**电子科技大学计算机科学与工程学院**

**标 准 实 验 报 告**

**（实验）课程名称 数据结构与算法**

**电子科技大学教务处制表**

**电 子 科 技 大 学**

**实 验 报 告**

**学生姓名：陶浩轩 学 号：2023080902011 指导教师：陈端兵**

1. **实验室名称：**

学知三组团10栋139

1. **实验项目名称：**

基于决策树的分类与随机森林

1. **实验原理：**

决策树是一种常用的分类算法，其原理是利用信息增益或信息增益比等指标选择最优特征进行划分，将数据集逐步划分为不同的子集，直到每个子集都属于同一类别或无法继续划分。ID3算法是决策树的一种实现方式，它使用信息增益作为特征选择的依据。随机森林则是由多个决策树构成的集成学习模型，通过随机选择特征和样本数据构建多棵决策树，并对它们的预测结果进行投票表决，从而提高分类的准确率和泛化能力。

1. **实验目的：**

理解决策树和随机森林的基本原理和实现方式，掌握决策树和随机森林在分类中的应用。

1. **实验内容：**

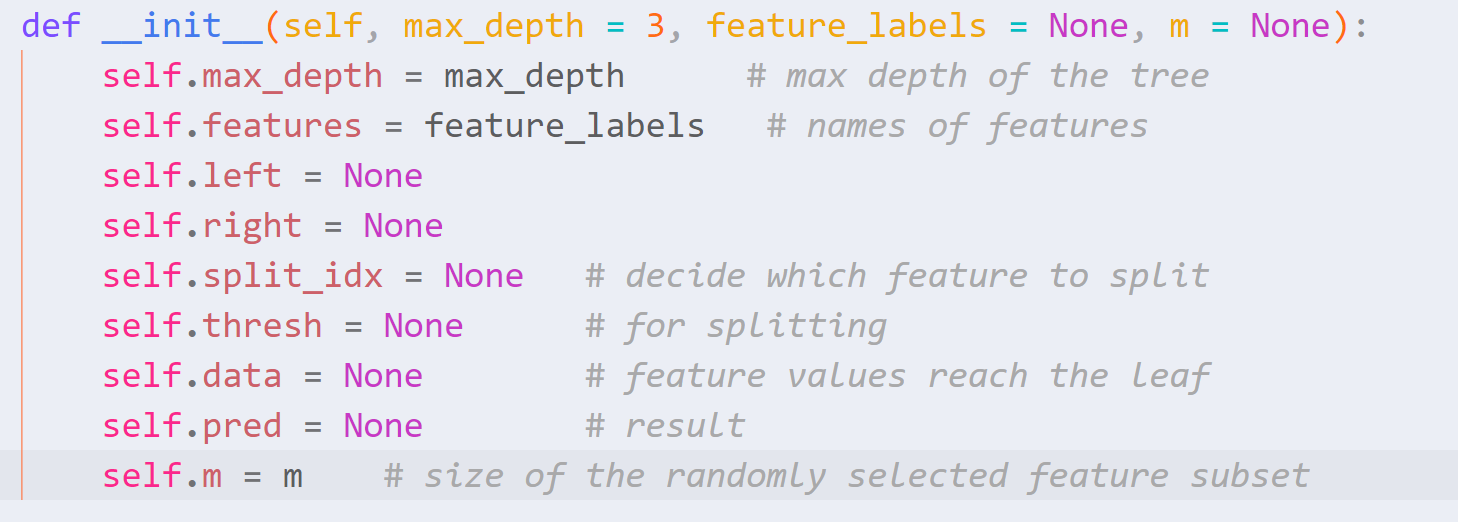
使用ID3算法构建决策树与随机森林，对鸢尾花数据集进行分类，并评估其准确率。

1. **实验器材（设备、元器件）：**

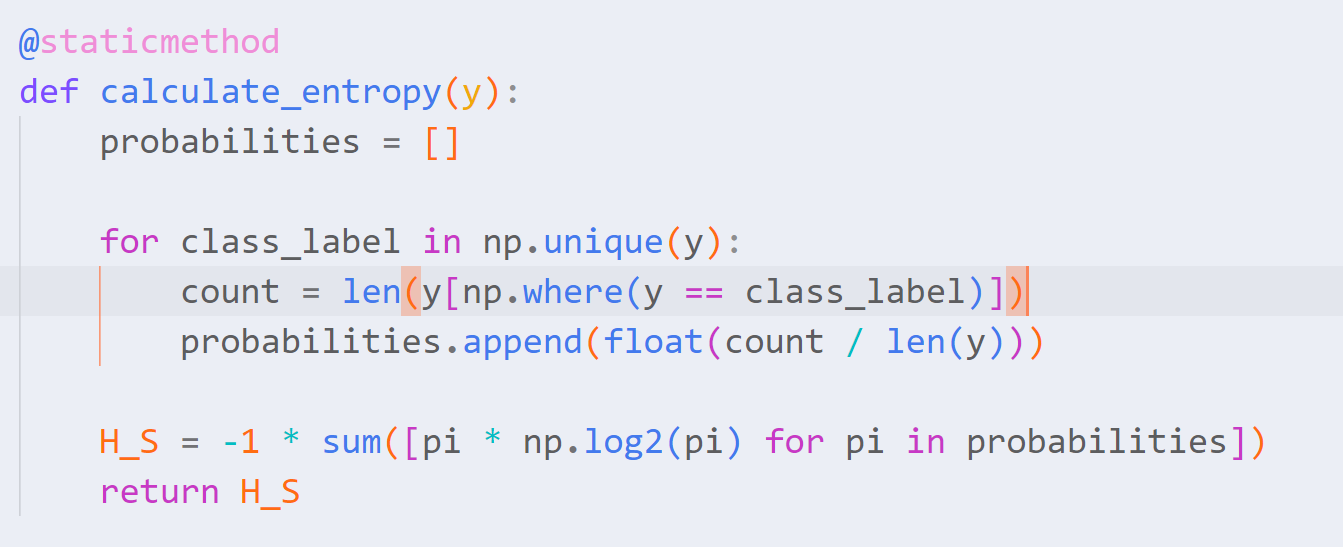
Window10，Python 3.12.6，NumPy 2.1.3

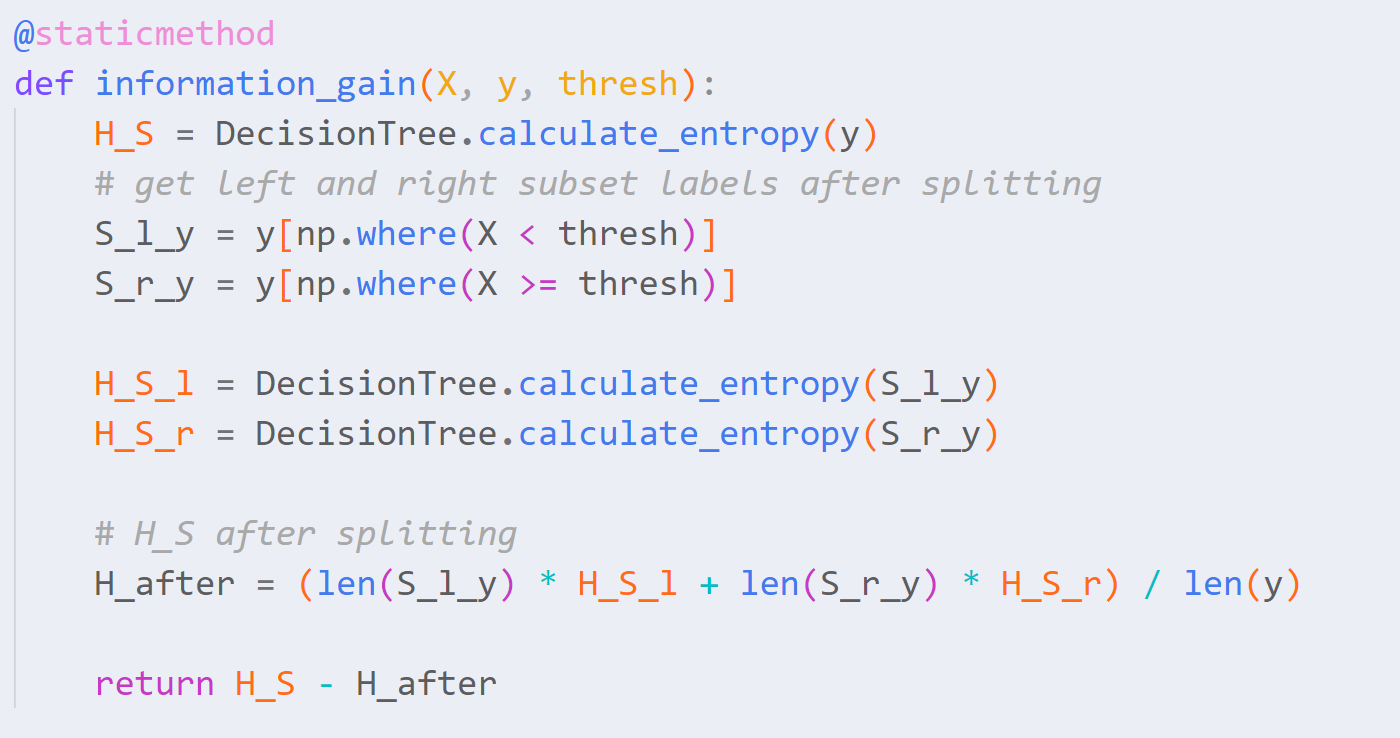
**七、实验步骤：**

1. 构建 DecisionTree 类，设置最大深度max\_depth，特征标签feature\_labels，左右子树left和right，分割特征索引split\_idx，分割阈值thresh，到达叶节点的数据data，叶节点的预测结果pred，以及随机特征子集的大小m

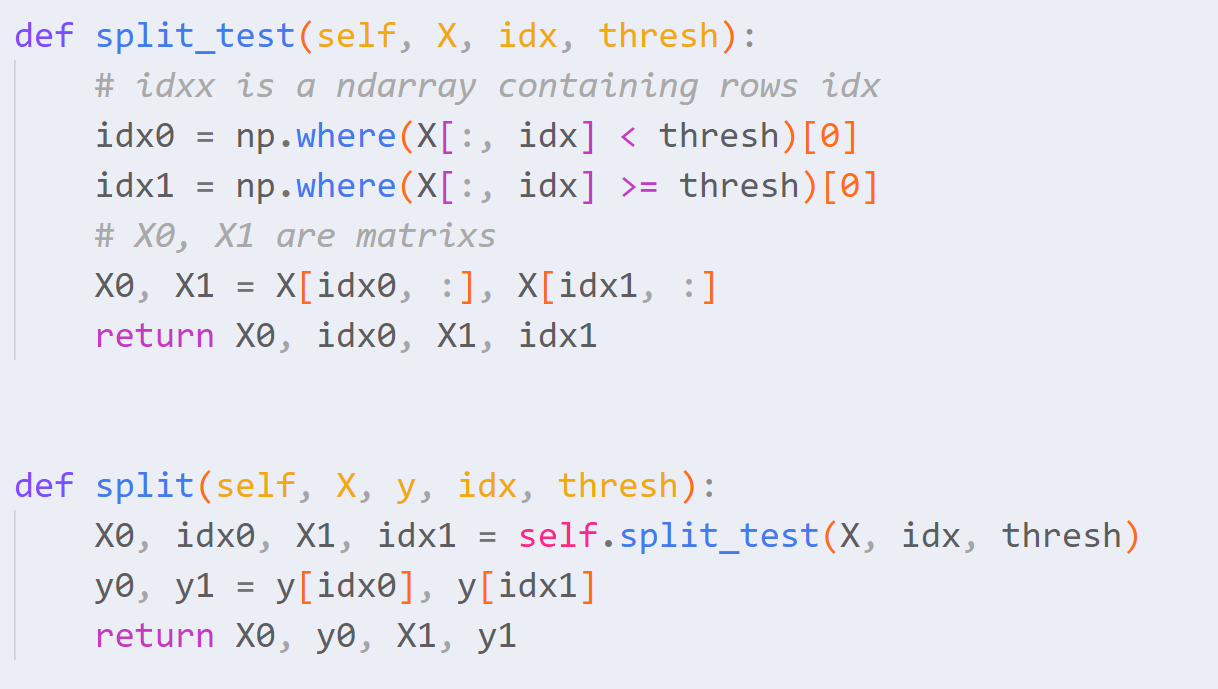
****

2. 使用公式计算信息熵，基于给定的特征和阈值计算信息增益

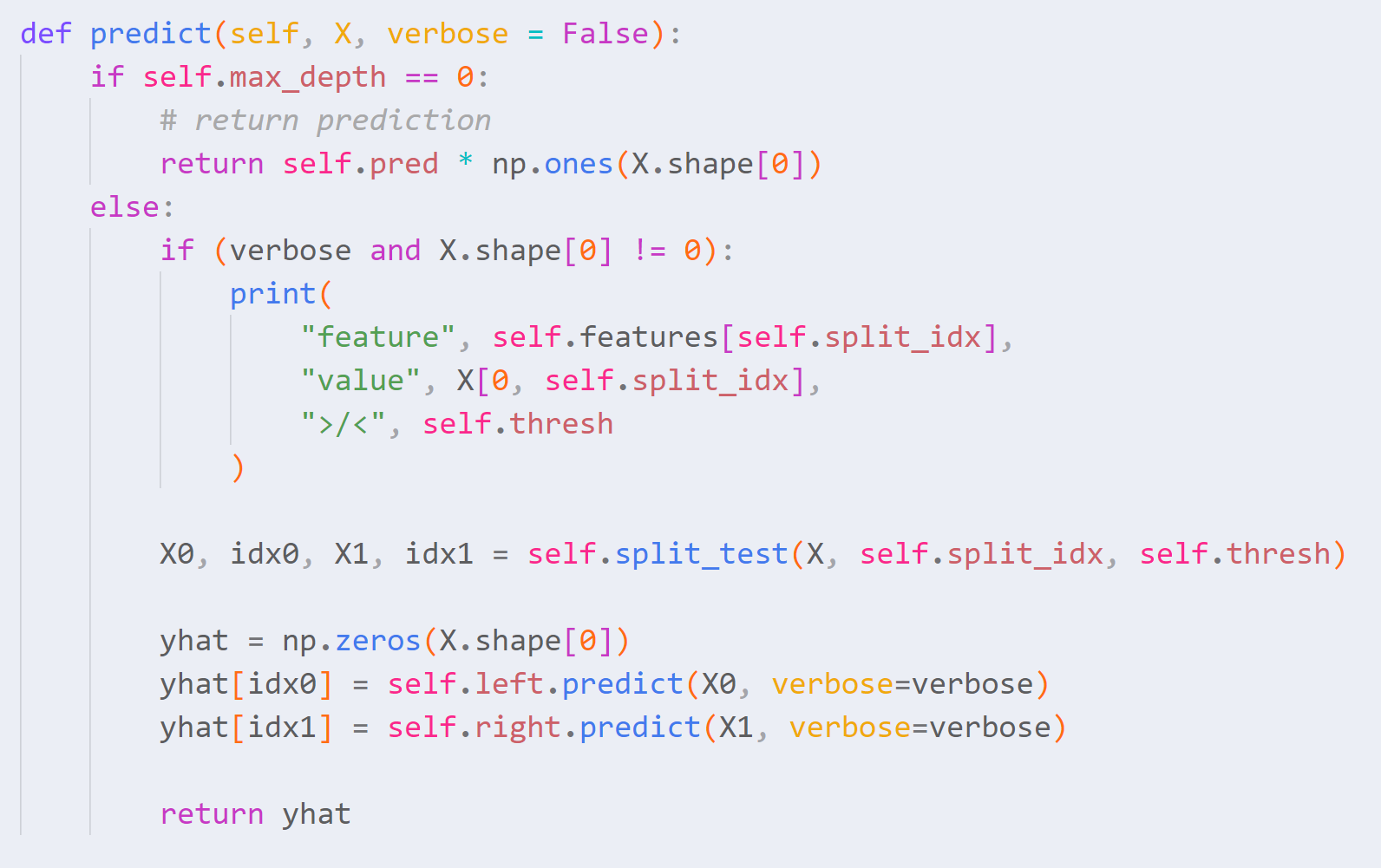
