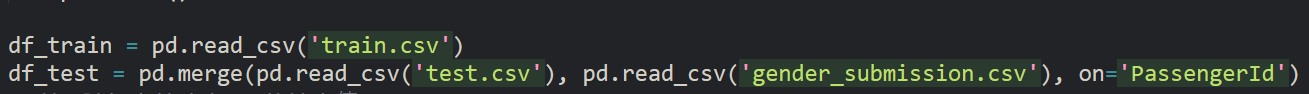
**数据挖掘大作业三：分类与聚类**

王鹏 2120171069

一、对数据集进行处理

本次作业选用数据集为泰坦尼克号人员情况，利用Python中的pandas库进行csv数据文件的读取，对数据集进行预处理，以适合分类和聚类算法。

1. 读取训练数据集和测试数据集：

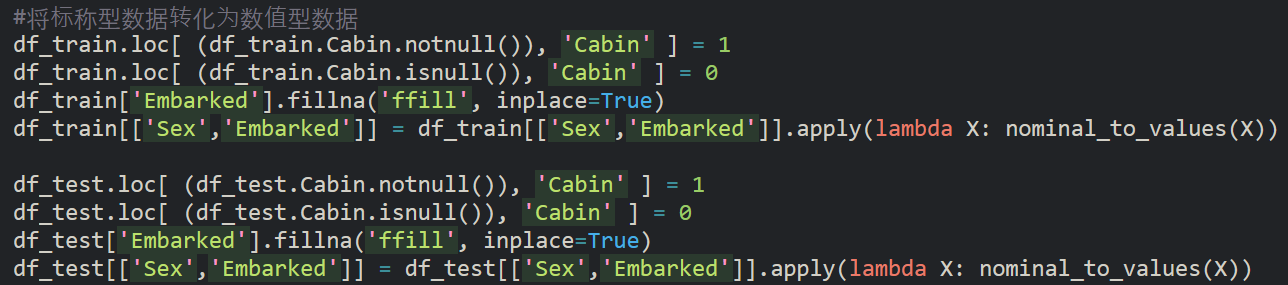


2、处理缺失数据；

用scikit-learn中的随机森林算法填充age缺失值，



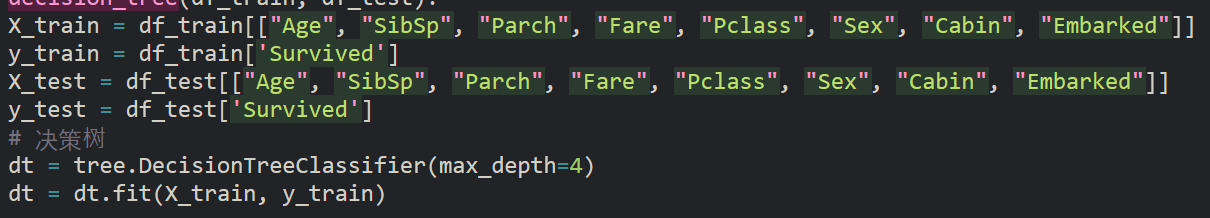
3、按Cabin有无数据，将这个属性处理成Yes和No两种类型，并将“Sex”,”Embarked”属性序列化：



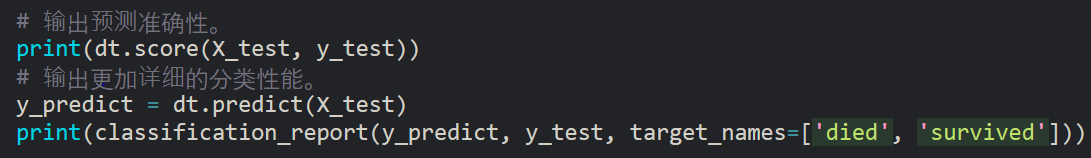
二、分类算法

2.1 决策树

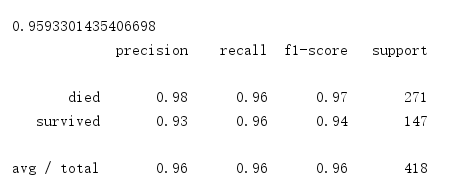
（1）选取数据中用于决策树模型训练的字段，用Sklearn中DecisionTreeClassifier进行决策树模型的训练：



