一、 实验名称：决策树

决策树与 K-邻近算法有着相同的作用，可以对数据进行分类预测，不同的是 K-邻近算法基本不存在对模型进行数据训练的过程，而决策树算法会预先对数据进行训练得到一颗能帮助做决策的“树”。决策树算法通过对训练数据集进行数据集划分，让原本无序的数据集变为多个有序的子集。

二、 数据描述：

Lenses 数据集包含很多患者眼部状况的观察条件以及医生推荐的隐形眼镜类型。Lenses.txt 数据集一共 24 条数据，每条数据为一个患者的观察情况（age，prescript，astigmatic，tearRate）。隐形眼镜类型包括硬材质、软材质以及不适合佩戴隐形眼镜（hard，soft，no lenses）。

三、 实验内容：

1. 分析数据
2. 训练决策树
3. 绘制决策树

四、 实验环境和编程工具：

系统环境：Windows all/Linux

程序语言：Python

编译工具：PyCharm

第三方包： pickle，matplotlib，operator

五、 实验步骤：

1 分析数据集

根据数据描述，lenses 数据集一共有 4 个特征：age，prescript，

astigmatic，tearRate 和一个类别标签信息。根据 lenses.txt 了解数据在文件中的存储格式。

2 计算给定数据的香农熵

数据集划分前后的信息变化称之为信息增益，衡量信息的方式是香农熵。

$$\mathcal{H} = -\sum_{i=1}^{n} p(x_i)\log_2 p(x_i)$$

$\mathcal{H}$ 为给定数据集的香农熵，简称熵。x_i为（特征），n 为特征个数，x_i的信息定义为 $\ell(x_i) = -\log_2 p(x_i)$，熵定义为信息的期望值。熵值越高，信息越分散无序，反之则越有序。

还可以使基尼不纯度（Gini impurity）度量集合无序程度。

3 根据香农熵选取最优的数据集划分方式

划分数据集的原则是最优化信息增益，让无序（无规律）的数据在划分后可以变成多个有序的子集。故通过计算数据集划分前后的熵值变化决定数据集的划分方式。

4 递归生成决策树

使用 python 的字典变量<key, value>，key 表示决策树节点，保存特征名称；value 表示节点分支情况。

限定数据划分的边界：整个数据集高度有序则表示不能再划分；所有的特征都已经被划分。

递归构建决策树：对已经划分的数据子集分别进行划分。

5 绘制决策树

对模型训练结果进行可视化，绘制决策树树型结构