依据人口普查数据预测个人的收入状况（分类问题）：

一、目标：

根据美国1994年的人口普查数据，建立分类模型，能正确判别某个人的收入是否超过5万美元/年（二分类问题）

二、数据集:

census\_income.csv

【属性】

数据集包含14个属性，其中，年龄、final weight、最高学历、资本收入、资本指出和每周小时数是数值属性，其余是标称属性。

目标属性为： <=50K, >50K



1. 要求（部分要求如果难以实现，允许不完成，只要把数据挖掘过程描述清楚即可）：
2. 应用合适的数据探索方法，更好地了解数据。
3. 数据预处理：
   1. 检查数据集是否存在缺失值和异常值，如果存在，则消除它们（任意方法均可）。
   2. 根据选用的模型，自行决定是否需要归一化数据，另外，将目标类别标设置为0类（<=50K）和1类（>50K）
   3. 是否需要特征工程，如何实施？
   4. 标称属性需要进行数值处理（参考我们此前对german\_credict数据集的处理方法）
   5. 其他预处理步骤。
4. 选用合适的分类模型，划分训练数据和测试数据，构建分类模型。
   1. 选用一种合适的分类模型，解释原因。并且，至少和一种对比模型(baseline model)进行实验结果对比。
   2. 解释模型的参数选择方法和最优参数结果。
   3. 选择合适的性能评价指标评价模型的性能。
   4. 绘制有关训练结果或训练过程的图。
5. 完成“xxx数据挖掘分析报告”。
6. 准备5分钟左右的ppt报告，在第17周随堂讲述本设计的主要工作。