**数据挖掘大作业三：分类与聚类**

1 对数据集进行处理

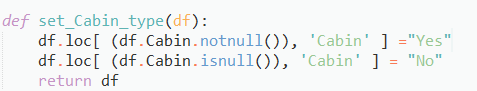
本次作业选用数据集[https://www.kaggle.com/c/titanic/data]，利用Python中的pandas库进行csv数据文件的读取，对数据集进行预处理，以适合分类和聚类算法。

（1）处理缺失数据；

用scikit-learn中的RandomForest来拟合一下'Age','Fare', 'Parch', 'SibSp', 'Pclass'等字段的缺失数据：



按Cabin有无数据，将这个属性处理成Yes和No两种类型：



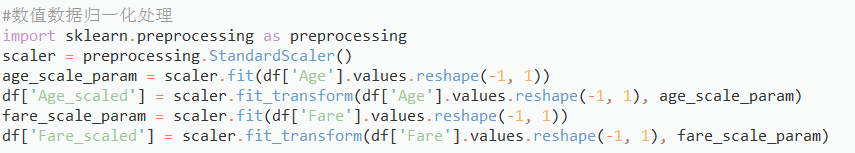
（2）将标称数据转化为数值数据；



（3）对类目型的特征因子化；



（4）数值数据归一化处理；



2 分类算法

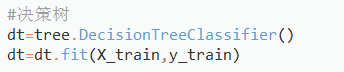
使用Sklearn中train\_test\_split随机划分训练集和测试集，从样本中随机的按比例选取train data和testdata，用于测试分类算法的预测准确率。以下使用决策树和逻辑回归算法进行模型训练和预测。



