**本 科 实 验 报 告**

实验名称： 决策树的python实现

学 员： 孙源辰 学 号： 201802001067

培养类型： 无军籍本科学员 年 级： 大三

专 业：人工智能与大数据 所属学院： 计算机学院

指导教员： 刘新旺 职 称： 副教授

实 验 室： 305-205 实验日期： 2021春季学期

国防科学技术大学训练部制

《本科实验报告》填写说明

1．学员完成人才培养方案和课程标准要所要求的每个实验后，均须提交实验报告。

2．实验报告封面必须打印，报告内容可以手写或打印。

3．实验报告内容编排及打印应符合以下要求：

（1）采用A4（21cm×29.7cm）白色复印纸，单面黑字打印。上下左右各侧的页边距均为3cm；缺省文档网格：字号为小4号，中文为宋体，英文和阿拉伯数字为Times New Roman，每页30行，每行36字；页脚距边界为2.5cm，页码置于页脚、居中，采用小5号阿拉伯数字从1开始连续编排，封面不编页码。

（2）报告正文最多可设四级标题，字体均为黑体，第一级标题字号为4号，其余各级标题为小4号；标题序号第一级用“一、”、“二、”……，第二级用“（一）”、“（二）” ……，第三级用“1.”、“2.” ……，第四级用“（1）”、“（2）” ……，分别按序连续编排。

（3）正文插图、表格中的文字字号均为5号。

一、实验目的和要求

通过python实现决策树的建立与预测

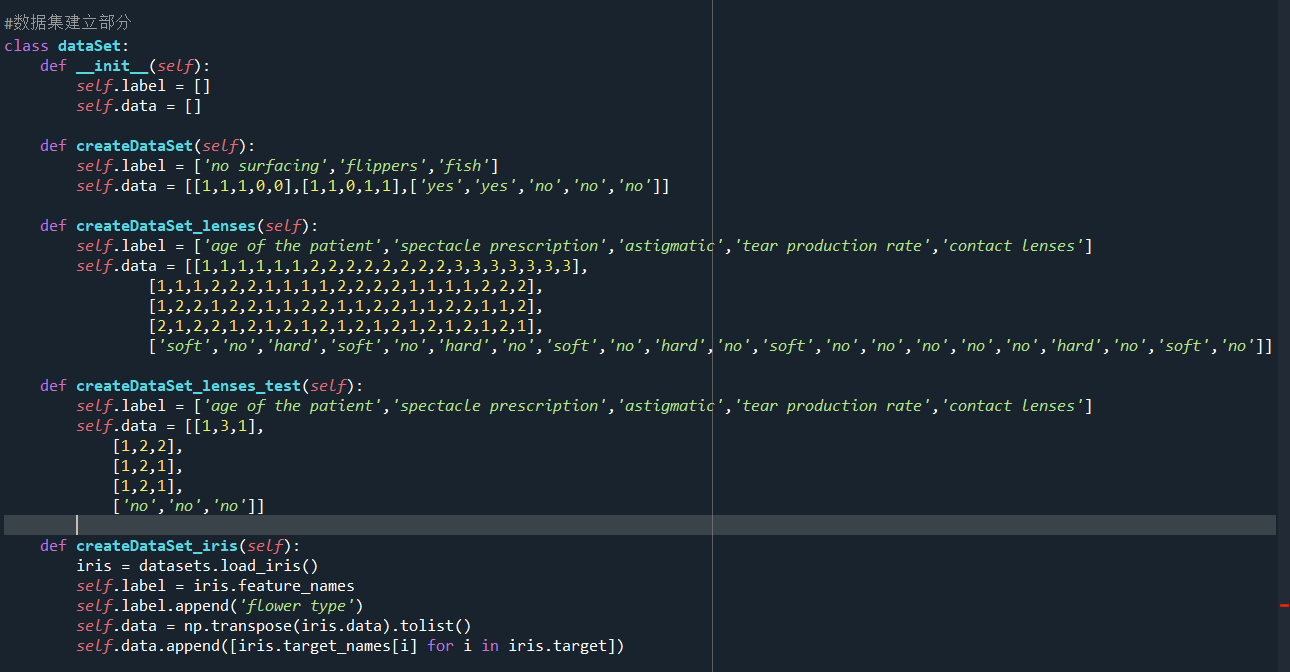
二、实验内容和原理、

决策树是通过训练将分类方法以树的形式表示的分类器，它使用贪心算法，找到信息熵最大的分类方法以实现准确性。

三、操作方法与实验步骤

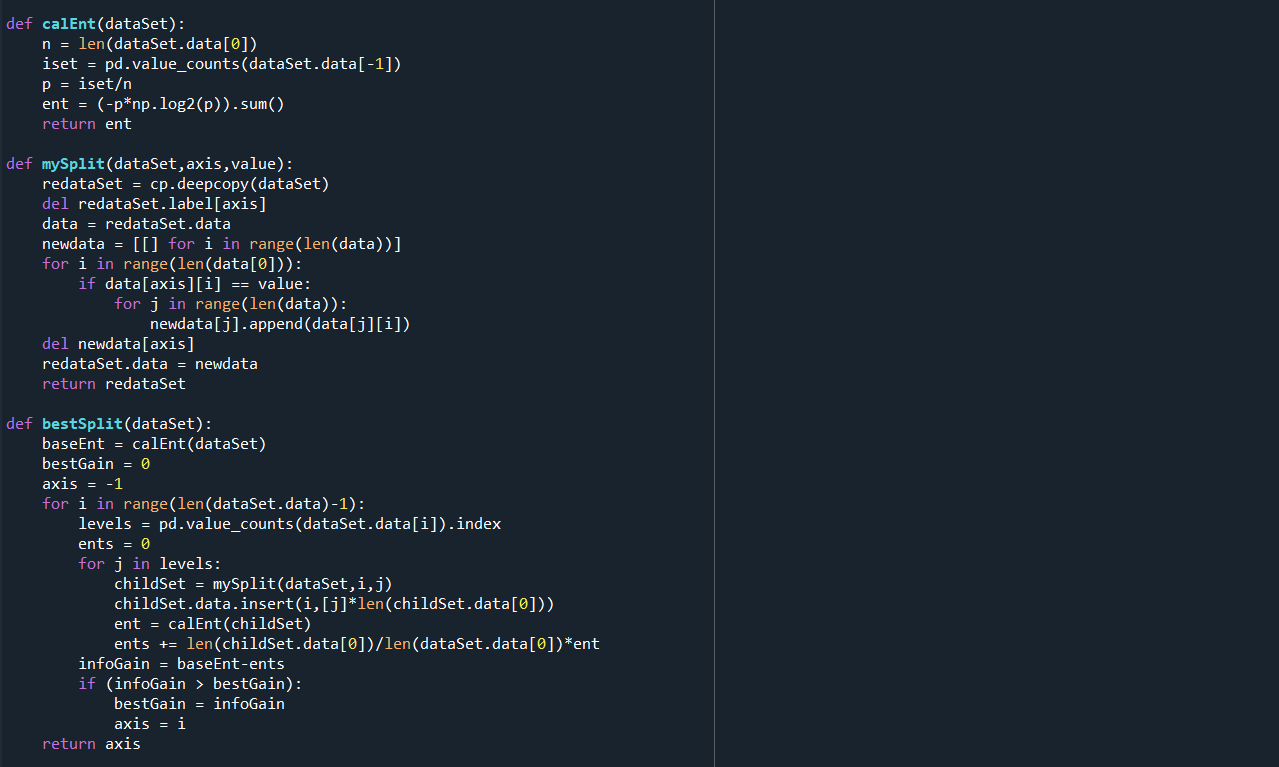
1.编程思路：

(1)定义数据集类，实现函数生成多种数据集



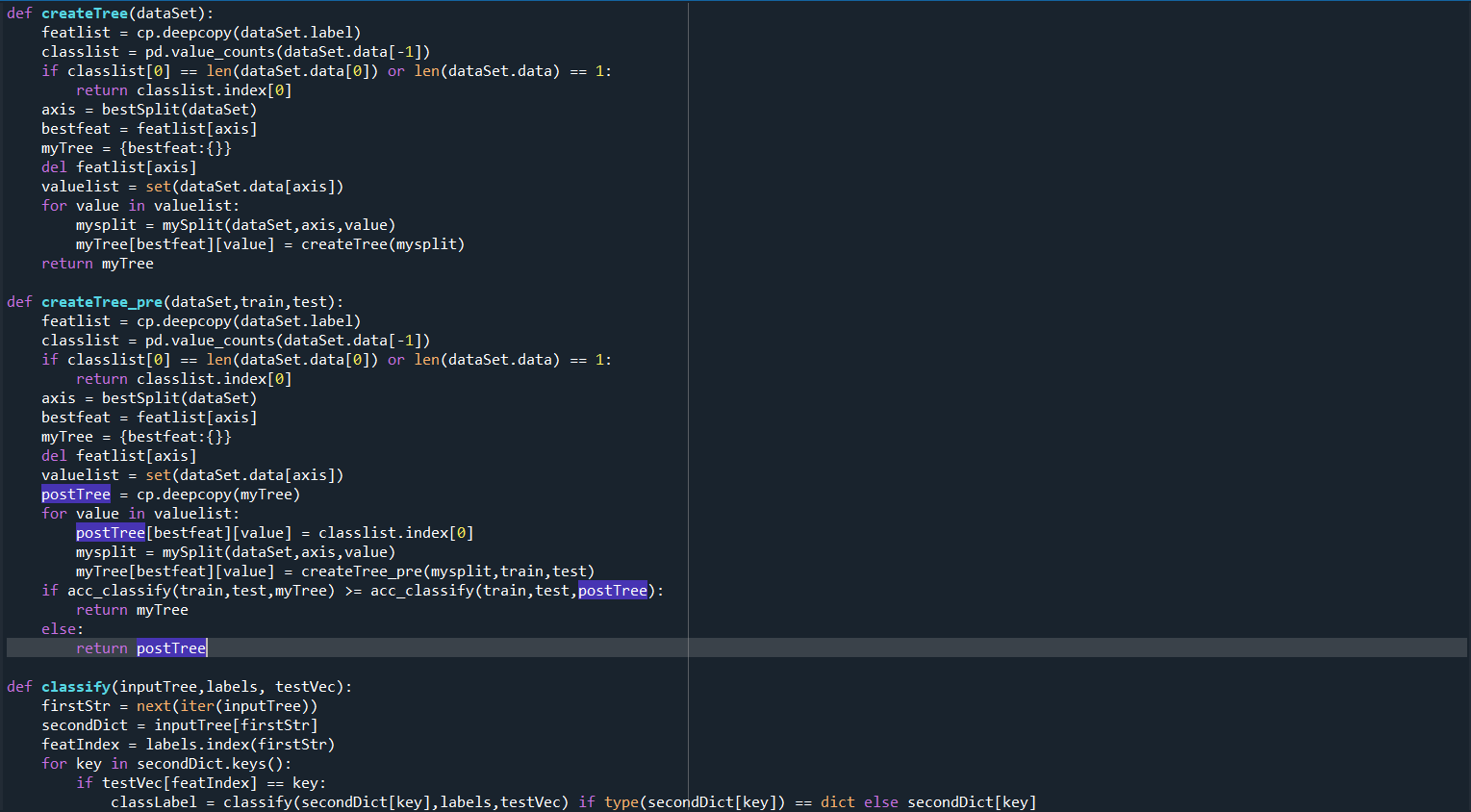
如图所示，每个数据集类由类标签和数据两部分组成，数据部分为矩阵形式，每一行对应一个标签，最后一行为类型。此类实现了fish,lenses和iris数据集的建立。

(2)针对离散数据，编写计算信息熵的函数，以及最佳分类和根据最佳分类划分数据集的函数



calEnt函数是求数据集信息熵的函数，通过选取最后一行的类型计算；mySplit通过传入的标签索引与标签值，将数据集划分成标签对应取值等于标签值的子集；bestSplit则选出最大信息增益的标签，返回索引。

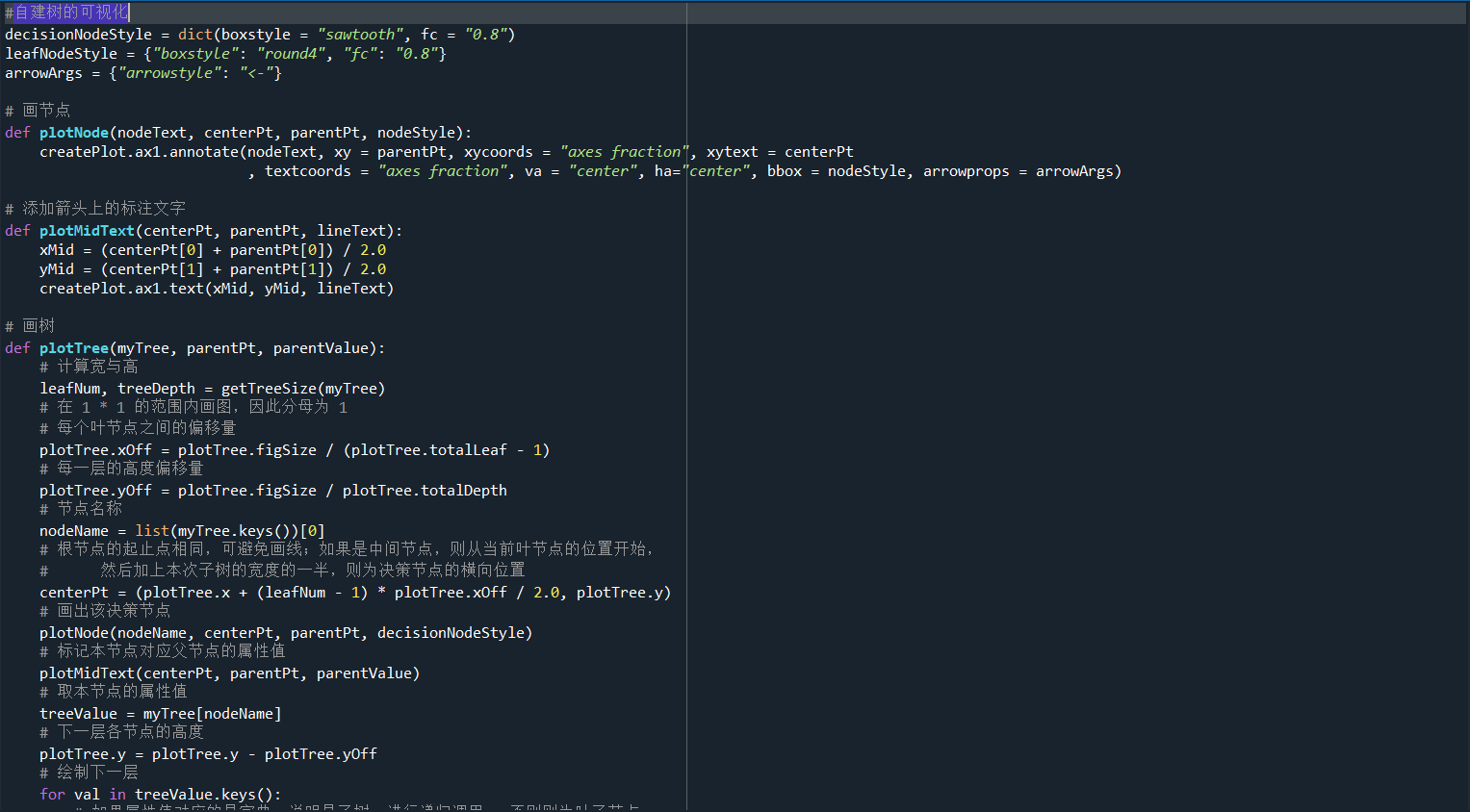
(3)编写离散值条件下建树，以及求预测准确率的函数，同时编写预剪枝建树函数

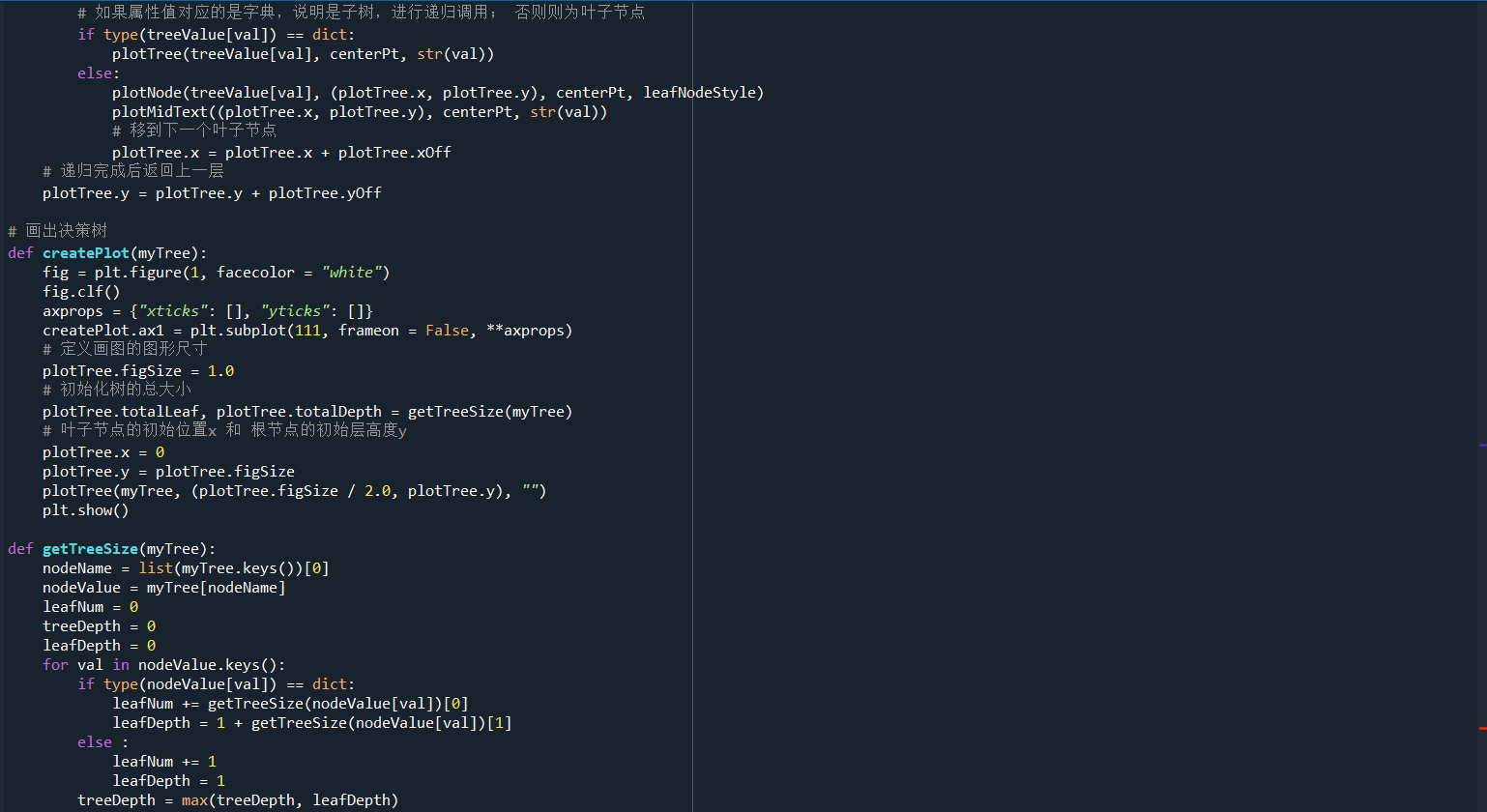


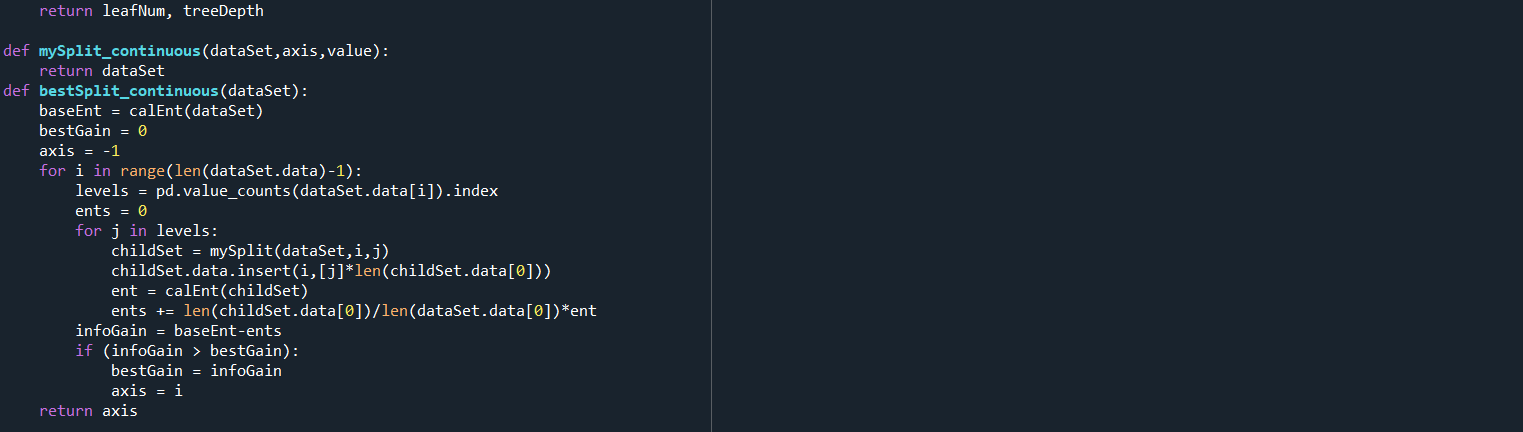


createTree函数通过递归，对每个子结点进行循环调用函数来建树；classify是根据决策树和标签对测试集数据进行分类预测的函数；而acc\_classify通过比较预测和实际错误的数目来计算正确率；createTree\_pre函数是预剪枝建树函数，每次比较扩展前和扩展后的预测准确率来决定是否扩展，为了兼容预剪枝建树，acc\_classify函数增加了可变Tree参数，用来计算不完全树的预测准确率。