****

****

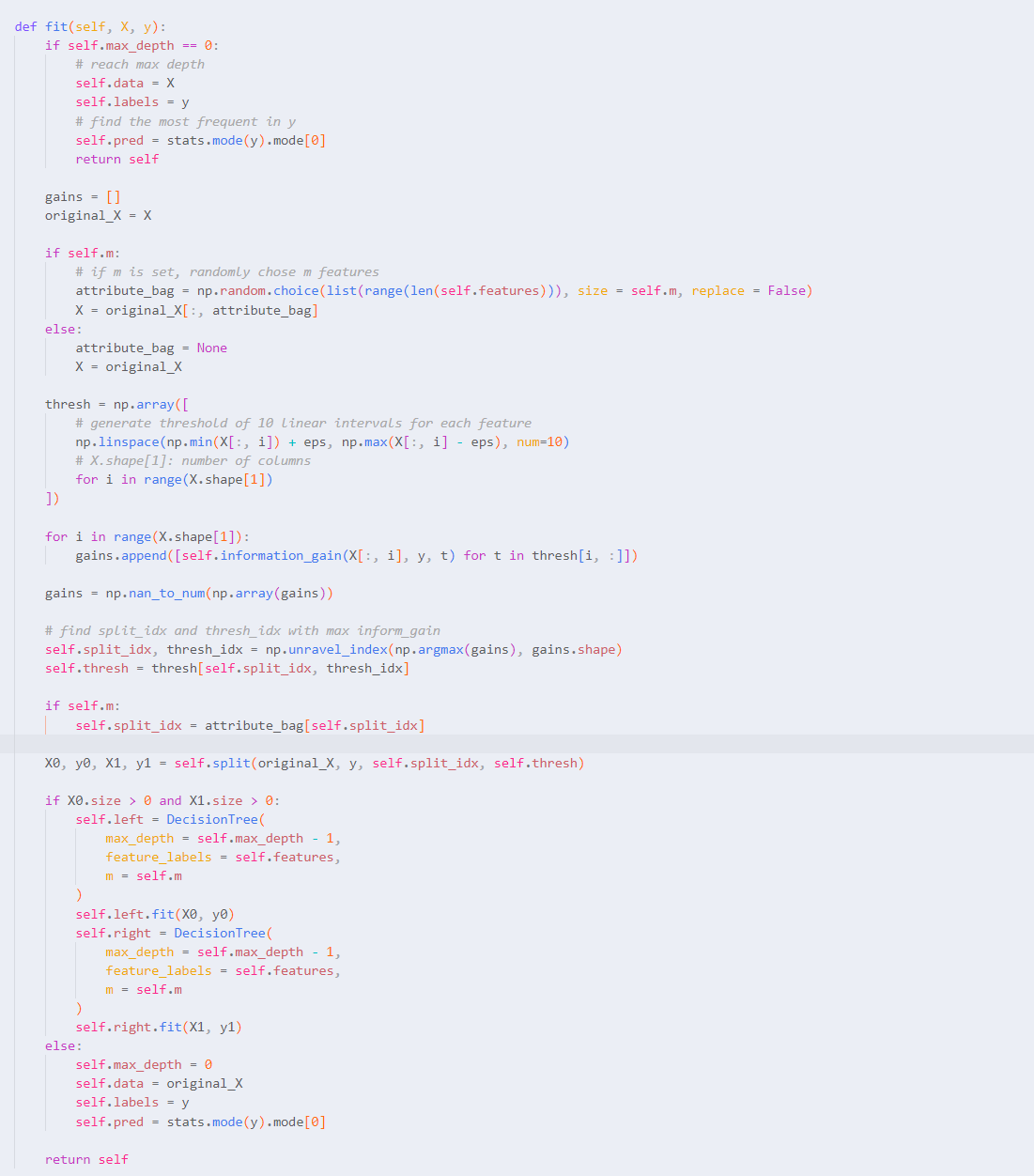
3. 编写 split\_test 方法，根据给定的特征索引和阈值进行数据分割；split 方法调用前者分割数据集，并返回分割后的特征和标签