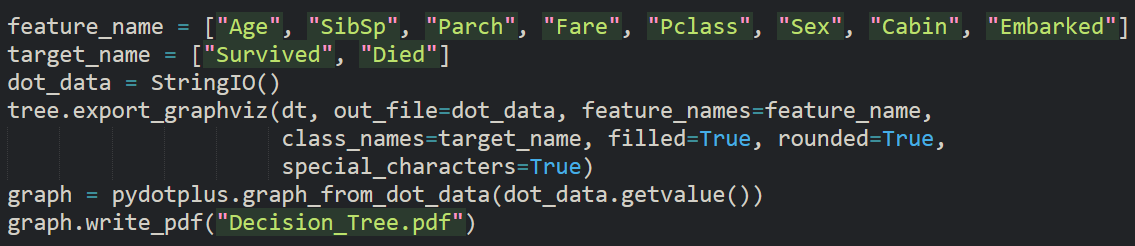
（2）输出预测准确性和详细的分类性能

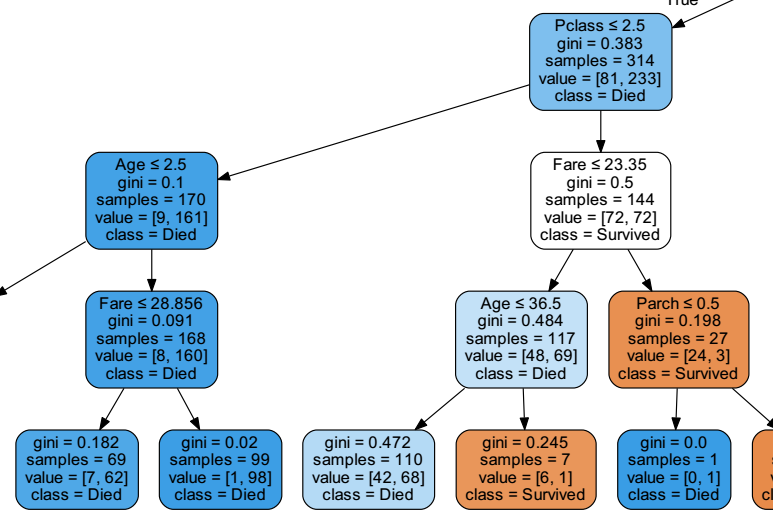


（3）预测准确性和详细的性能结果



（4）决策树可视化代码及效果

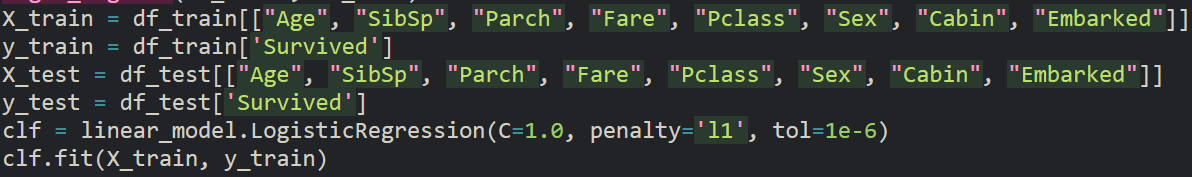




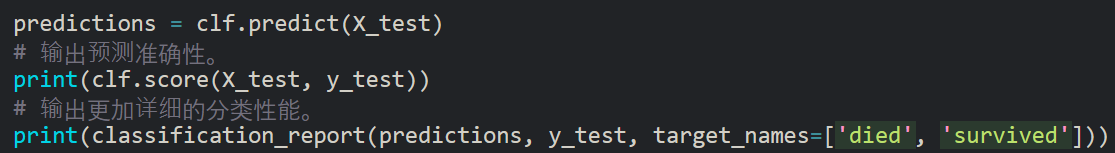
图保存在Decision\_Tree.pdf中。

2.2 逻辑回归

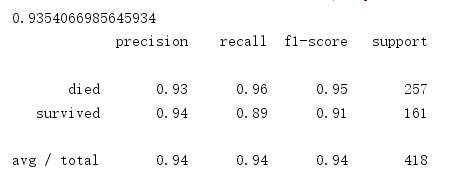
（1）选取数据中用于逻辑回归模型训练的字段，用Sklearn中LogisticRegression进行逻辑回归模型的训练：



