# 实验四

中国科大2024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

乳腺癌数据集（Breast Cancer Dataset）构建于1988年，来源于南斯拉夫卢布尔雅那肿瘤研究所大学医学中心。该数据集记录了286个乳腺癌患者的疾病复发情况和部分个体属性值（包含患者年龄、肿瘤大小、是否放疗等9种类别型特征）。现欲挖掘该数据集各属性特征之间的频繁项集与关联规则，为乳腺癌的疾病预后提供有用的信息模式，请你按要求编写 Python 代码实现任务列表中的内容。

## 任务列表

1.（25%）读取数据集`data2.csv`，存储到变量`df`中，进行数据预处理。

- Q1.（5%）原始数据表存在部分缺失值，请指出哪些特征含有缺失值，并删除所有含空缺值的行。
- Q2.（10%）当前数据表未能正确处理部分数据值的文本与日期表示类型，使得`tumor-size`与`inv-nodes`含有大量异常值，请使用`value_counts()`方法验证，并参照`variables.xlsx`修正所有异常值。
- Q3.（10%）数据表中的特征多为文本属性，不便于后续的关联分析处理过程，请导入`variables.xlsx`，用数字索引替换之，并展示索引与属性值的对应关系字典`ind2val`。

  例如，`Class`属性含`no-recurrence-events`与`recurrence-events`两种可能值，可分别用`0,1`代替，`age`含`10-19`，`20-29`等可能值，可分别用`2,3,...`替代之，以此类推。 相应地，可建立字典类型变量：`ind2val = {0: 'Class=no-recurrence-events', 1: 'Class=recurrence-events', 2: 'age=10-19', 3: 'age=20-29', ... }`。

2.（75%）基于预处理后的数据集`df`，编写算法代码进行关联规则分析。

- Q1.（45%）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，以最小支持度阈值为0.4，挖掘`df`中的频繁项集。

**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.（20%）基于提取出的频繁项集，以最小置信度阈值为0.75，提取形如`X->{0}`的强关联规则，并分别输出它们的置信度和提升度。

- Q3.（10%）参考`ind2val`中索引与属性值的对应关系，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。