****

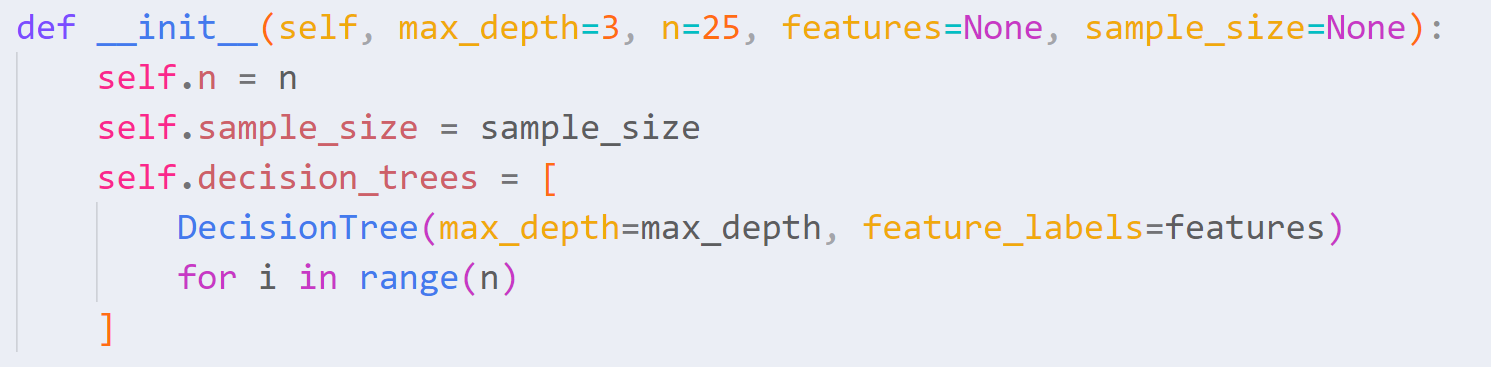
4. 实现 predict 方法，递归地遍历树，根据特征值和阈值决定向左还是向右子树进行预测，直到达到叶节点。用于对数据的分类进线预测

