# 关联规则挖掘

2120150993 胡健

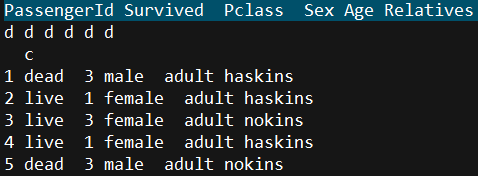
## 一、数据描述

本次实验选择的是Titanic存活数据来进行实验。训练数据包含12个属性，分别是PassengerId（乘客ID），Survived（是否生还），Pclass（船舱等级），Name（姓名），Sex（性别），Age（年龄），SibSp（船上的兄弟姐妹及配偶个数），Parch（船上的父母及子女个数），Ticket（船票号码），Fare（票价），Cabin（船舱号），Embarked（登船港口）。数据集中共有891条乘客记录。

## 二、数据预处理

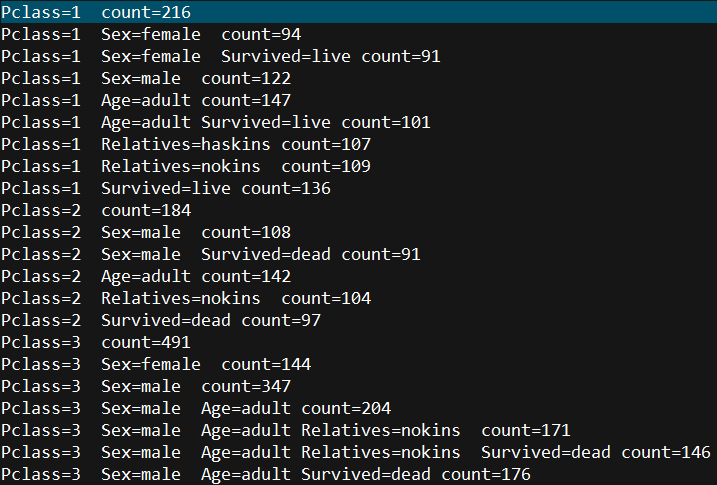
在全部的12个属性中，Name，Ticket，Fare，Cabin，Embarked属性属于乘客的个人信息，不做处理。在Survived属性里用dead代替0，用live代替1。在Age属性里，用“teen”代替小于18的元素，用“adult”代替18到55之间的元素，用“old”代替大于55的元素。SibSp和Parch两个属性都表示乘客的亲属个数，因此对二者求和得到乘客的总亲属个数，用Relatives属性表示。

预处理后的数据包含6个字段，PassengerId（乘客ID），Survived（是否生还），Pclass（船舱等级），Sex（性别），Age（年龄），Relatives（亲属个数）。



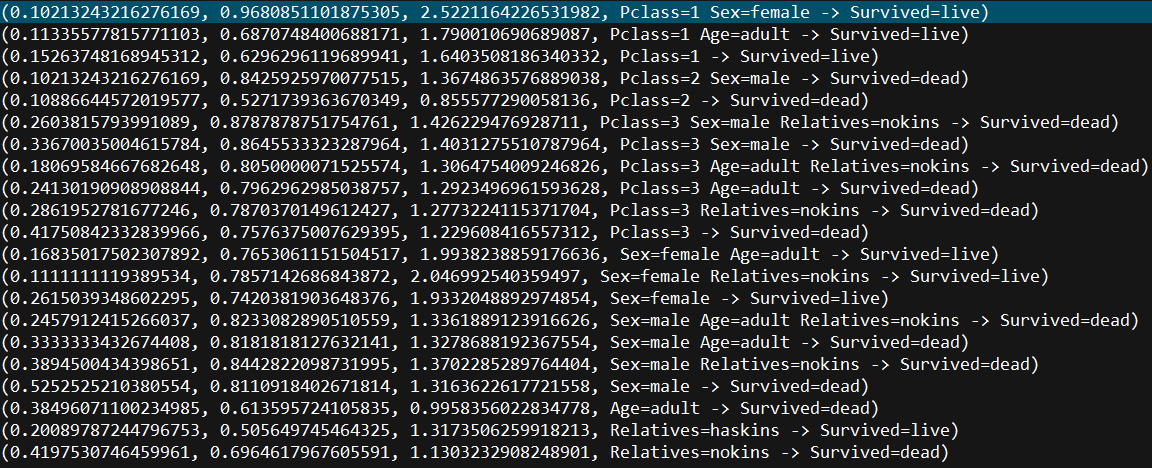
## 三、频繁项集

实验中测试了多组支持度和置信度，最终选择最小支持度为0.1，得到的频繁项集为freqItems.csv。



## 四、关联规则

得到的关联规则为AssociateRulesR.csv。



## 五、规则评价

使用Lift对生成的关联规则进行评价：

最高的Lift值对应的规则为“Pclass=1 Sex=female -> Survived=live”，说明这个规则比较合理。

## 六、可视化与分析

在此次实验中发现了比较有趣的现象，就是单身中年男性且坐的是三等舱，遇难的可能性更高。从Lift数值中也可以看出，Pclass=3 Sex=male Relatives=nokins -> Survived=dead这条规则的Lift值较高，为1.426229477。