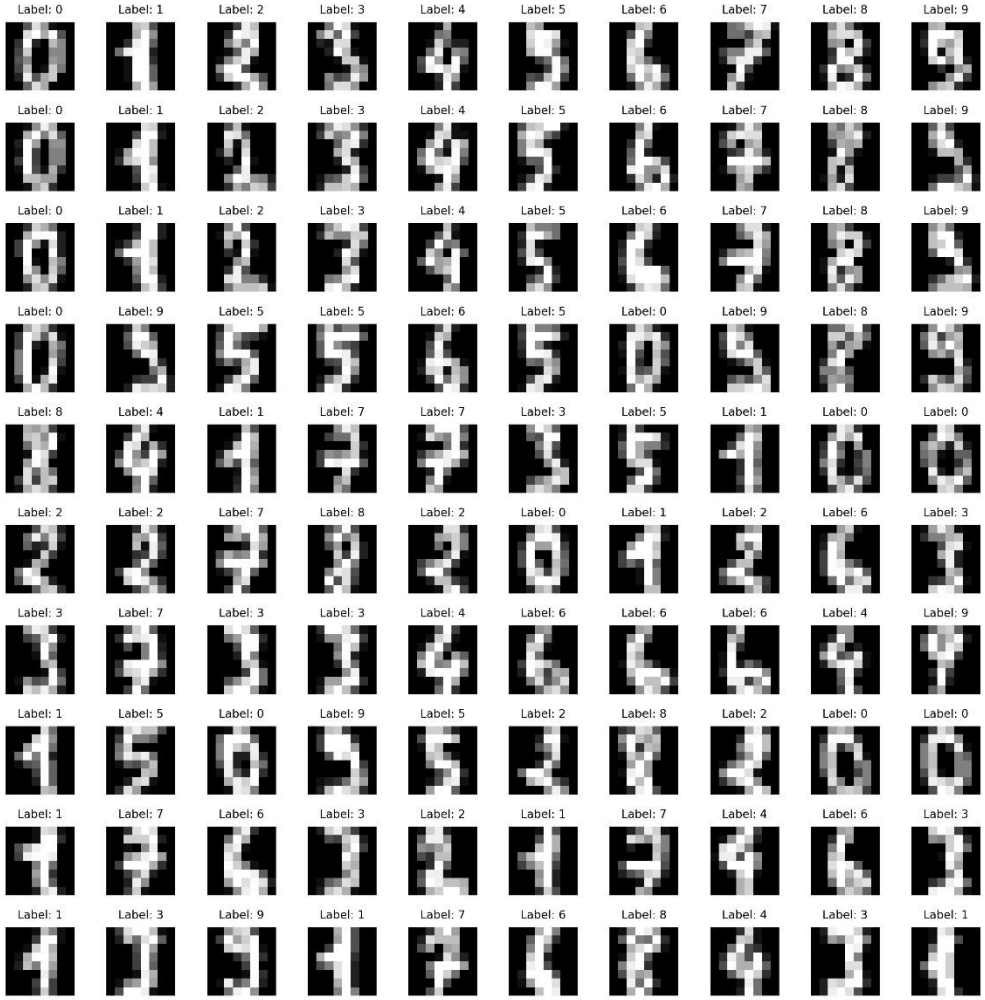
数据集说明

本次任务采用 sklearn.datasets 中的 Digits 数据集。Digits 数据集是机器学习领域中常用的一个多类分类数据集，它来源于 UCI 机器学习库。这个数据集包含了 1,797 个 8x8 像素的手写数字图像，涵盖了从 0 到 9 的十个数字类别。每个图像都是灰度的，并且已经被中心化，即图像的中心位于网格的中心位置。下图展示了数据集中部分图片：



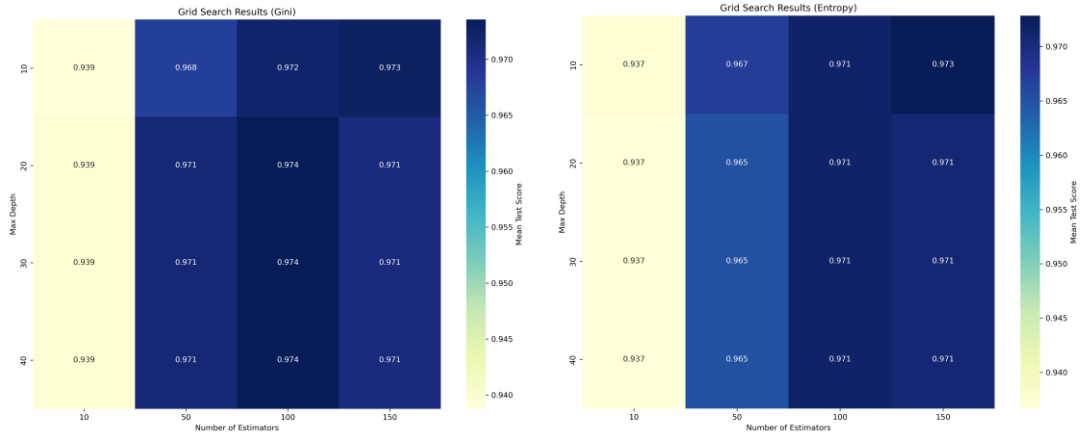
实验过程

按照 8：2 的比例对数据集进行划分，其中 1437 个样本作为训练集，剩下的 360 个样本作为测试集。分别使用 Gini 指数和信息增益来构造随机森林，并且使用网格搜素的方式找到模型的最佳参数。其中主要有两个超参数：“

1. n_estimators：对原始数据集进行有放回抽样生成的子数据集个数，即决策树的个数。若 n_estimators 太小容易欠拟合，太大不能显著的提升模型。
2. max_depth：决策树最大深度。若等于 None,表示决策树在构建最优模型的时候不会限制子树的深度。通常情况下，如果模型样本量多，特征也多的情况下，推荐限制最大深度；若样本量少或者特征少，则不限制最大深度
3. criterion：表示节点的划分标准。这里我们分别采用"gini"基尼指数和"entropy"信息增益作为节点的划分标准。

实验结果分析

通过网格搜索的方式得到最佳参数：




```
最佳参数： {'criterion': 'gini', 'max_depth': None, 'n_estimators': 100}
最佳性能： 0.9735554587688734
```

下面是训练好的最佳模型在测试集上预测结果可视化展示: