**杭州电子科技大学计算机学院**

**数据仓库与数据挖掘**

**实验2：分类和预测**

时间：2021年12月8日，学号：19151633姓名：应宇杰

# 一、实验目的

1、掌握数据挖掘中数据预处理的方法；

2、了解数据转换的过程和方法；

3、理解分类的一般过程和基本原理；

4、巩固分类算法的算法思想，能够进行分类操作；

3、学会分类预测问题中的性能评估方法。

# 二、实验原理

**1、数据预处理**

现实世界中的数据库极易受噪音数据、遗漏数据和不一致性数据的侵扰，为提高数据质量进而提高挖掘结果的质量，产生了大量数据预处理技术。数据预处理有许多方法：

(1) 数据清理: 数据清理是完成格式的标准化、对空缺值进行处理、清除重复的数据以及对异常数据进行错误纠正和清除等操作；

(2) 数据集成: 数据集成是将来自不同数据源的数据合并为统一一致的数据存储中, 这种数据存储可以是数据库或数据仓库;数据集成主要包括:包含相同字段属性的纵向追加和具有相关属性叠加的横向合并。

(3) 数据归约: 数据归约是针对原始数据集中地属性和记录, 实现有效的数据采样与对应属性选择, 进一步降低数据规模, 在数据归约过程可以采用聚集、聚类以及将冗余特征值删除等形式, 达到既能最大限度的保持数据的原有特征, 又能够有效的精简数据量的目的。数据归约主要通过数据立方体技术、维消减、数据压缩、数据块消减、离散化和概念层次生成等方法实现。

(4) 数据变换: 数据变换是根据需要将数据压缩到较小的区间中, 也就是对数据进行规格化处理, 将数据压缩到特定的范围之内。

以上几种数据预处理方法, 相互之间不仅关联而且是独立的, 各个预处理方法的实施并没有先后顺序的严格制约, 并且相互贯通, 例如消除数据冗余的过程既可以看做是数据清洗过程的一项工作, 也可以认为是数据归约工作中的一种方法

**2、常用的预测模型：决策树、朴素贝叶斯分类器等。**

**3、评价预测结果，常用的性能评价指标：**

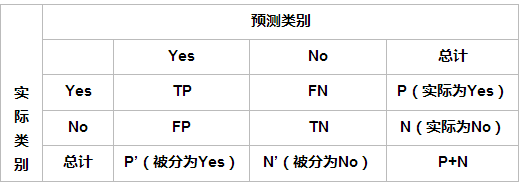
混淆矩阵

True Positive(真正，TP)：将正类预测为正类数

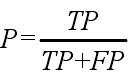
True Negative(真负，TN)：将负类预测为负类数

False Positive(假正，FP)：将负类预测为正类数(误报)

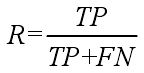
False Negative(假负，FN)：将正类预测为负类数 (漏报)

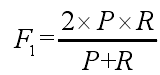


（1）**精确率**P(Precision)是指被正确预测为正例数与所有被预测为正例数的比率，反映了预测模型的准确程度，也称为查准率。



（2）**召回率** R(Recall)是指被正确预测为正例数与实际正例数的比率，反映了一个有缺陷模块被正确预测出的概率，也称为查全率。



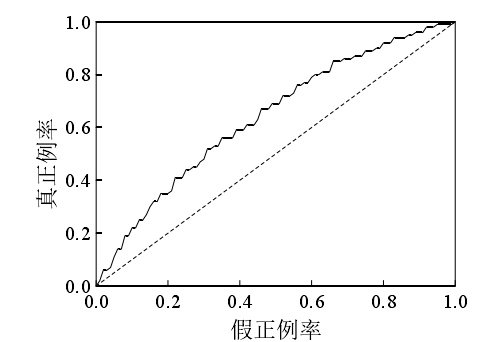
（3）**F-Measure** 是信息检索领域的一个评价指标，常用的是F1 度量，即为精确率与召回率的调和平均数。

（4）**AUC**

真正例率 TPR(True Positive Rate)与召回率相同，也是指被正确预测为正例数与实际正例数的比率。

假正例率 FPR(False Positive Rate)是指被错误预测为正例数与实际负例数的比率。

接受者操作特征(Receiver Operating Characteristic，ROC)曲线是描述分类模型真正例率 TPR 和假正例率 FPR 之间关系的一种图形化方法，如下图所示。



横坐标表示假正例率，纵坐标表示真正例率。对于一个特定的预测模型和训练数据集，其预测结果对应于 ROC 曲线上的一个点，通过调整该模型的阈值即可得到一条经过(0, 0)和(1, 1)的曲线，曲线下方的面积即为AUC(Area Under the Curve)的值。其中，AUC的取值范围为0~1，当 AUC 为 0.5 时，表示随机猜测模型的性能，如上图中的虚线所示。AUC 值越大，说明该模型的性能越好。因此，好的预测模型应尽可能地靠近坐标系的左上角。

# 三、实验内容：

**背景与挖掘目标**

泰坦尼克号沉船事故被认为是20世纪世界十大灾难之一。1912年4月15日凌晨一块像岩石般坚硬的冰块刺进了泰坦尼克号船体，号称“不沉之船”的泰坦尼克号邮轮最终沉入了北大西洋海底。泰坦尼克号沉船事故导致1500多人遇难。根据泰坦尼克号的伤亡记载，女性的生存率高于男性，小孩的生存率高于成人。

通过已有数据，分析乘客的各个属性及其与生存率之间的关系，探索乘客的哪些属性影响生存率，建立新的属性特征并将其加入原始训练集数据，利用处理后的数据建立多个不同的模型，对各个模型进行评估，利用预测精确度最高的模型对测试集中的各个样本进行生存率预测。挖掘目标包括:

(1) 预测船上人员的每个属性和生存率的关系；

(2) 探索与生存率有关的新特征；

(3) 建立预测精确度高的模型；

(4) 预测测试集样本的生存率。

**实验要求：**

运用数据挖掘技术对收集到的数据进行数据探索与预处理，进行数据建模；最后采用分类预测算法，挖掘并预测乘客生存率泰坦尼克号乘客生存率。

(1) 初步分析乘客属性，探索影响生存率的新特征；

(2) 对原始数据进行**预处理**，例如：数据填充、属性归约、数据变换等；

(3) 利用第(2)步形成的建模数据，分别采用**决策树算法**、**朴素贝叶斯算法**预测乘客生存率，计算预测精确度；

(4) 对模型进行分析，最后输出对测试集的预测结果。

# 四、实验步骤

**（1）实验环境**

Python 3.6，使用Pycharm和jupter 作为编辑器

Sklearn 1.1.3

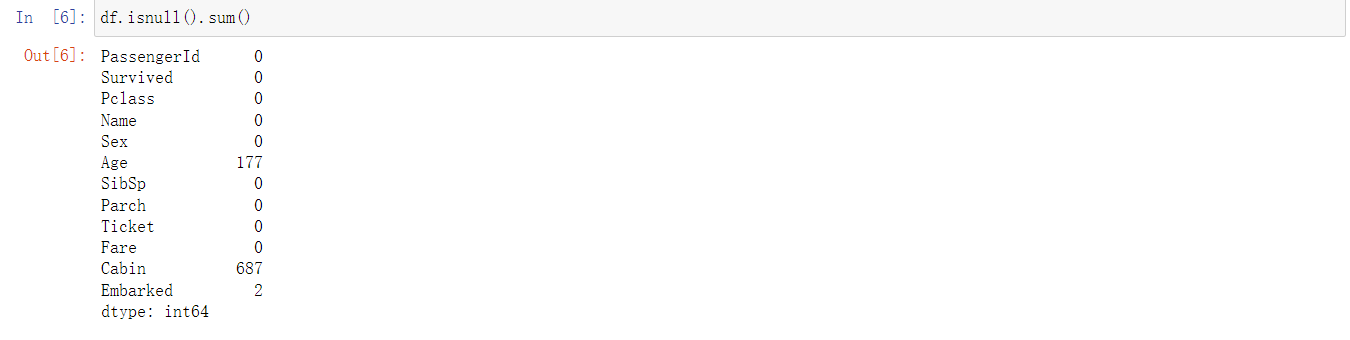
Pandas 2.0.3

**（2）步骤说明（文字+代码）**

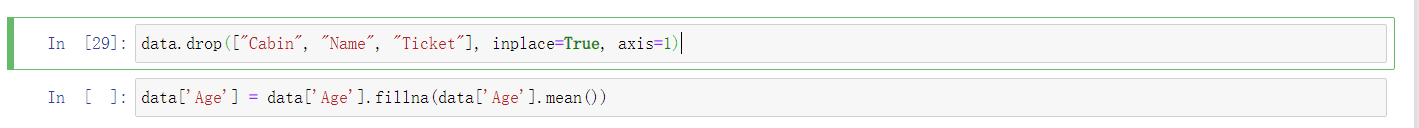
1、导入数据



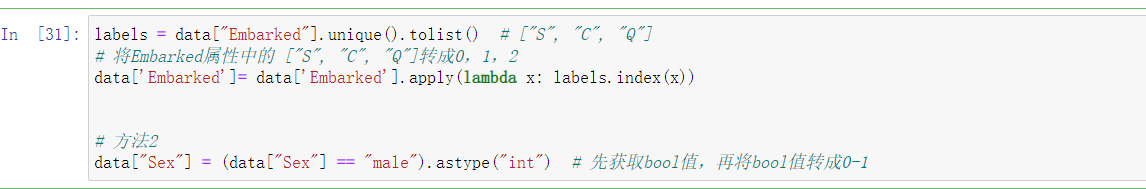
2、查看数据缺失情况，发现船舱编号和年龄大量缺失



3、数据预处理，年龄用平均年龄填充，船舱编号删除，Embarked用最多的港口替代



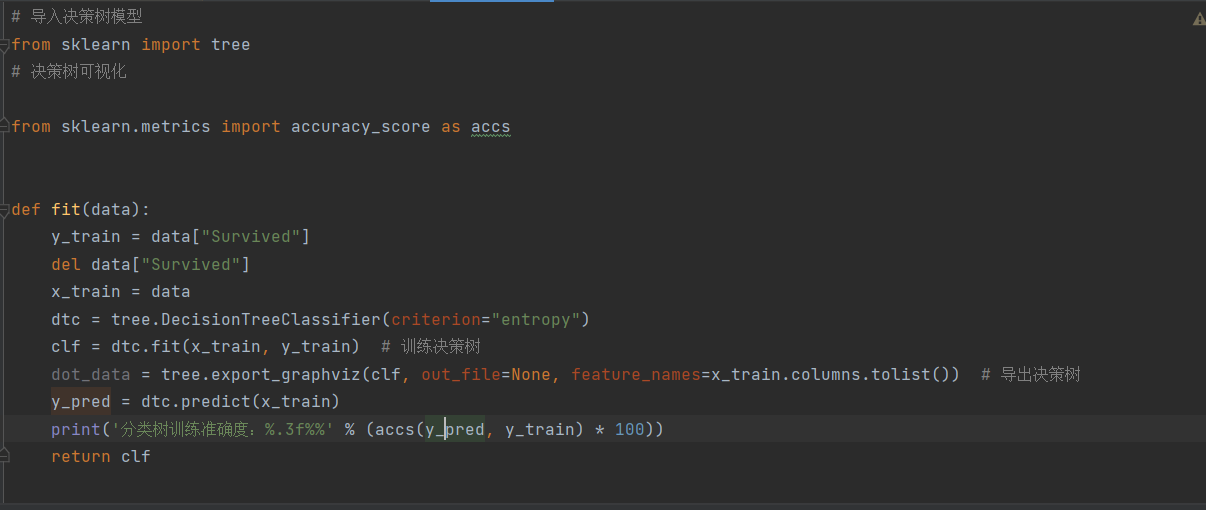
4、将Embark 和sex用数字替代，方便未来计算



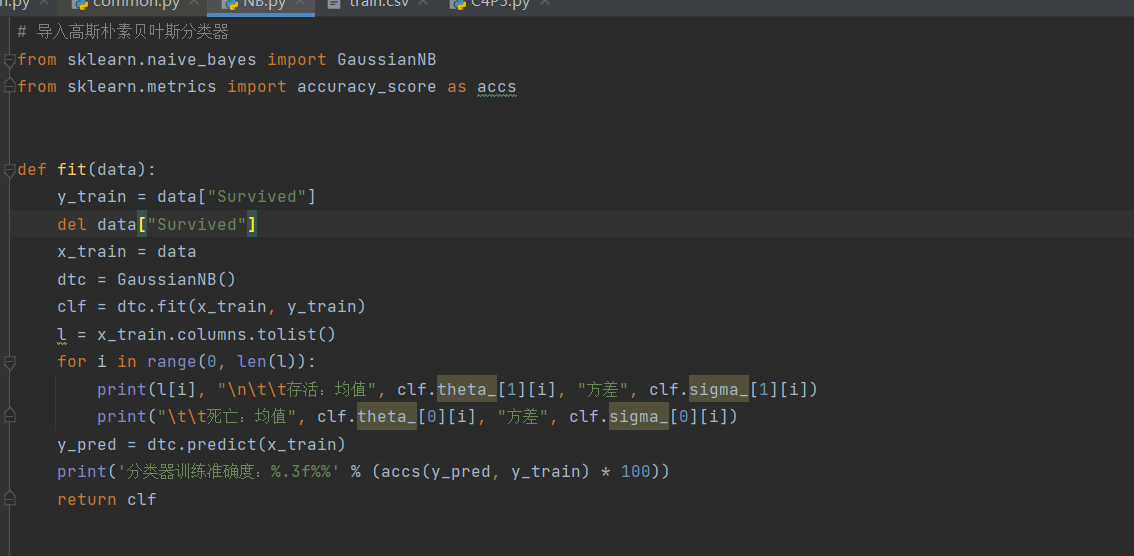
5、删除结果无关属性



6、运用决策树算法进行生存预测



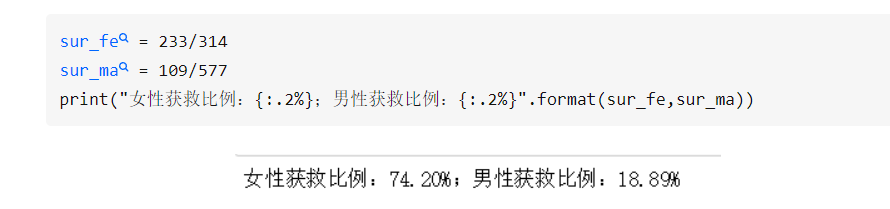
7、利用朴素贝叶斯算法进行生存预测



# 五、实验结果与分析

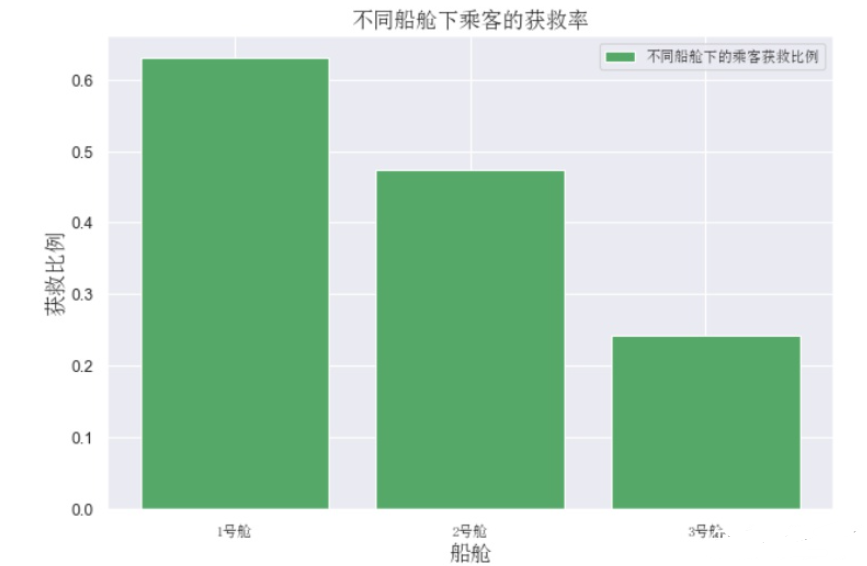
**（1）实验结果**

1、统计分析性别对获救几率的影响



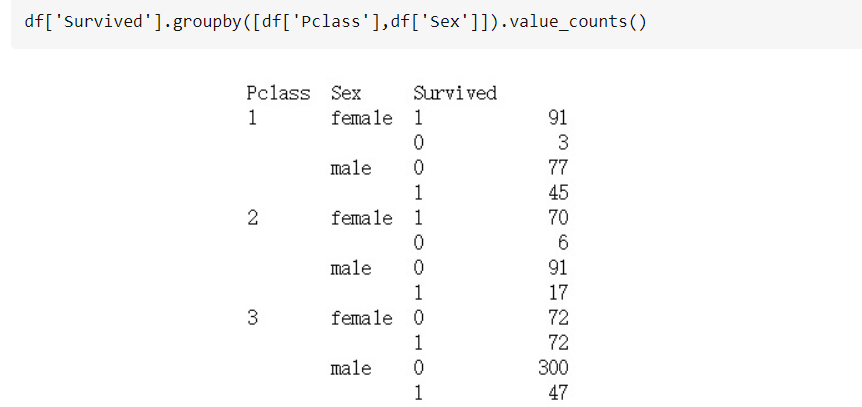
2、统计研究船舱对获救几率的影响



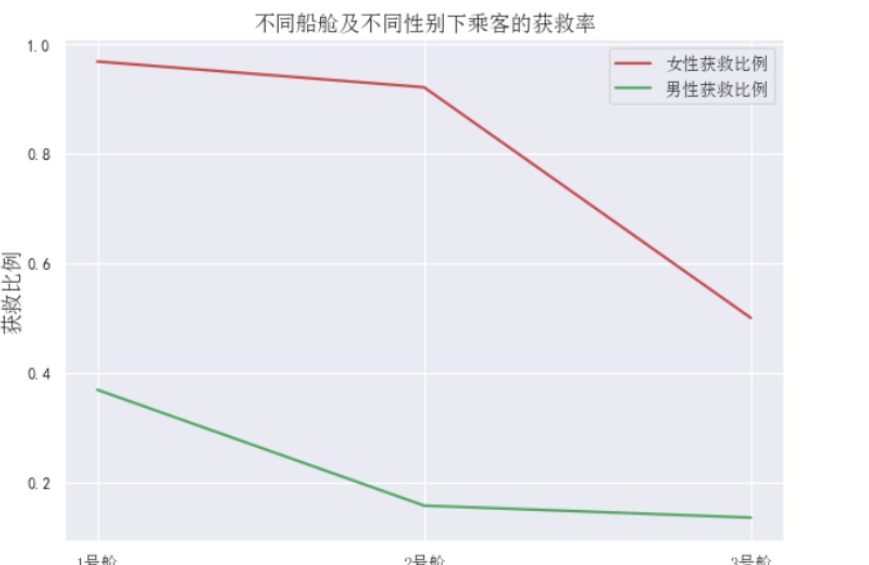


3、研究船舱和性别双因素影响获救几率

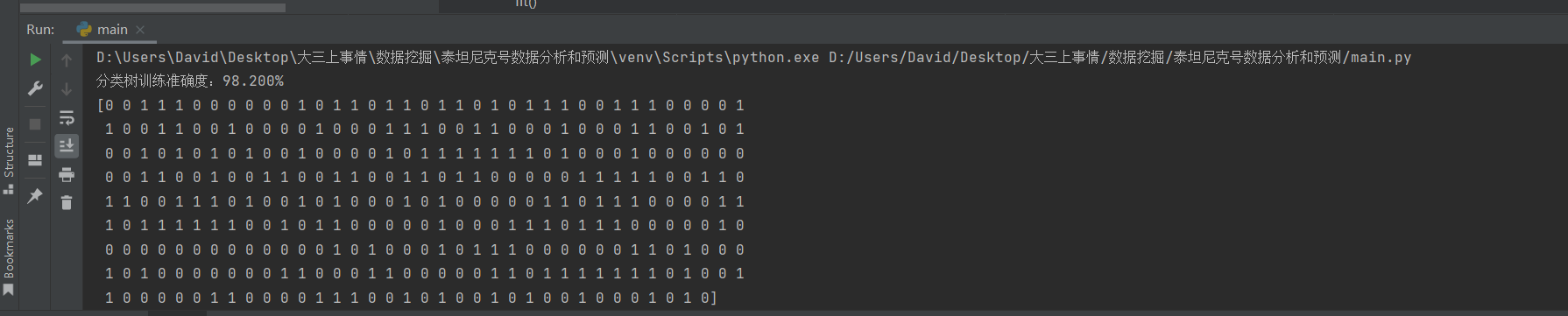
第一步、统计不同船舱的男女人数



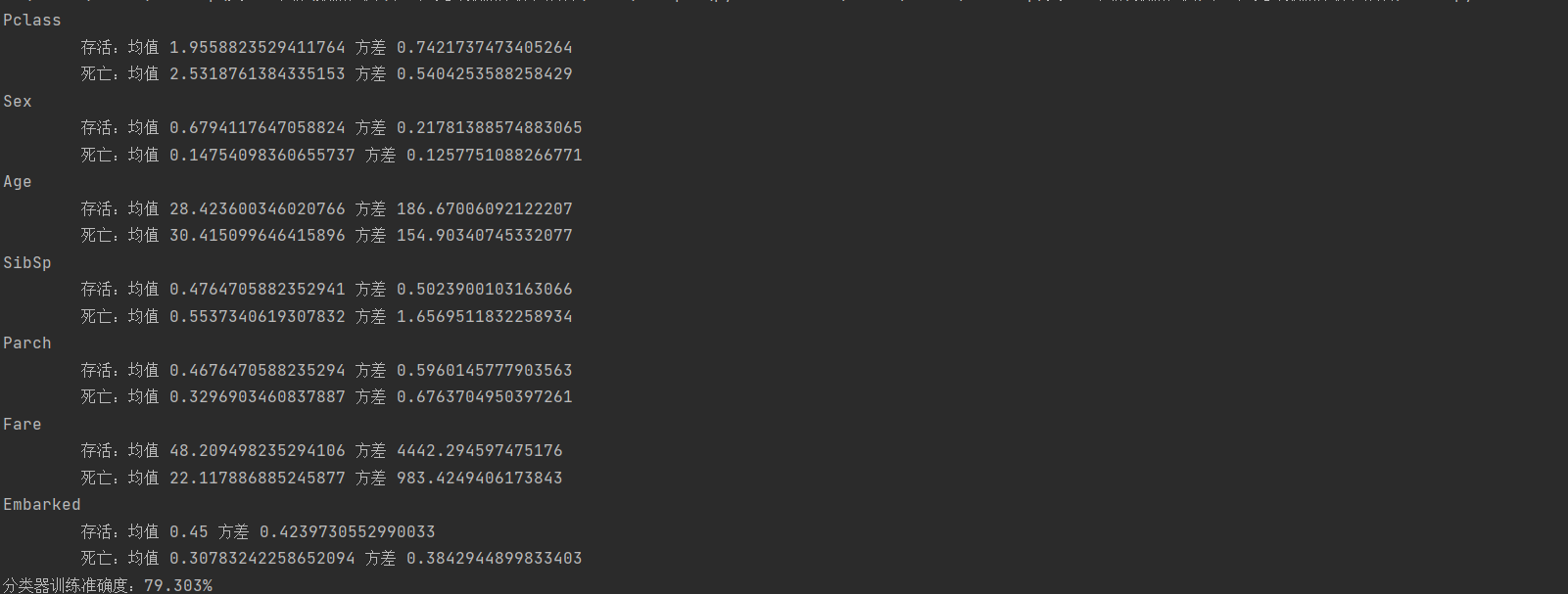
第二步、可视化

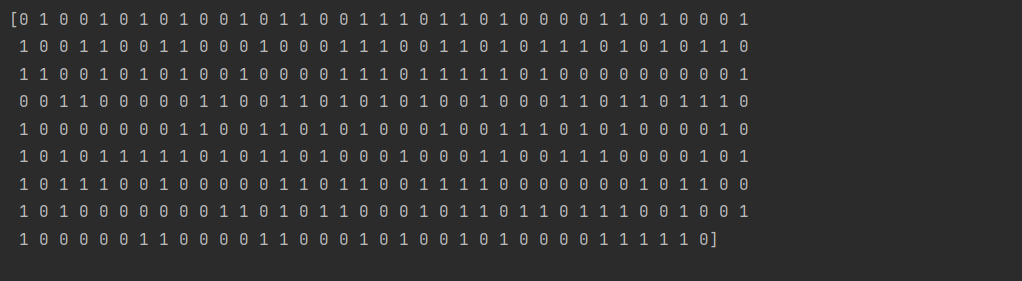


4、决策树模型对训练集的判断正确性，对测试集的预测结果



5、朴素贝叶斯模型进行的结果和对测试集的预测结果





**（2）对实验结果的分析**

发现确实是女性相对于男性更容易获救，而且高舱位的女性比低舱位的女性更容易获救，所以高仓位的女性会有更大的生存几率，用决策树法的通过预测train数据预测成功率为98.2%，而朴素贝叶斯法预测的成功率为79.3%

# 六、心得体会

通过这次实验，我学习了在开始数据挖掘前的数据预处理，包括数据清洗和数据填充和无关参数删除，在数据挖掘时使用决策树法和朴素贝叶斯来预测生存的几率，通过比较两个算法的正确率发现在这个数据集中决策树算法更优。