****

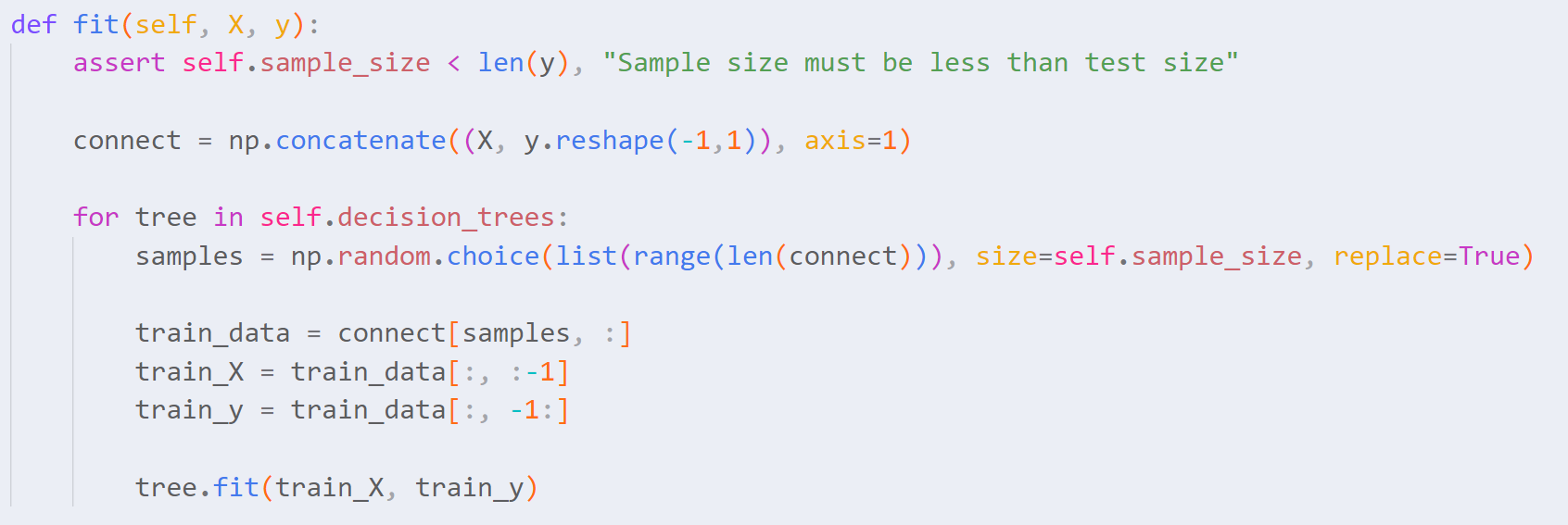
5. 完成训练决策树模型的方法 fit，递归地选择最优特征和阈值进行分割，直到达到最大深度或数据无法进一步分割。如果设置了随机特征子集的大小m，则随机选择特征进行分割

****

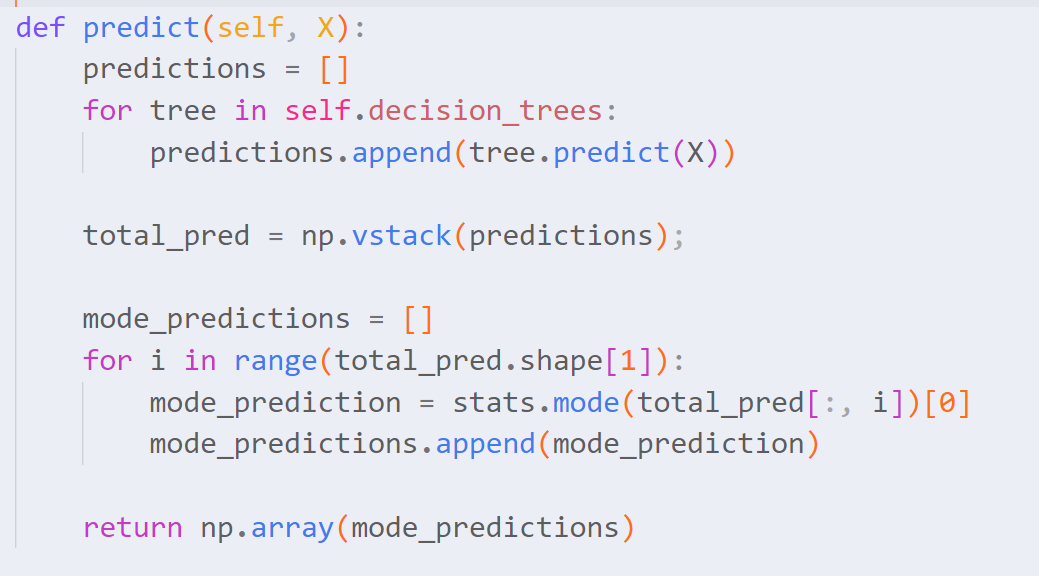
6. 在实现决策树后，编写随机森林类 RandomForest。该类初始化方法设置了森林中树的数量n，样本大小sample\_size，并为每棵树创建一个DecisionTree实例

