机器学习

第2次实验：
使用决策树和随机森林对数据分类

实验目的

01

会用Python提供的sklearn库中的决策树算法对数据进行分类

02

会用Python提供的sklearn库中的随机森林算法对数据进行分类

03

会用Python提供的方法对数据进行预处理

实验内容

使用决策树算法和随机森林算法对income_classification.csv的收入水平进行分类。训练集和测试集的比例是7:3，选取适当的特征列，使得针对测试样本的分类准确率在80%以上，比较2种分类方法的准确率。

特征列

age：年龄，整数
workclass：工作性质，字符串
education：教育程度，字符串
education_num：受教育年限，整数
maritial_status：婚姻状况，字符串
occupation：职业，字符串
relationship：亲戚关系，字符串
race：种族，字符串
sex：性别，字符串
capital_gain：资本收益，浮点数
capital_loss：资本损失，浮点数
hours_per_week：每周工作小时数，浮点数
native_country：原籍，字符串

分类标签列：income

imcome > 50K
Imcome ≤ 50K

实验内容

题目1：读入数据并显示数据的维度和前5行数据

题目2：对连续变量年龄进行离散化，并显示前5行数据离散化后的结果

题目3：对属性是字符串的任意特征进行数字编号处理，显示前5行编码后的结果，

每个特定的字符串用一个整数来表示，整数序列从0开始增长。

实验内容

题目4：对预处理后的数据用决策树算法和随机森林算法分类

实验步骤

1) 选择合适的若干特征字段

2) 按7:3划分训练集和样本集

3) 使用训练集训练一个决策树分类器

4) 使用测试集计算决策树分类器的分类准确率

5) 使用训练集训练一个随机森林分类器

6) 使用测试集计算随机森林分类器的分类准确率