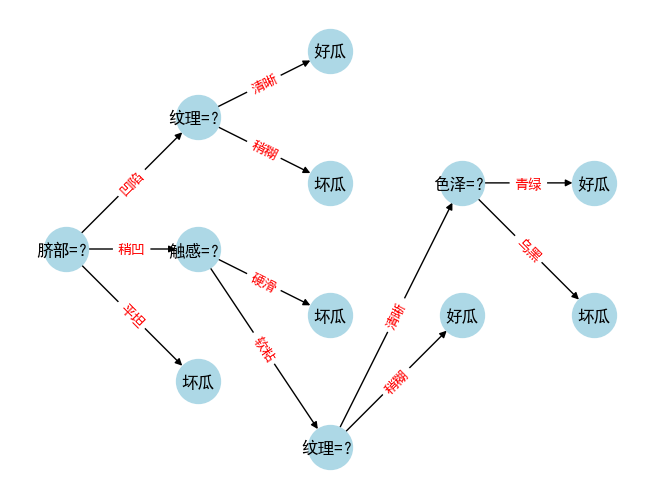




使用“信息增益”作为度量样本集合纯度，决策树如上图所示，测试集的准确率为85.71%（**其中在做根节点最佳特征选择时，其实按“脐部”和“色泽”分割所得到的指标是一致的，但按照“色泽”分割最终的准确率只有28%左右，于是，我选择了“脐部”作为最佳特征**）。

**②信息增益率**

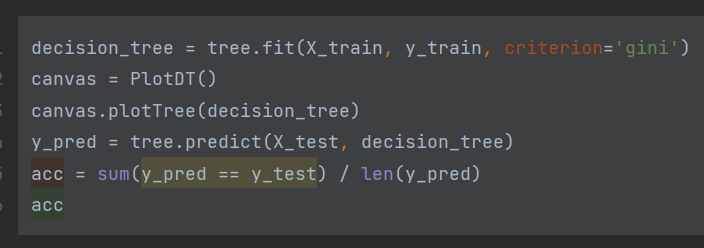


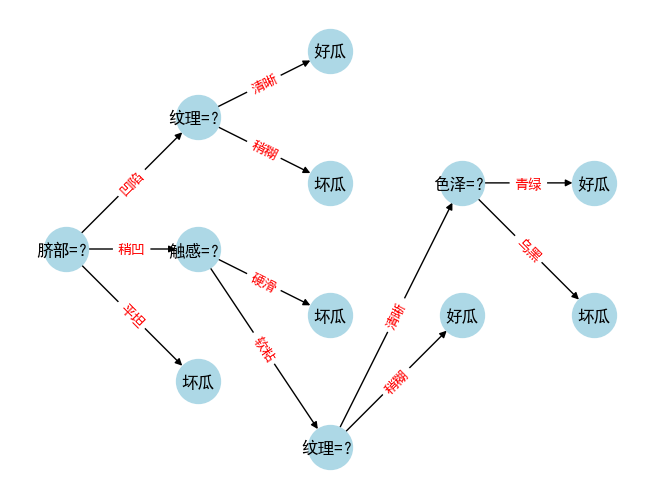




使用“信息增益率”作为度量样本集合纯度，决策树如上图所示，测试集的准确率为85.71%（**其他情况原因同上**）。

**③基尼指数**







使用“基尼指数”作为度量样本集合纯度，决策树如上图所示，测试集的准确率为85.71%（**其他情况原因同上**）。

1. **实验心得**

这次实验让我更深入地了解了决策树算法。数据预处理的部分让我意识到了数据清洗和转换对算法成功运行的重要性，特别是将中文特征转化为数字表示的过程，这让我更好地理解了数据预处理的必要性。

在构建决策树的过程中，我学会了选择最佳特征来分割数据以降低不纯度。不同的衡量标准对特征选择有着明显的影响，我尝试了信息增益、信息增益率和基尼指数，发现它们在构建树结构和在测试集上的表现上有些许差异。

绘制决策树图是一种很好的可视化方法，它让我更直观地了解了树的结构和节点之间的关系。这种方式对于展示和理解决策树模型是很有帮助的。

总的来说，这次实验丰富了我的对决策树算法的理解，让我更深入地认识了特征选择的重要性以及不同指标对决策树构建的影响。