****

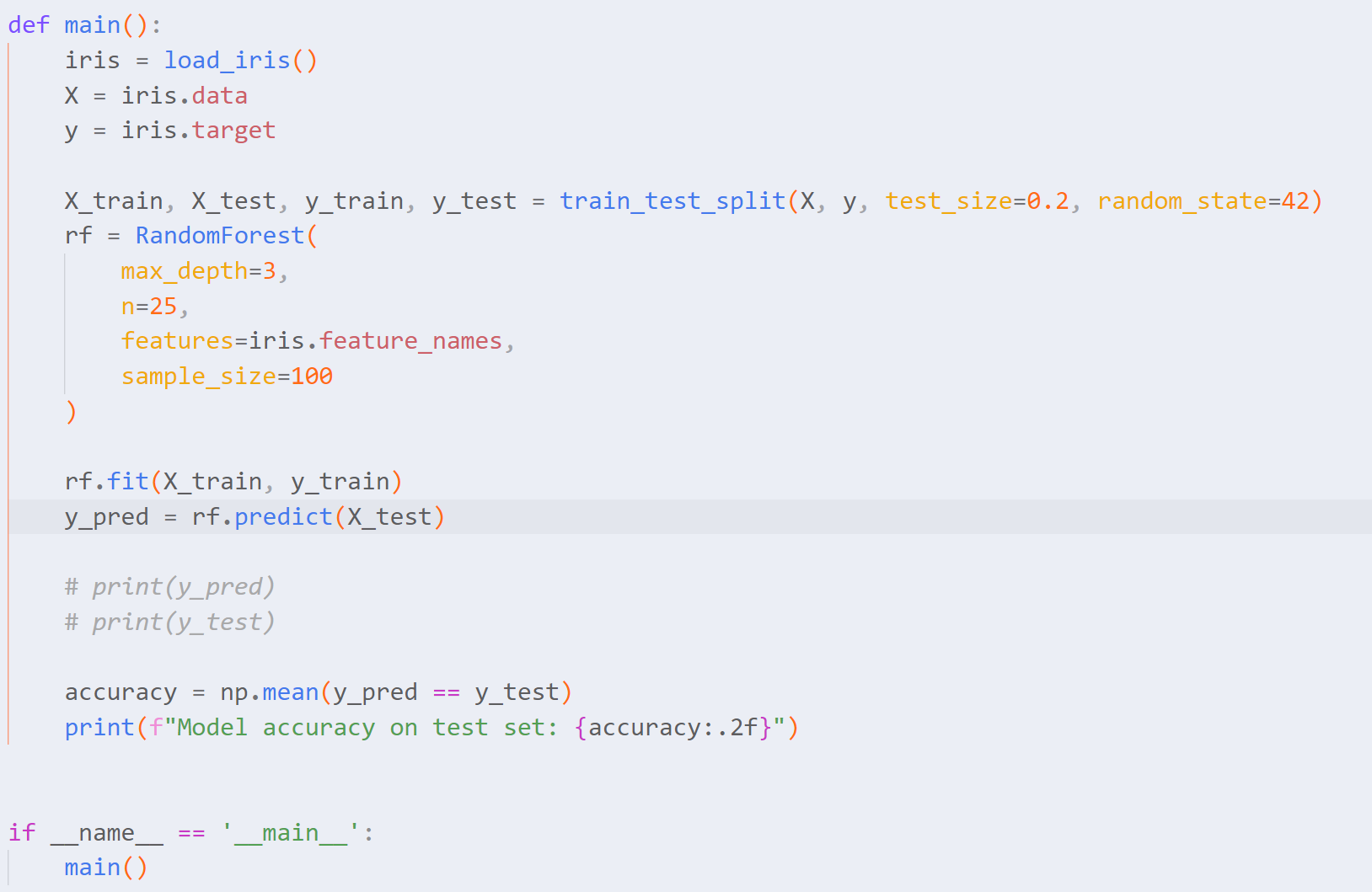
7.实现训练随机森林模型的方法 fit，具体是对数据集有放回的随机抽样，并用抽样的数据集训练森林中的树

****

8. predict 方法，用森林中的每一棵树得出的预测结果，“投票”得出最终的预测

****

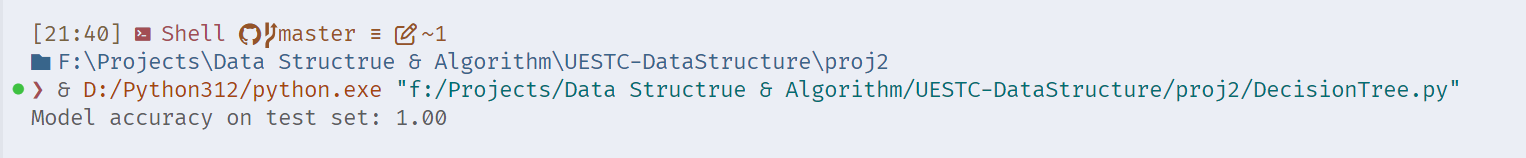
9. 最后编写主函数，加载鸢尾花数据集，并划分训练集与测试机。随后创建随机森林实例，训练模型。最后使用测试集进行预测和评估