(4)针对自主生成的决策树实现可视化

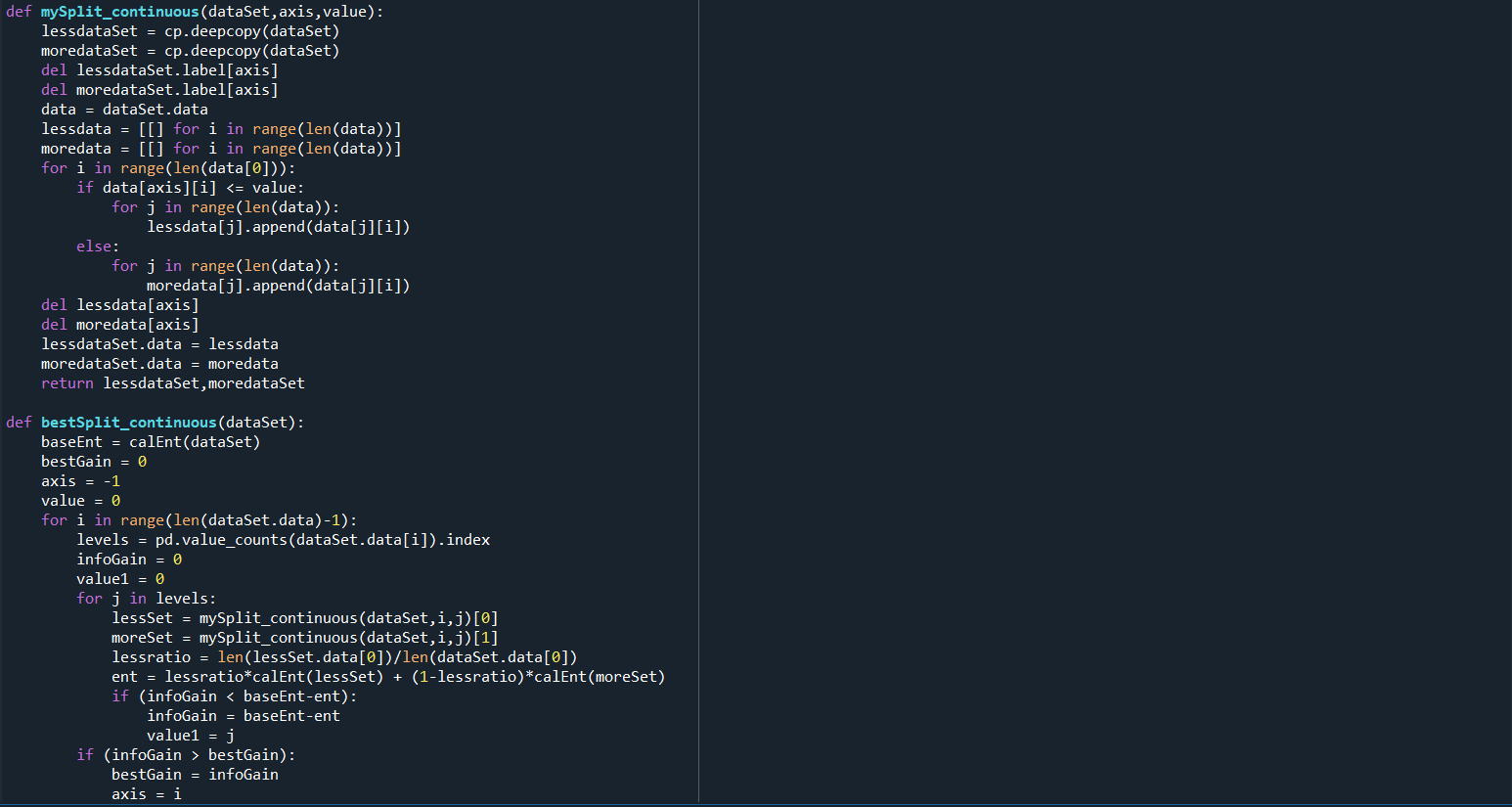


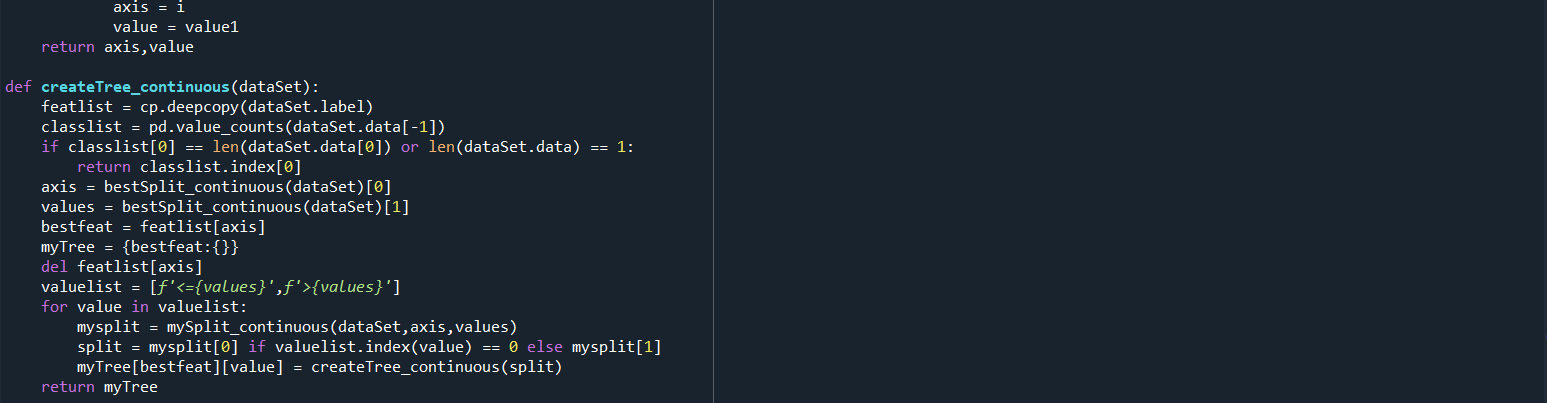




通过对以嵌套列表类型储存的树建立树的模型，包括结点，判断条件等等，最后通过matplotlib.pyplot库进行输出

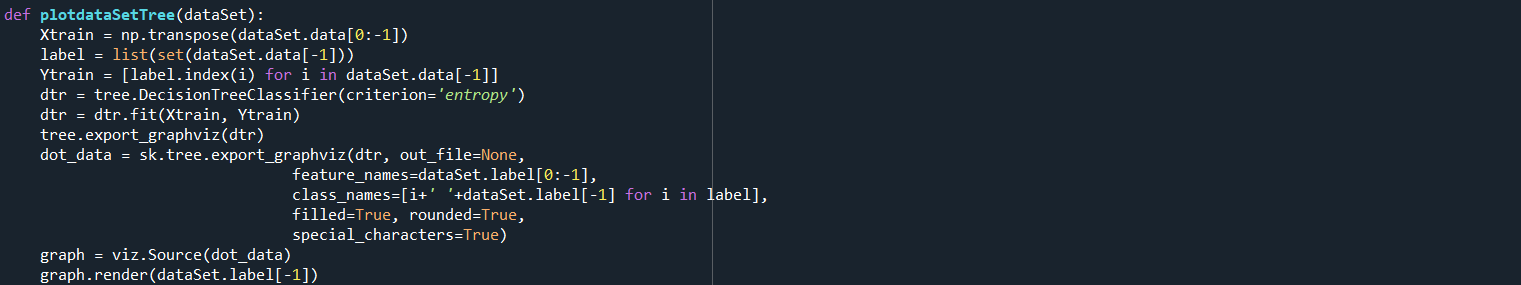
(5)针对iris数据集的连续性编写连续性数据处理建树函数





连续数据处理使用了二分类方法，mysplit将数据集根据特征分为大小两部分；bestsplit则计算总熵返回最佳分点索引和取值；createTree则对二分类树分别递归建树，并且将标签修改为大小于。