（2）输出预测准确性和详细的分类性能：



（3）预测准确性和详细的性能结果如下

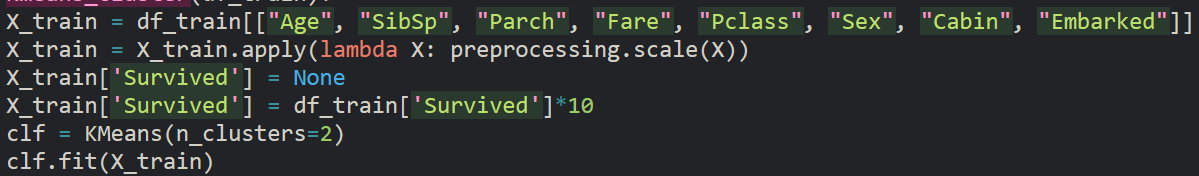


1. 聚类算法

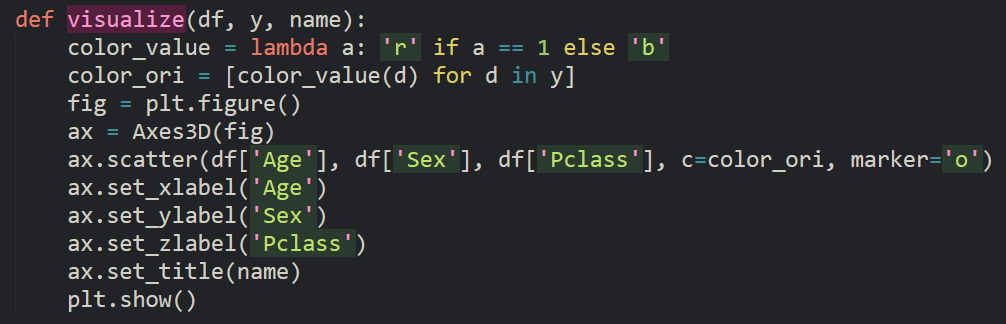
在聚类算法的实验中，我将survived字段作为训练标签加入到数据集中，并将其权重设置为10（其余字段权重为1），期望在聚类过程中聚类结果能表达出其余字段与survived字段的关联