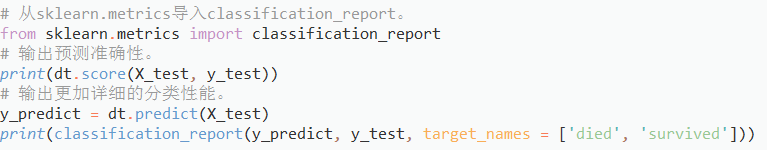
2.1 决策树

（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：

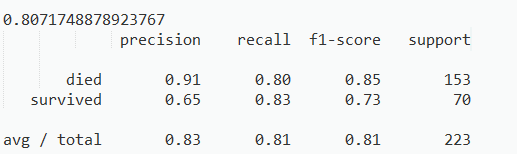




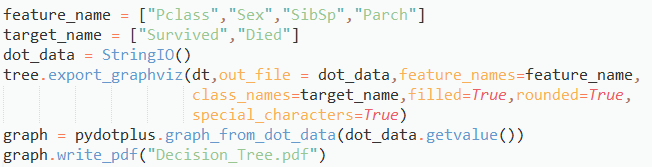
（2）输出预测准确性和详细的分类性能

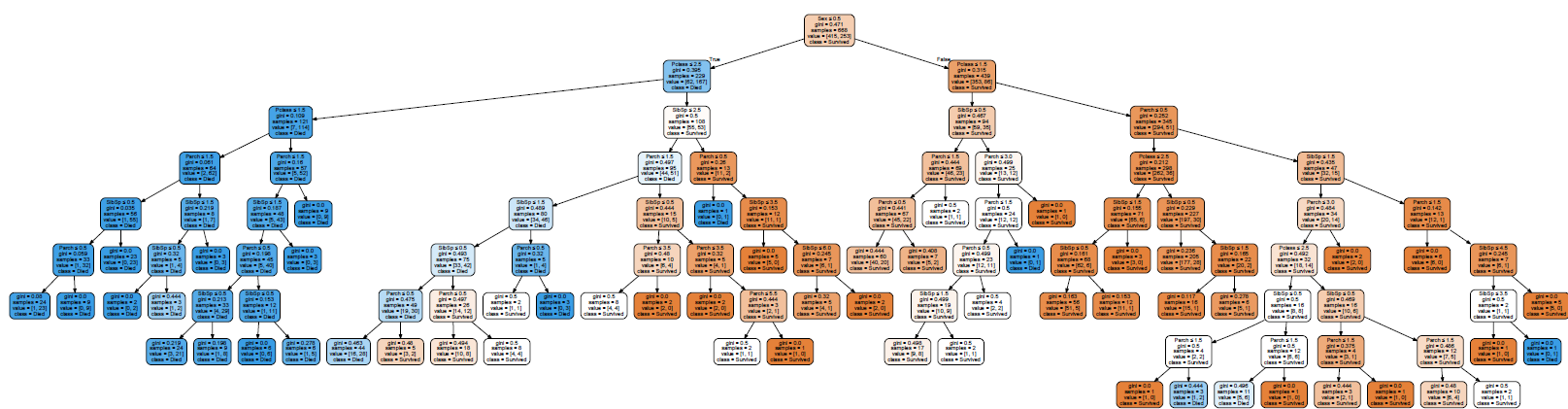


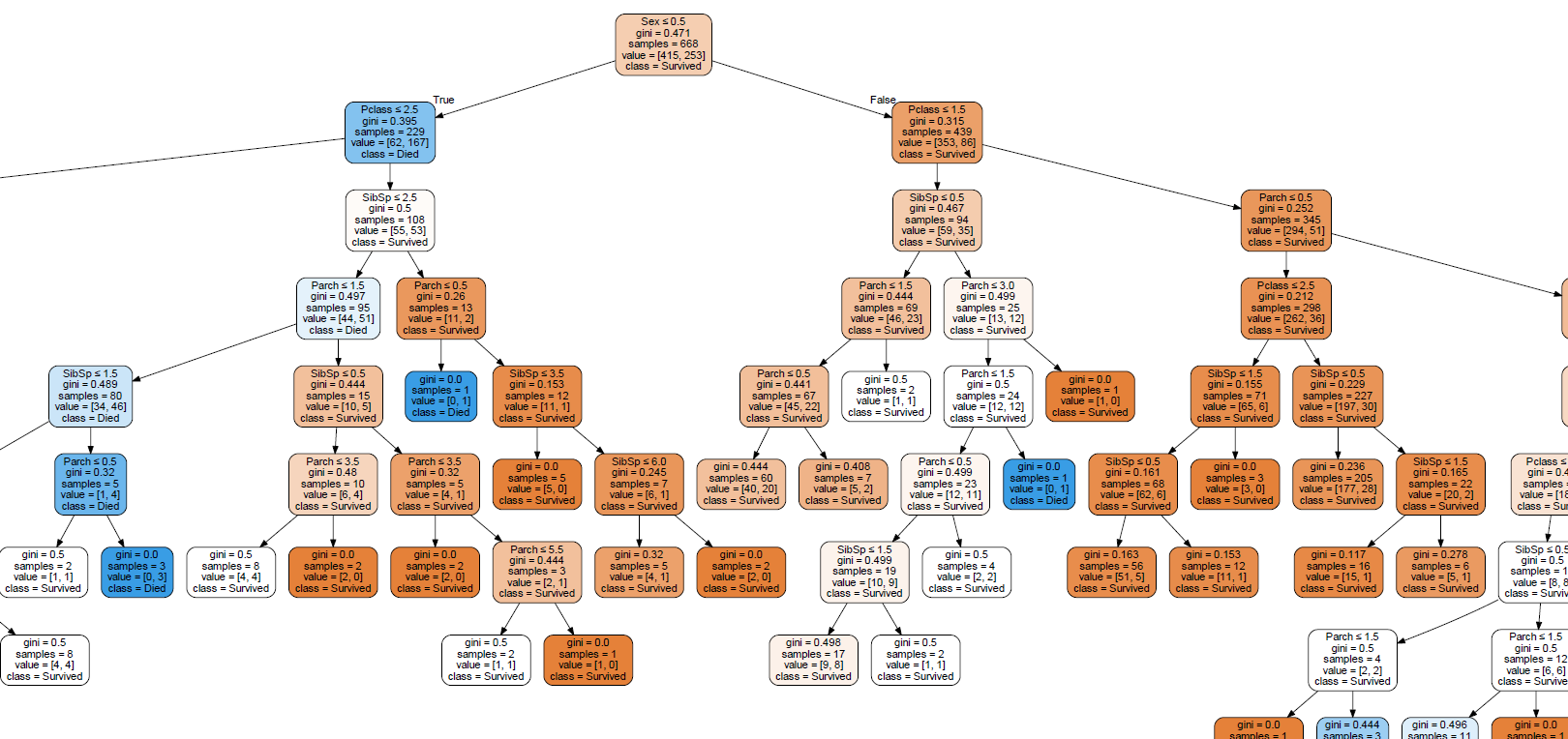
（3）预测准确性和详细的性能结果



（4）决策树可视化代码及效果



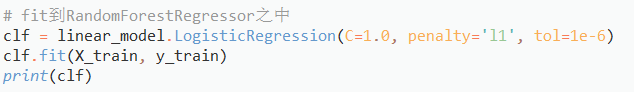