(6)编写基于sklearn建立决策树并可视化的函数



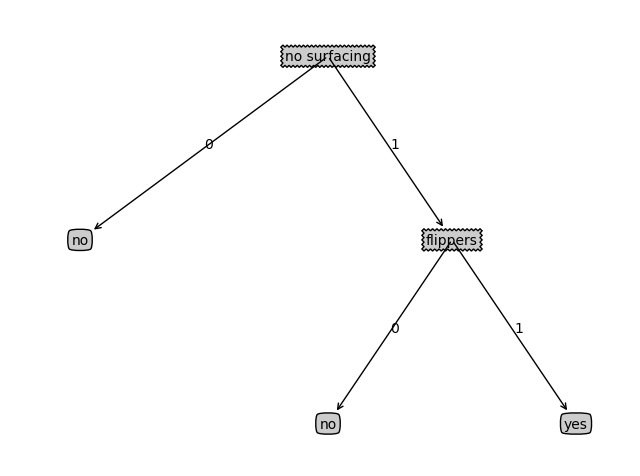
这个函数针对任何数据集均有效，它提取数据集的数据，并提取数据集的标签生成类标签与分类名称。

四、实验结果及分析

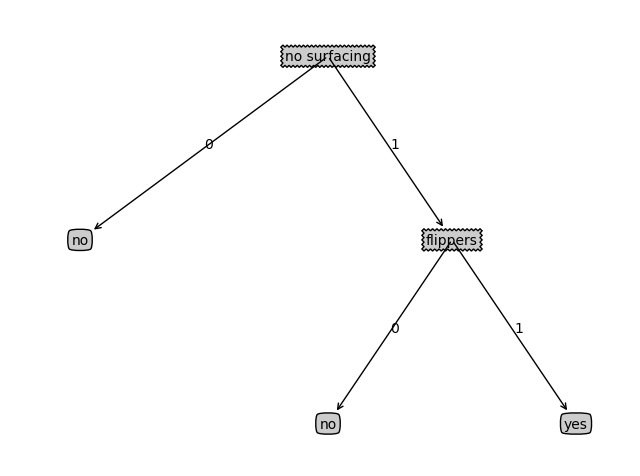
(1)自建决策树与准确率

Fish：

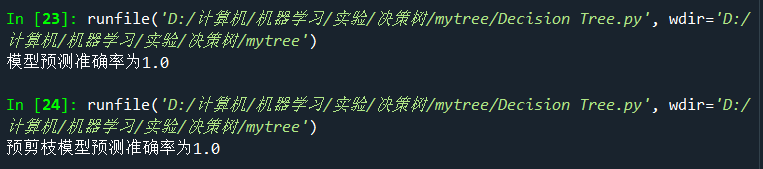
完全树:



剪枝后：

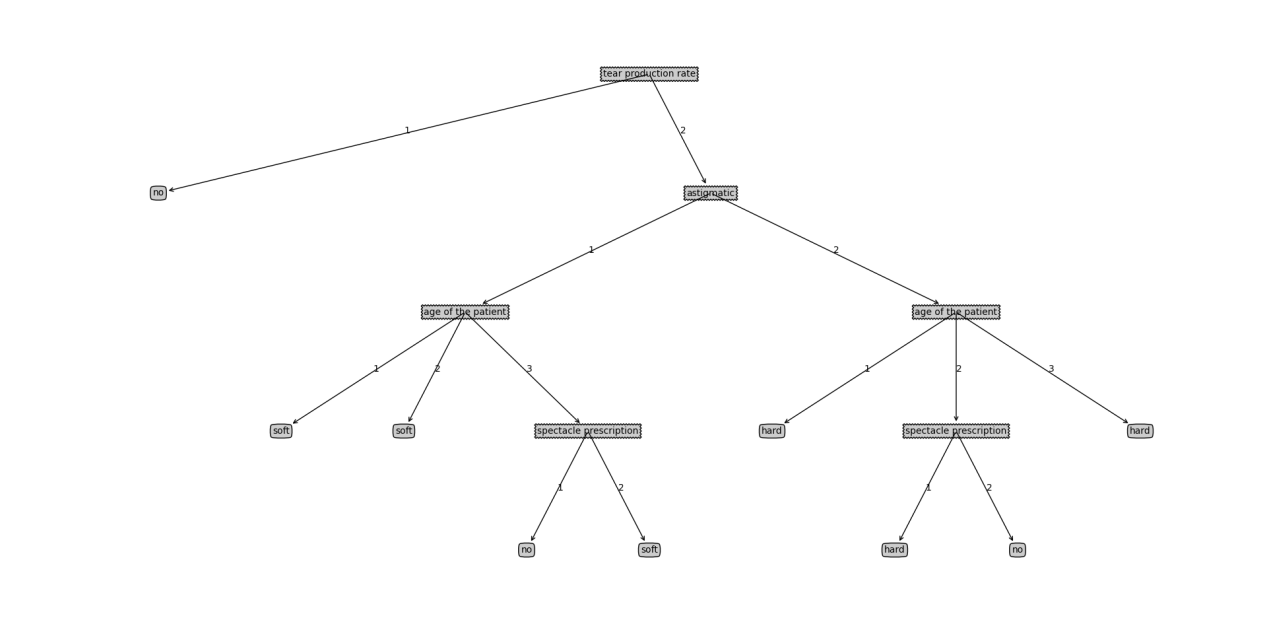


准确率:

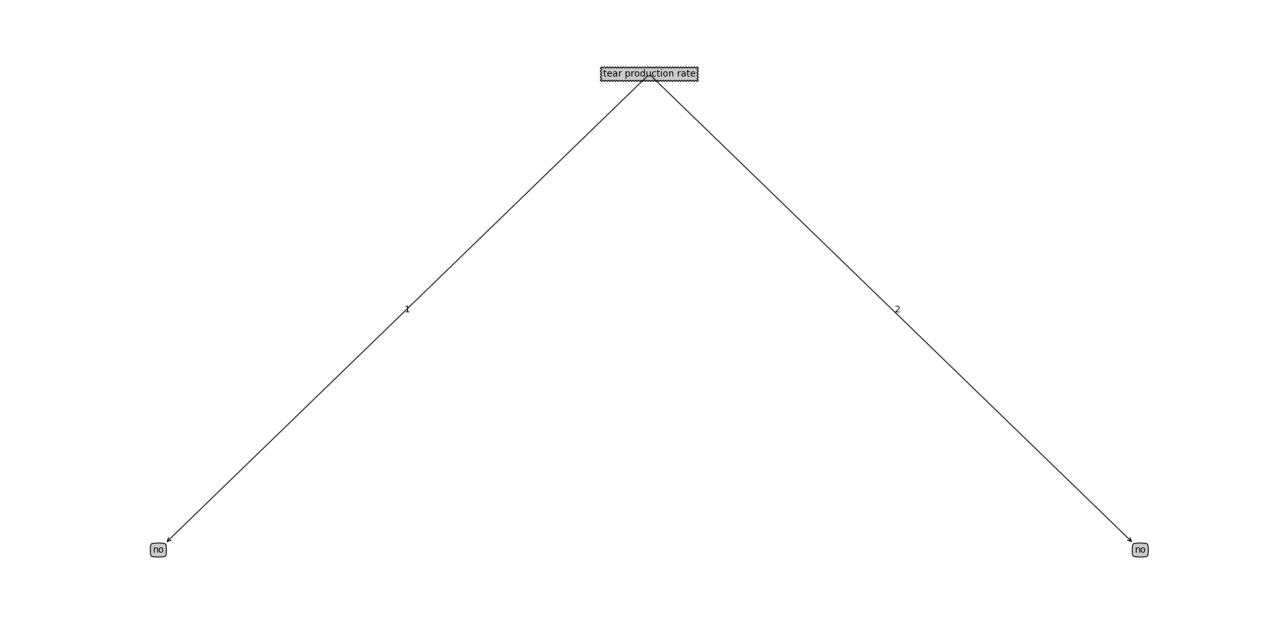


Lenses:

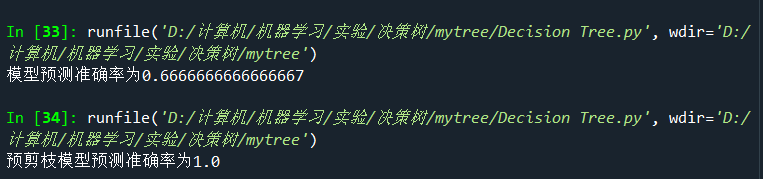
完全树：



剪枝后：

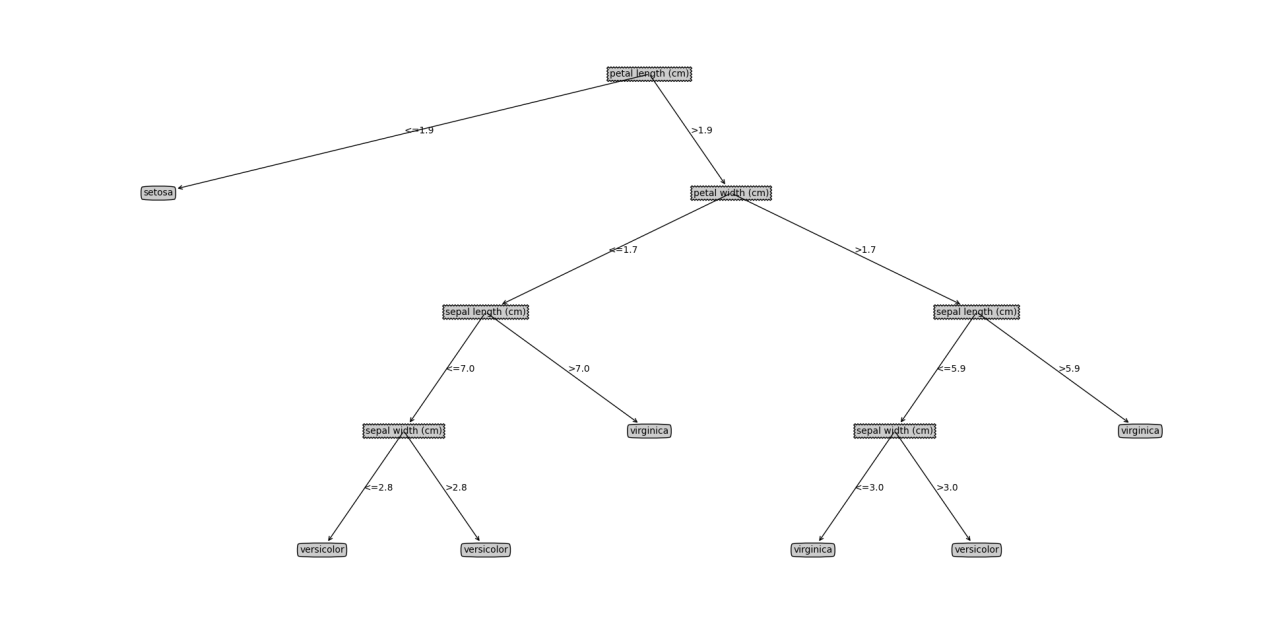


准确率;