3.1 K-means

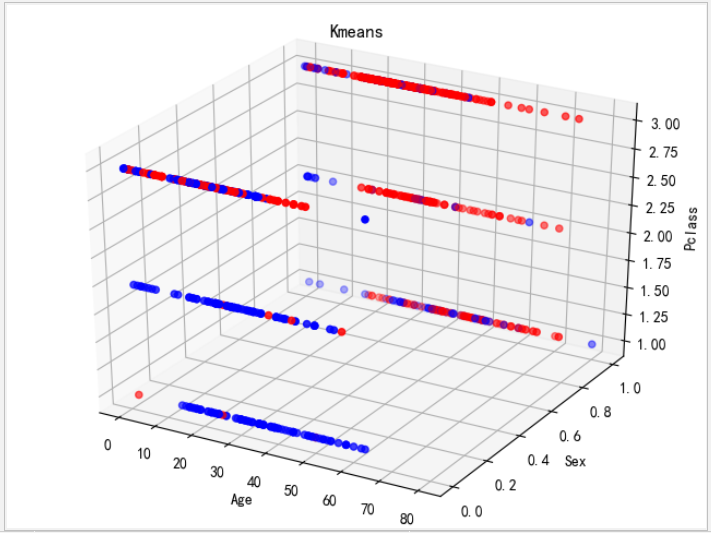
（1）选取数据中用于K-means聚类的数据字段，进行归一化并聚类：



（2）利用Axes3D进行可视化；



可视化效果如下

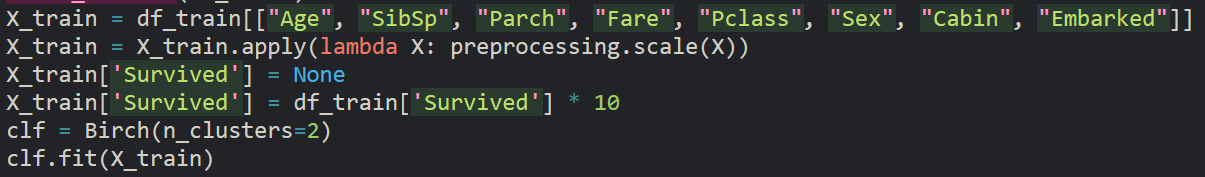


1. 可视化结果解释

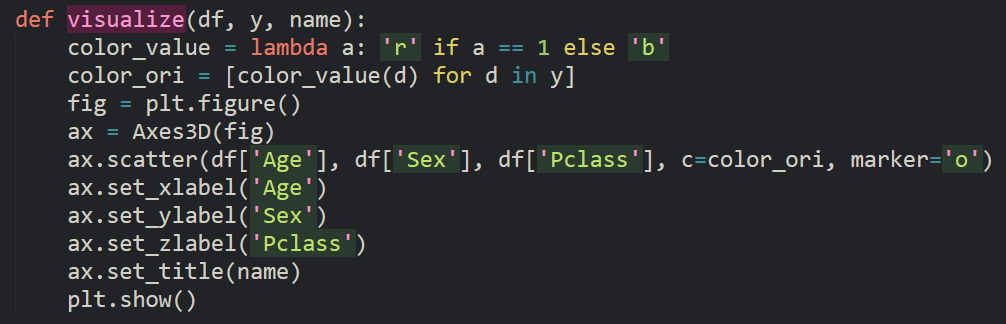
需要先解释一下坐标轴Sex, 0表示女性，1表示男性，从中明显可以看出，女性和男性大致上各聚成一类，说明在泰坦尼克沉没时，让女性先走的的意见成为主流。除此以外，我们还可以看到，船舱等级较高的乘客获救概率也比较大，看来有钱人还是率先得到了救助。在年龄上，年龄与获救的关系不如前面那两个属性这么明显，但也大致可以看出，在10岁以下的孩童获救概率较高。此外，值得注意的时，在图上左下角出现了一个1号仓的女童未获救，与我们的推论相悖，是值得研究与调查的一个点。

3.2 Birch

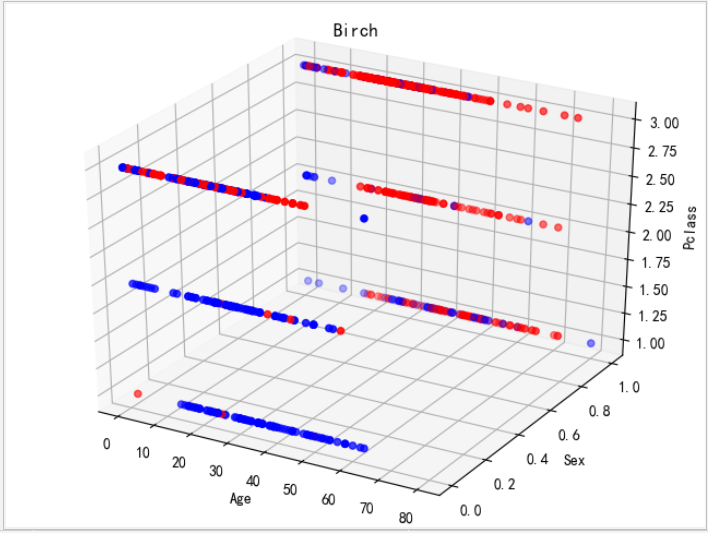
（1）选取数据中用于K-means聚类的数据字段，进行归一化并聚类：



（2）利用Axes3D进行可视化；



可视化效果如下



（3）利用Axes3D进行可视化；

用Birch的聚类结果与Kmeans差不多，也可以得到上面的解释。