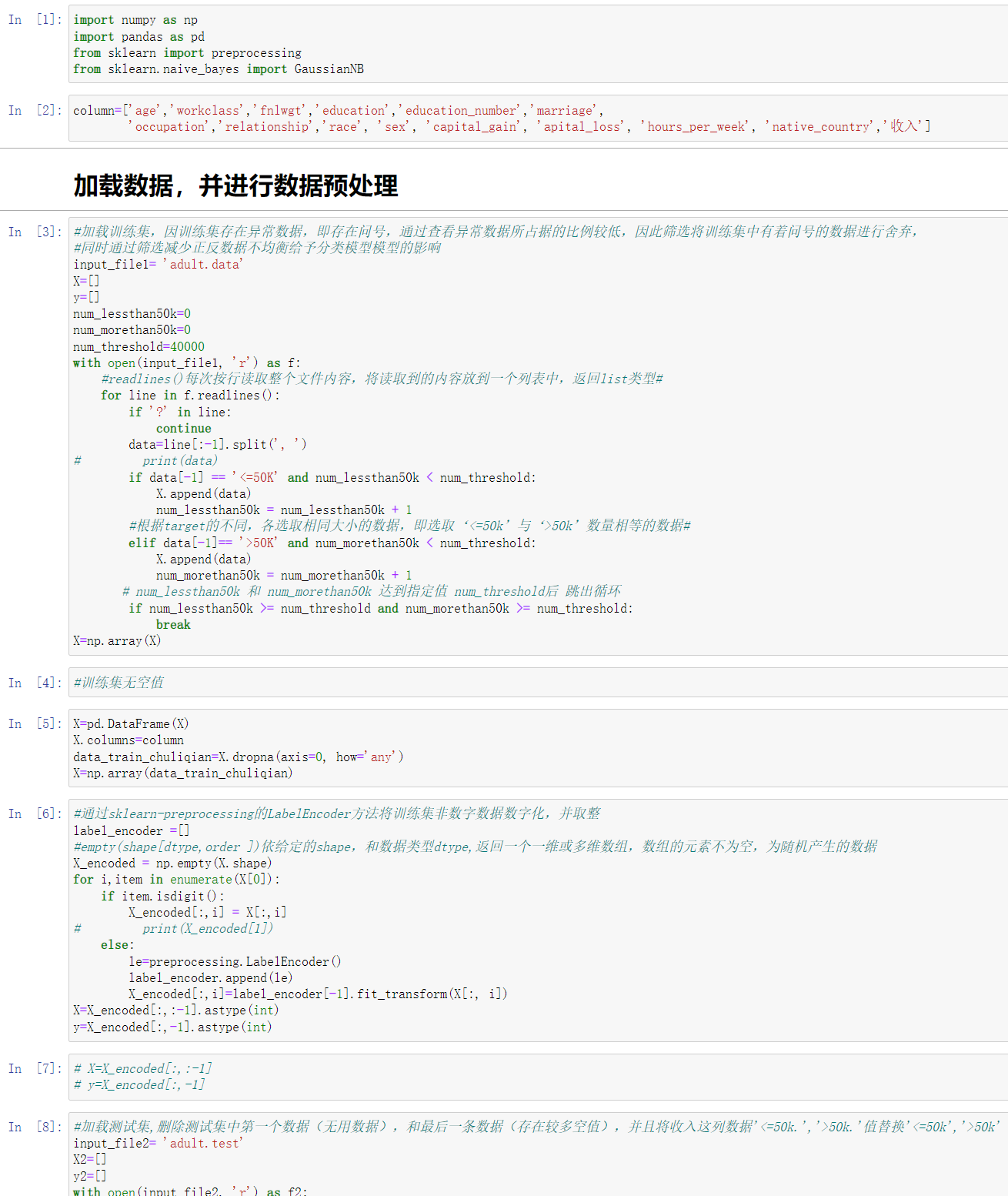
设计要求：

1. 实验说明中1的预处理，写明处理过程或方法，及部分原始数据处理后的结果（数据集的5%）

利用python，导入训练集和测试集，在adult.data存在一定数量的异常数据，例如个别数据中存在“？”内容，因为此异常数据比例较第，因此通过字符串筛选的方法将存在“？”的数据进行去除，同时进行正反数据数据量的筛选，减少后面因数据不均衡给与模型的影响。通过dropna方法查看到数据集没有出现空值的数据，之后通过LabelEncoder方法将训练集字符串特征进行硬编码并取整。其训练集具体实现过程如下

