计算机与信息工程学院实验报告（一）

姓名：王赫 学号：1828070097 专业：数据科学与大数据技术(明德计划) 年级：2018级

课程：机器学习与数据挖掘 主讲教师：罗慧敏 辅导教师：\_\_\_\_\_\_\_

实验题目： 作业2

实验目的： 基于信息熵决策树构建（西瓜数据集3.0a）

剪枝决策树构建及显著性检验（西瓜数据集3.0a）

实验环境（硬件和软件） Anaconda/Jupyter notebook/Pycharm

实验内容：

课后习题：

基于信息熵决策树构建（西瓜数据集3.0a）

剪枝决策树构建及显著性检验（西瓜数据集3.0a）

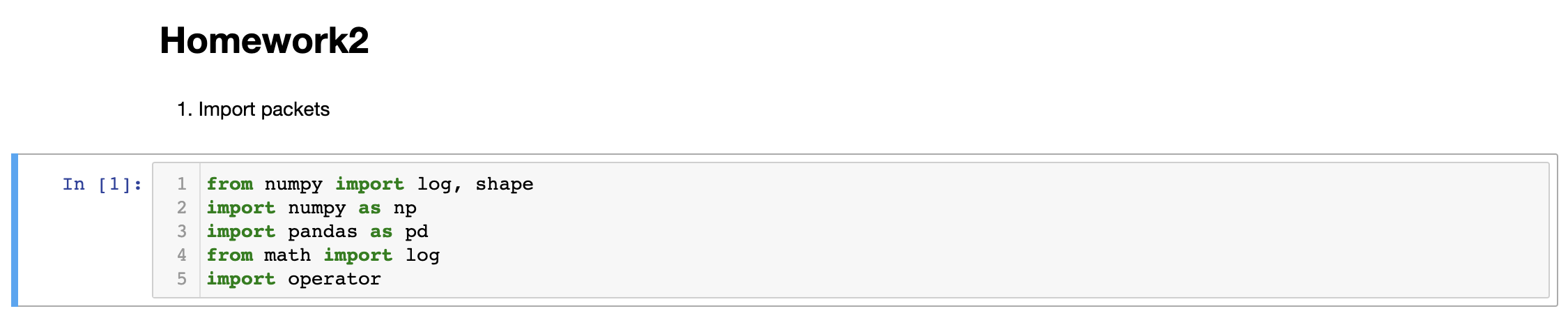
实验步骤：

1. 数学公式推导
2. 编程实现
3. 数据记录
4. 实验总结

实验数据记录：

题目1:

1. 导入依赖包



1. 计算信息熵



1. 离散数据集划分



1. 连续数据集划分



1. 选择最优



这里对连续性数据的处理方案是：首先将连续性数据取相邻两两的数据取平均值，然后对处理后的数据按照离散型数据进行处理，计算信息熵即可



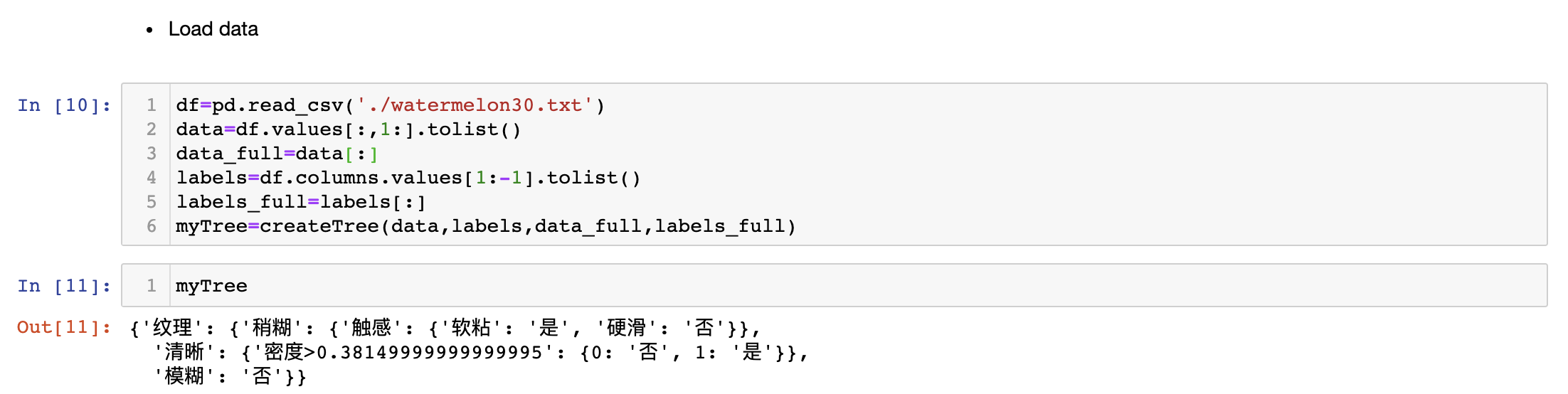
1. 投票



1. 创建树



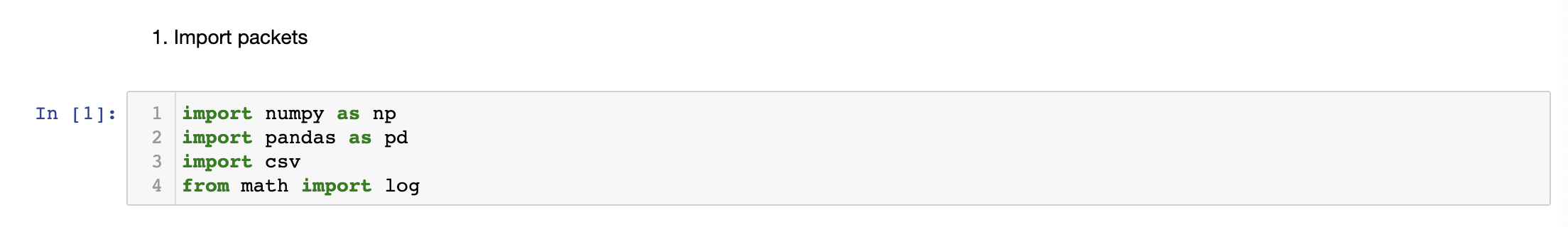
1. 读本地数据，输出创建的树



题目2:

本题主要是创建剪枝前后不同的树

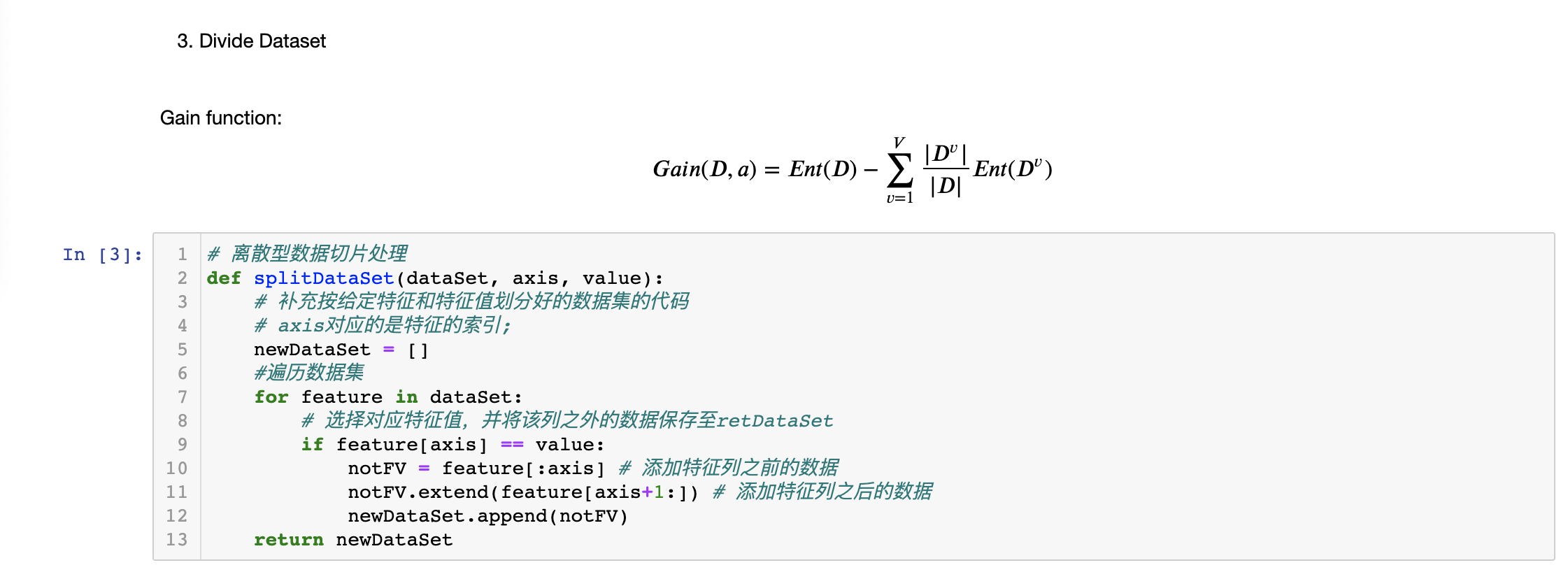
1. 导入依赖包



1. 计算信息熵



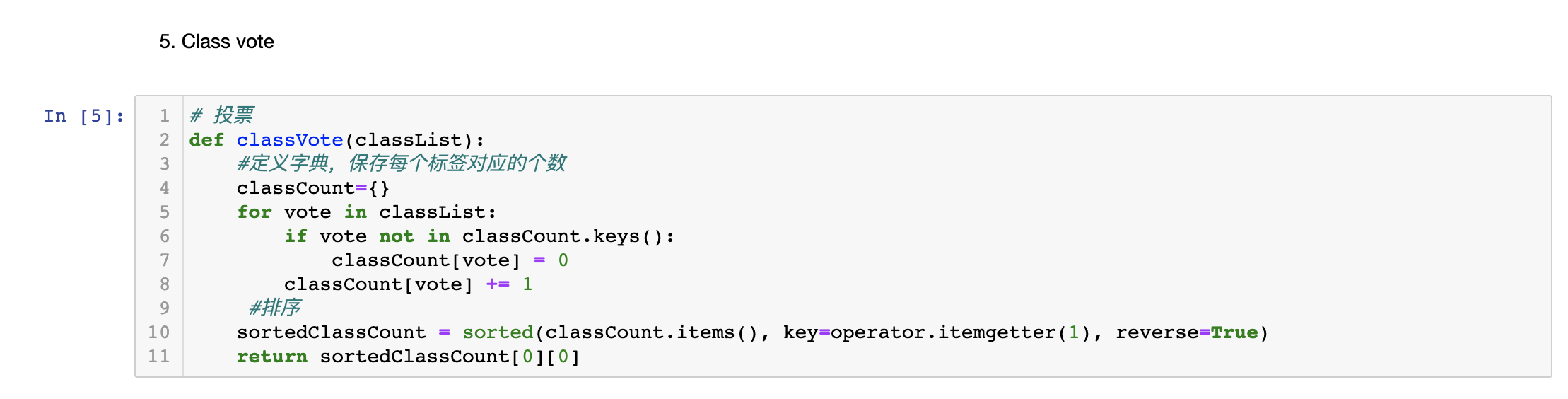
1. 离散型数据划分



1. 最有特征选择



1. 类别投票



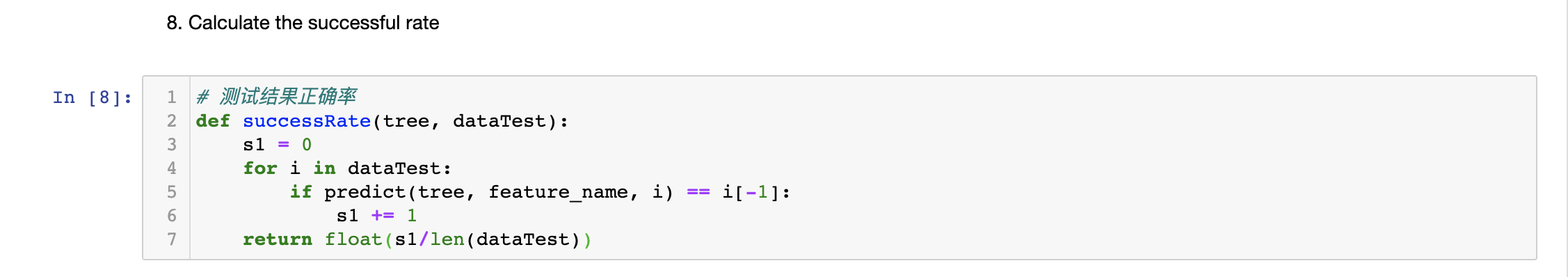
1. 无剪枝创建树



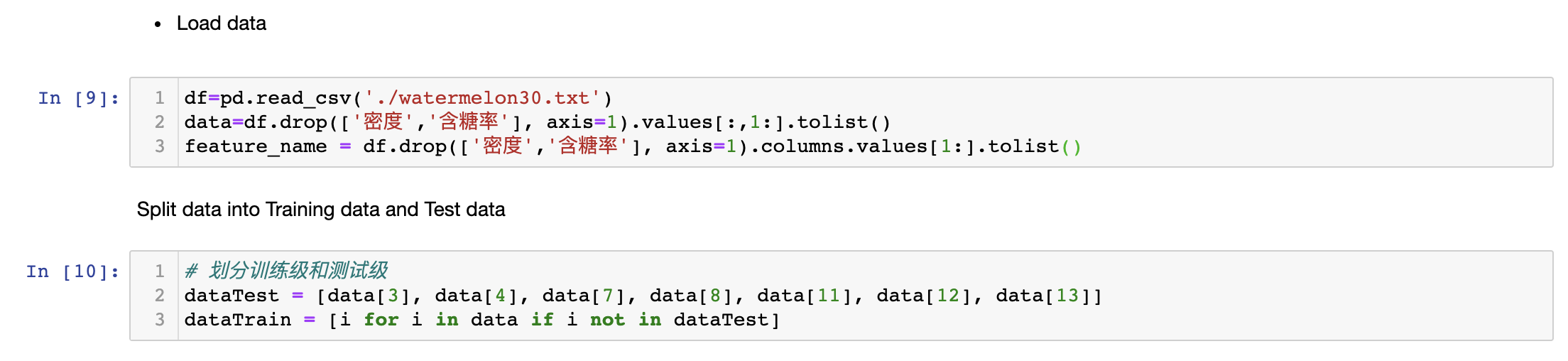
1. 预测函数



1. 准确度计算



