实验四

中国科大2025年春季学期"数据分析及实践"课程 - 实验四说明文档。

任务概述

本实验基于实验三所使用的PISA数据集，挖掘部分问卷属性的频繁项集与关联规则，请按要求完成实验任务。

任务列表

1.（15%）读取数据集`data.csv`，进行数据预处理。

- Q1.（5%）选取问卷中的`SC155Q01HA,SC155Q02HA,SC155Q03HA,SC155Q04HA,SC155Q05HA`5个离散性特征作为特征集，分别介绍这些特征所代表的含义和各自取值范围，注意到这些特征名称本身较为冗长且不易于理解，请对特征名进行简化修改，并删除存在缺失值的行。
- Q2.（10%）注意到选取的特征可能存在相同取值（如特征A和B都可能取值0），不便于后续的关联分析过程。请构建项集索引，并依据索引内容进行特征值替换。项集索引字典形式如下：

```
ind2val = { 0: '[COLUMN1]=[VALUE1]', 1: '[COLUMN1]=[VALUE2]', ... , }
```

基于所选项集索引字典进行单元格内容替换，以便于后续频繁项集挖掘和关联分析过程。

2.（60%）基于预处理后的数据集，编写算法代码进行频繁项集挖掘。

- Q1.（30%）请参考以下 Apriori 产生频繁项集的算法流程，自行编写相应代码，分别以最小支持度阈值为0.25和0.5，挖掘频繁项集。

算法 6.1 *Apriori* 算法的频繁项集产生

```
1:  k = 1
2:  F_k = {i | i∈I ∧ σ({i}) ≥ N × minsup}      {发现所有的频繁 1-项集}
3:  repeat
4:    k = k+1
5:    C_k = apriori-gen(F_{k-1})      {产生候选项集}
6:    for 每个事务 t∈T do
7:      C_t = subset(C_k, t)      {识别属于 t 的所有候选}
8:      for 每个候选项集 c∈C_t do
9:        σ(c) = σ(c) + 1      {支持度计数增值}
10:     end for
11:   end for
12:   F_k = {c | c∈C_k ∧ σ(c) ≥ N × minsup}      {提取频繁 k-项集}
13: until F_k = ∅
14: Result = ∪F_k
```

- Q2.（15%）当最小支持度为0.5时，频繁项集数量较少。请将各特征原始取值为1和2的单元格统一修改其值为0，取值为3和4的单元格统一修改其值为1。重复T1-Q2的项集索引构建过程，并以最小支持度阈值为0.5，挖掘频繁项集。
- Q3.（15%）分析Q1和Q2的结果，你有什么发现？请根据各特征定义，分析产生这种情况的原因。

3.（25%）基于T2-Q2得到的频繁项集挖掘结果，编写算法代码进行关联规则提取。

- Q1.（15%）请以最小置信度阈值为0.8，提取形如`X->{1}`的关联规则，并输出它们的置信度和提升度。
- Q2.（10%）参考项集索引的对应关系，对以上频繁项集和关联规则结果进行简要分析和总结。

格式和提交要求

1. 请按具体任务分步编写代码，存储于`.ipynb`格式文件中用于复现，必要时可增加注释。
2. 本实验可使用的外部库为`pandas`和`numpy`。
3. 实验报告必须涵盖任务列表中的所有内容和相应结果，并请存储于`.pdf`格式文件中。
4. 提交时，请将实验源代码和实验报告保存至一个压缩包中，命名为“学号-姓名-实验四.zip”，并于2025年5月8日之前发送至USTC_AD2025@163.com。