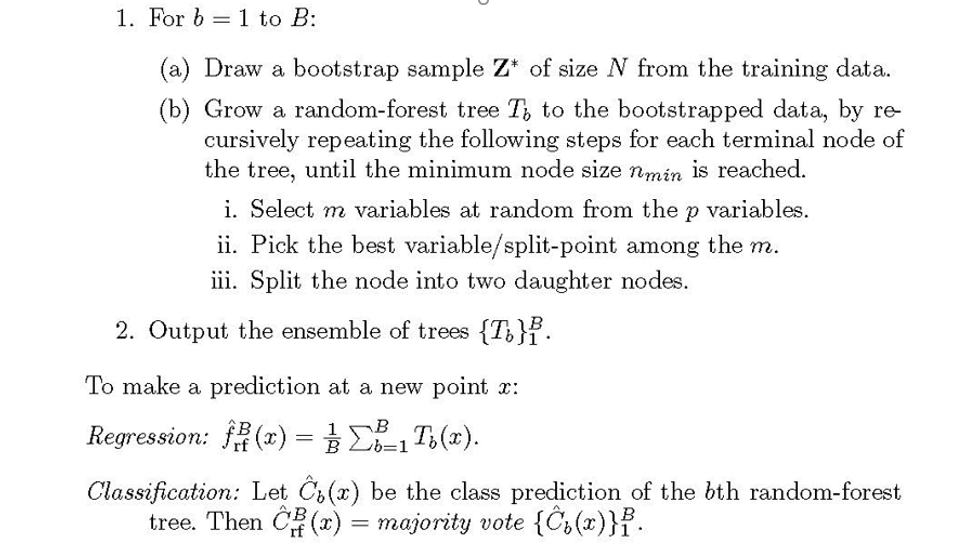


对于测试集，经过查询第一条数据属于异常数据，最后一条数据出现多个空值也属异常数据，并且测试集中的标签是“<=50k.”和“50k.”和训练集中出现的标签“<=50k”和“50k”不一致，无法进行同样的编码处理，因此在将异常数据删除并进行标签修改后将测试集数据进行同样的编码处理。其实现过程如下