1. 获取本地数据



1. 未剪枝决策树创建



1. 剪枝决策树创建
   1. 预剪枝决策树创建：

预剪枝决策树构建函数



创建结果

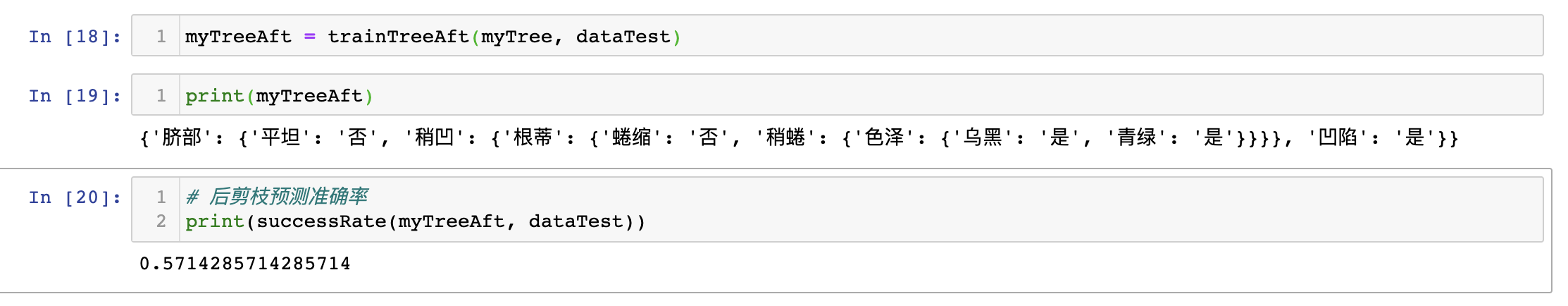


* 1. 后剪枝决策树创建

后剪枝决策树构建函数



创建结果



这里总结一下创建预剪枝和后剪枝决策树的方法：预剪枝创建决策树时，可以直接在创建过程中进行判断，最后返回准确率高的决策树即可；但是在创建后剪枝决策树时，需要首先创建好决策树，然后再对每一层的枝节点进行判断是否需要剪枝，计算量方面会比预剪枝决策树稍微大一点。