****

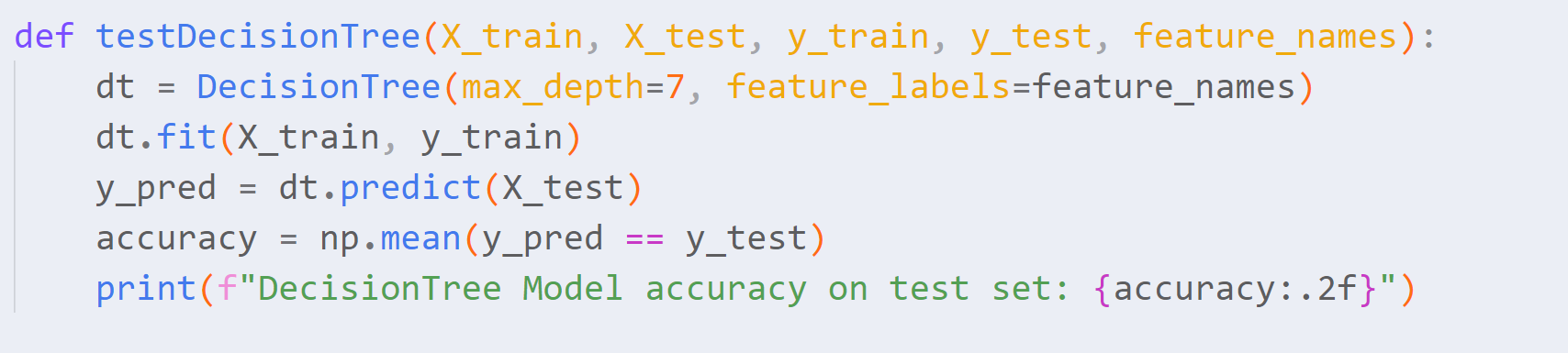
1. **实验数据及结果分析：**

本次实验采用了[鸢尾花数据集](https://zh.wikipedia.org/wiki/%E5%AE%89%E5%BE%B7%E6%A3%AE%E9%B8%A2%E5%B0%BE%E8%8A%B1%E5%8D%89%E6%95%B0%E6%8D%AE%E9%9B%86)，包含了150个样本，都属于鸢尾属下的3个亚属，分别是山鸢尾、变色鸢尾和维吉尼亚鸢尾。每个样本都包含4项特征，即花萼和花瓣的长度和宽度，它们可用于样本的定量分析

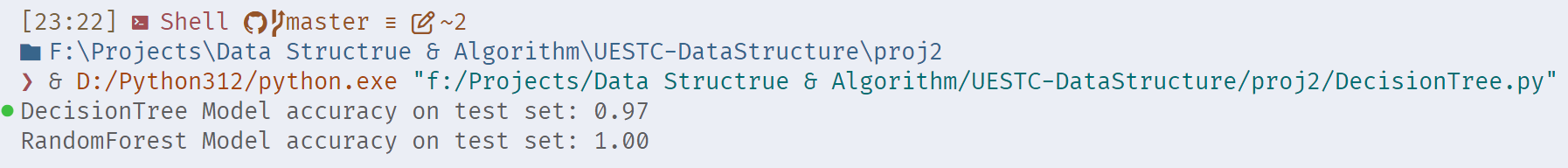
在 main 函数中，我将这 150 个样本中的 80% 用于训练随机森林，20% 用于测试。并使用 np.mean() 来求得预测的准确率，得到了理想的结果



接下来，我又用相同的数据集，添加了对决策树准确性的测试。



运行，得到决策树和随机森林预测的准确性。可以看到随机森林可以提供更准确的预测



1. **实验结论：**

通过本次实验，我们可以得出：决策树和随机森林是解决分类问题的两个可靠的方法。其中随机森林可以提供比决策树更加准确的预测。

1. **总结及心得体会：**
2. Python 在数据科学与机器学习方面提供了许多非常强大的库，比如 NumPy，SciPy，Sklearn 等等。在相关的工作中，比起其它语言，可以在保持程序高效运行的同时，提供更加便捷与快速的开发
3. 由于 Python 不像 C，Java 那样对数据类型进行限制，在写 Python 代码的时候应当更加注重方法与成员变量规范的编写，并在关键代码处使用 assert 实现 fast fail 方便 debug

**十一、对本实验过程及方法、手段的改进建议：**

1. 使用可视化工具。如 Matplotlib 和 Pandas，我们可以更加直观的观察决策树的构成，将“黑盒”变为“玻璃盒”
2. 可以尝试由一个节点，根据多个阈值范围分叉出多个节点。比较与二叉树的实现的准确性
3. 使用更多的数据集以及不同的划分方式，测试不同参数下的两个模型的准确性

**报告评分：**

**指导教师签字：**