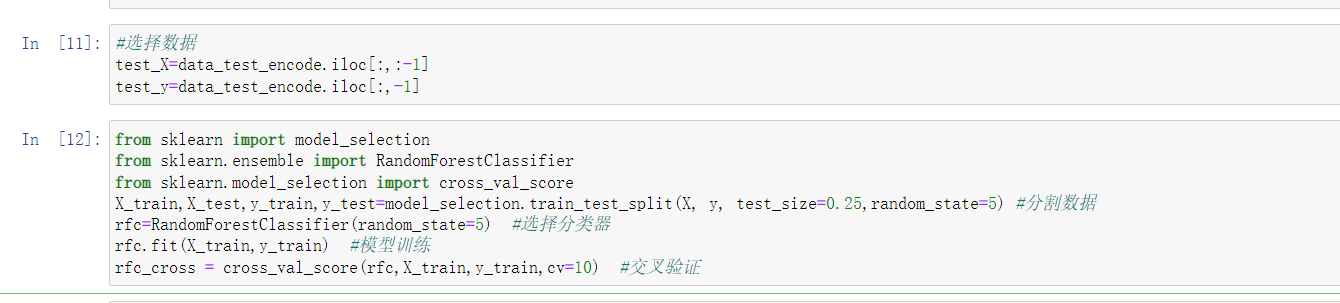
、其训练集和测试集进行预处理过后的结果（百分之五）保存在附件中

1. 分类模型建立的算法伪代码和源代码（并标注主要过程或方法）

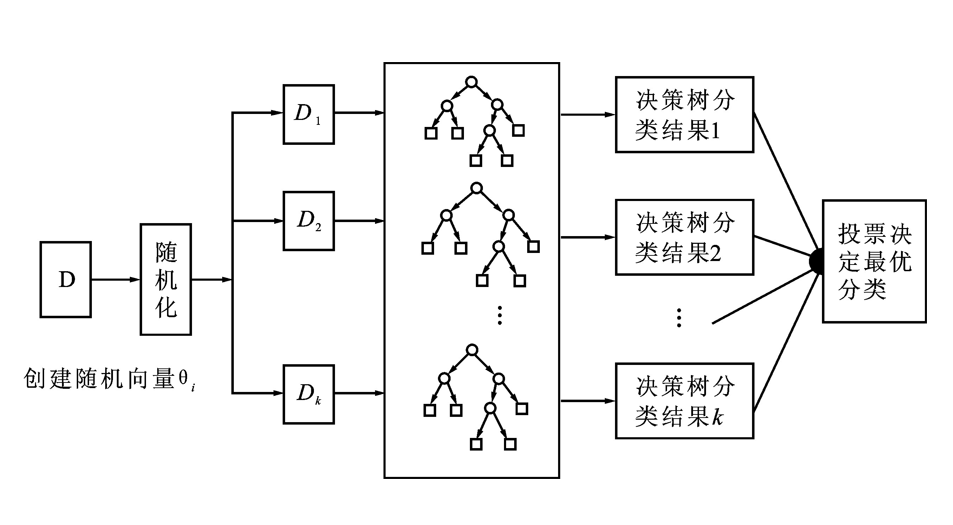
本次实验分类使用的是随机森林其伪代码如下：



源代码如下：



1. 展示20个数据对象的算法执行过程#

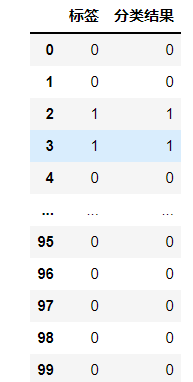


1. 写出形成的分类规则的源代码（标注主要过程或方法），以及该分类模型形成的分类规则

随机森林分类(RFC)是由很多决策树分类模型{h (X,) , k=1, . . .}组成的组合分类模型,且参数集{}是独立同分布的随机向量,在给定自变量X下，每个决策树分类模型都由一票投票权来选择最优的分类结果。RFC的基本思想：首先，利用bootstrap抽样从原始训练集抽取k个样本,且每个样本的样本容量都与原始训练集一样，其次，对K个样本分别建立k个决策树模型，得到k种分类结果;最后，根据k种分类结果对每个记录进行投票表决决定其最终分类。

1. 写出实验说明4）中对前100项测试数据进行测试的结果

其结果样例如下，其中“0“为<=50k,“1”为>50k。



1. 基于5）的结果，写出构建的混淆矩阵，并进行实验说明4）的评估，写出指标的求解过程

基与测试集前100条结果构建的混淆矩阵为，其中76为真负类（TN），4为假负类（FN），3为真正类（TP），17为假正类（FP），

准确率 （Accuracy）= (TP+TN)/(TP+TN+FP+FN)

精确率 （Precision） =(TP)/(TP+FP)

召回率 （Recall） =(TP)/(TP+FN)

F1=2\*(Precision\*Recall)/(Precision+Recall)

所以其准确率为0.93，精确率0.85，召回率为0.809523，f1值为0.8293

1. 写出实验说明4）所有测试数据集的指标值

准确率为: 85.0730411686587 %

精确度为： 73.26923076923076 %

召回率为: 61.78378378378378 %

f1值为： 0.6703812316715543 %