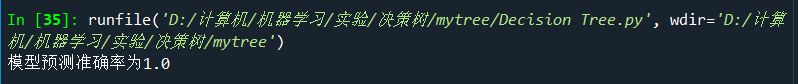


Iris：

完全树：

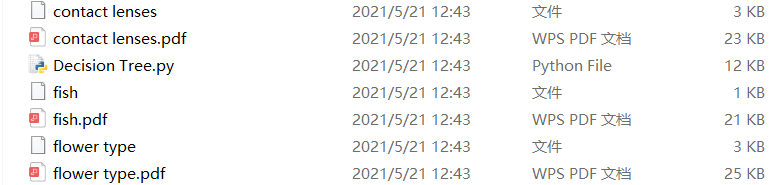


准确率：

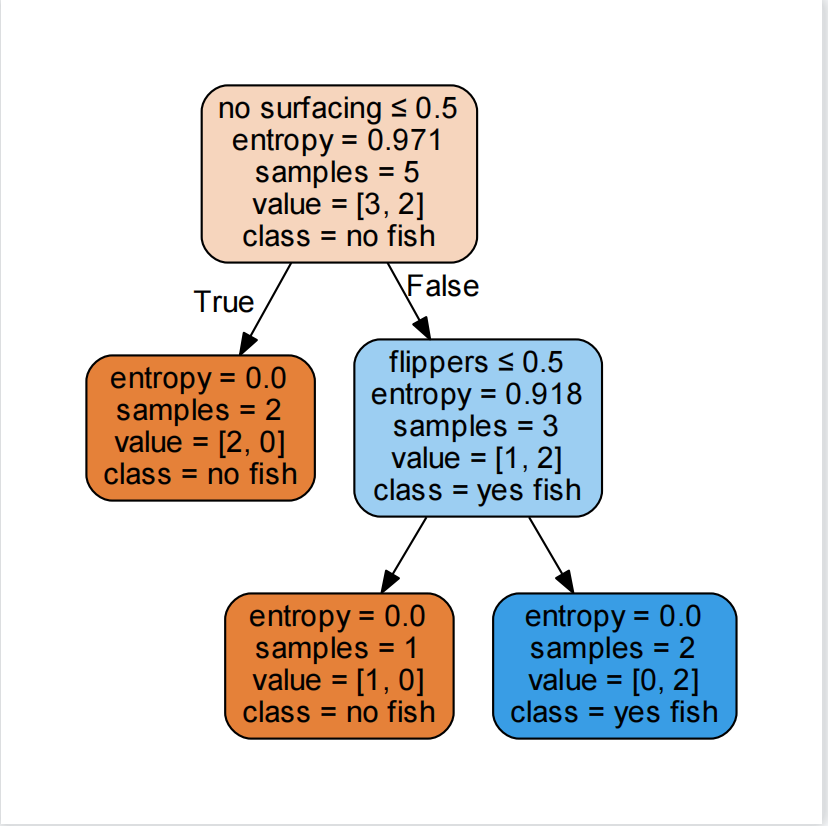


(2)基于sklearn建树并可视化结果

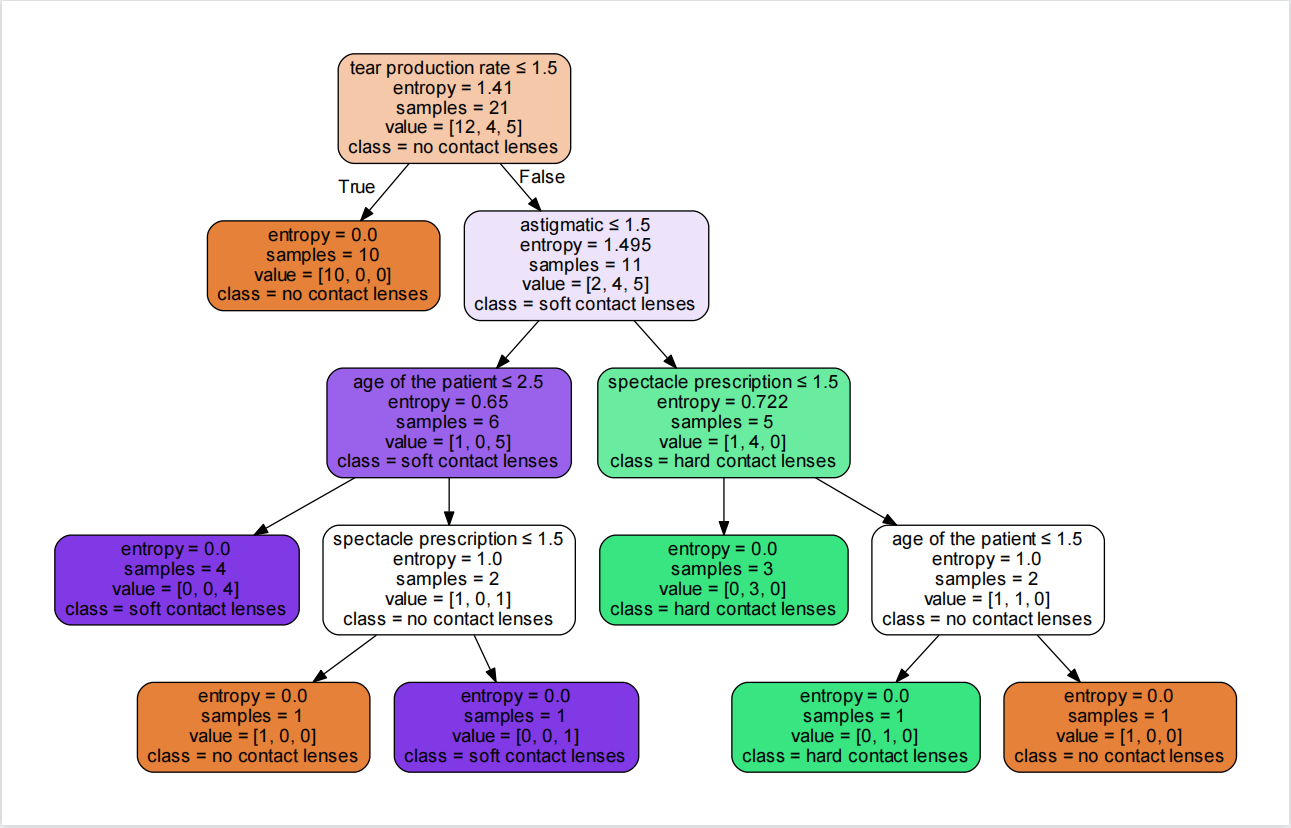
生成文件：



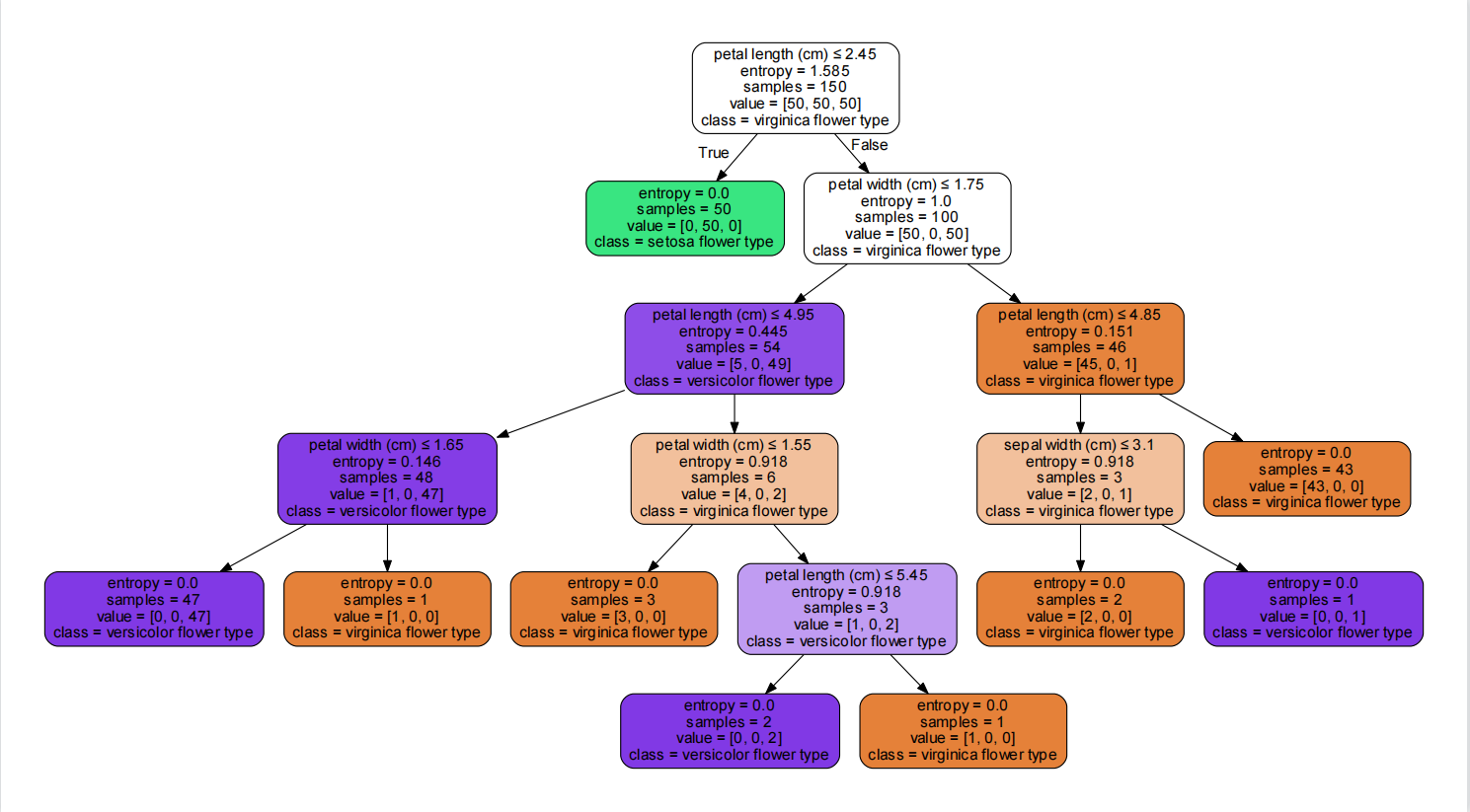
Fish:



Lenses：

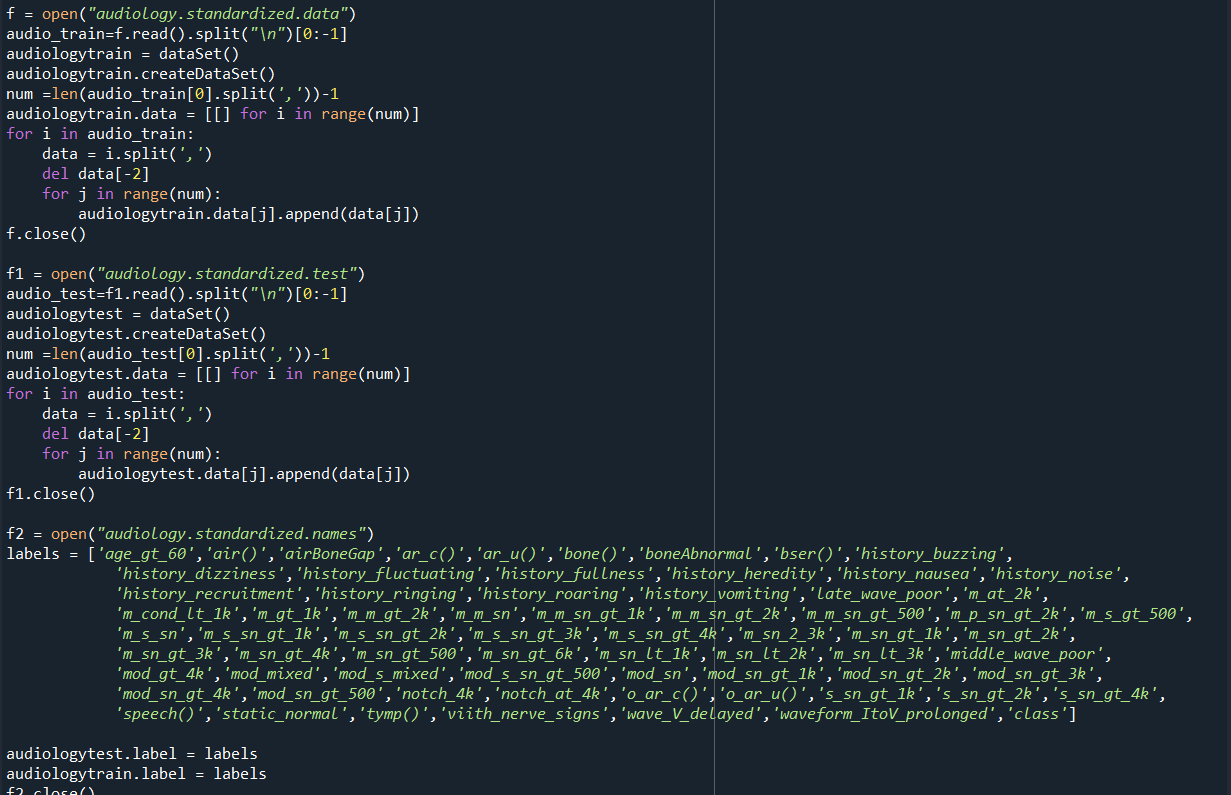


Iris:

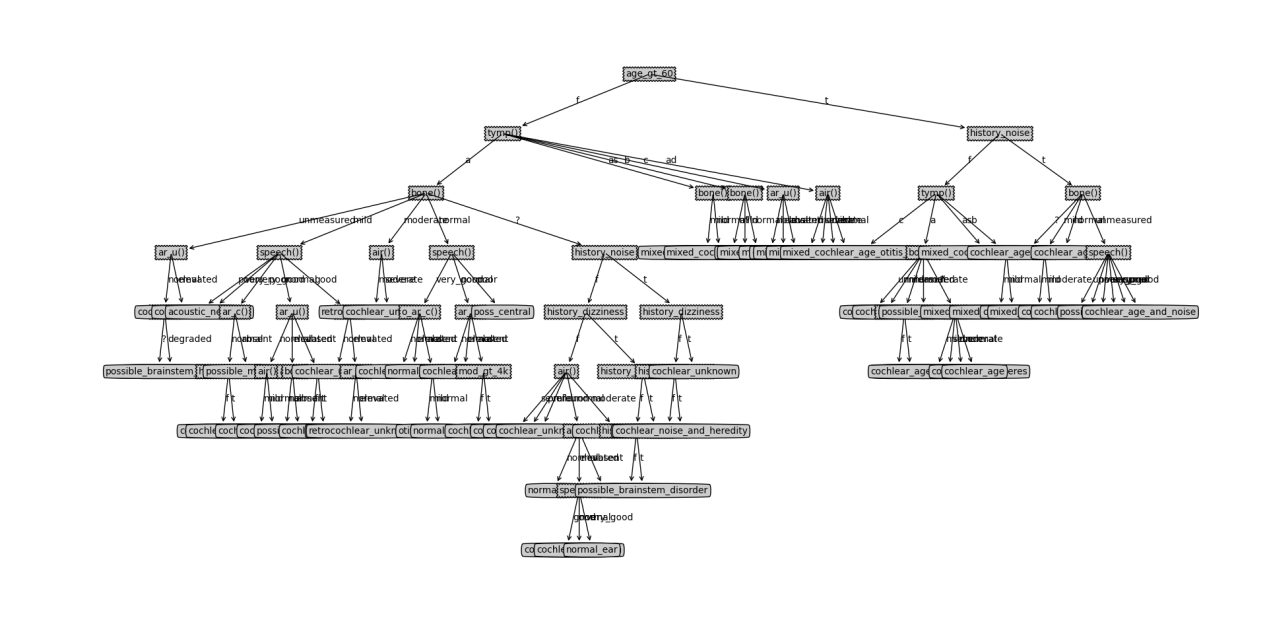


补：Audiology数据集

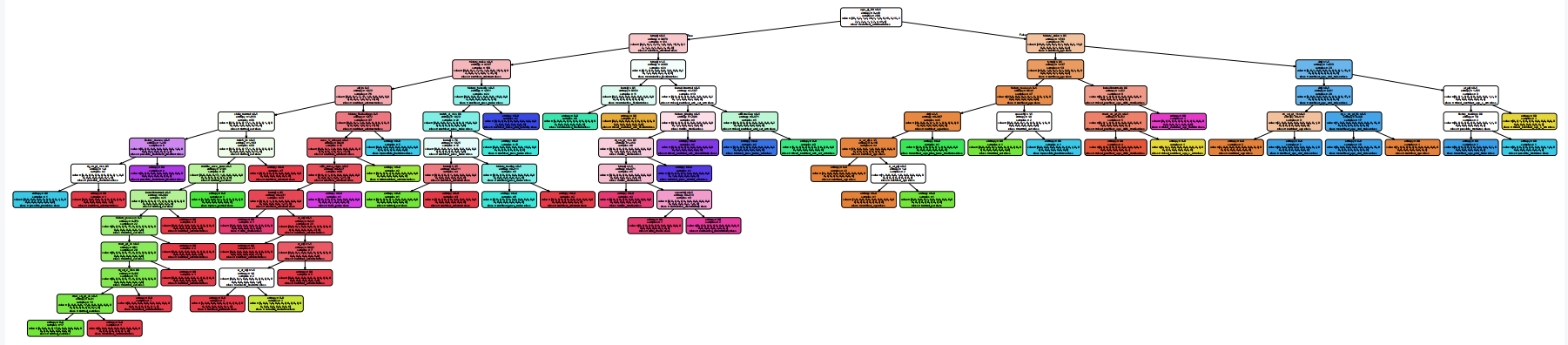
全代码：



完全树：



Sklearn树：



准确率：