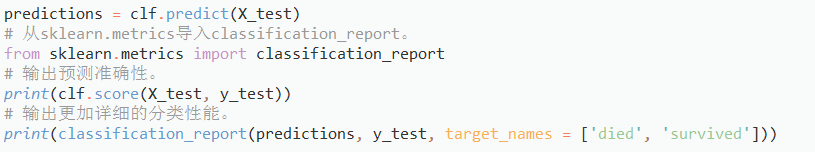
2.2 逻辑回归

（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：

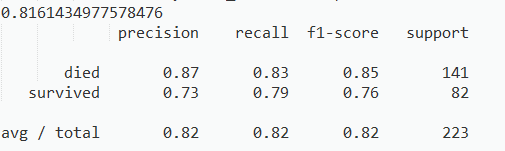




（2）输出预测准确性和详细的分类性能：



（3）预测准确性和详细的性能结果如下



3 聚类算法

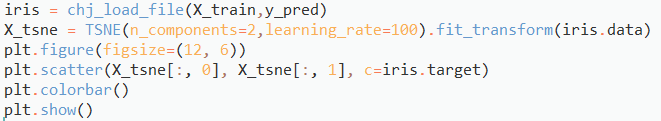
3.1 K-means

（1）选取数据中用于K-means聚类的数据字段，用Sklearn中Kmeans对数据进行K-means聚类：

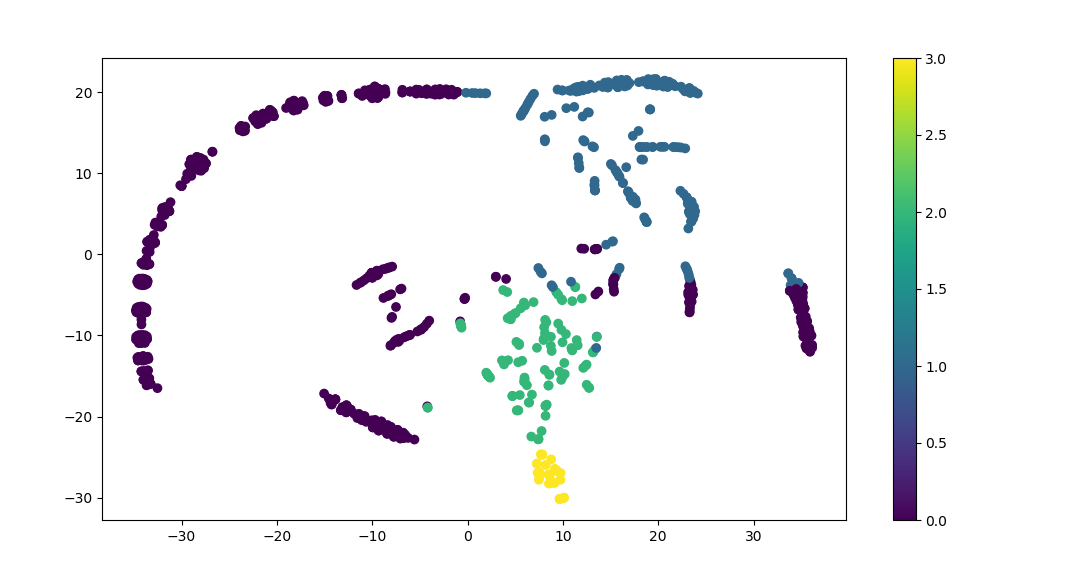




（2）利用TSNE进行降维处理并可视化聚类结果



可视化效果如下



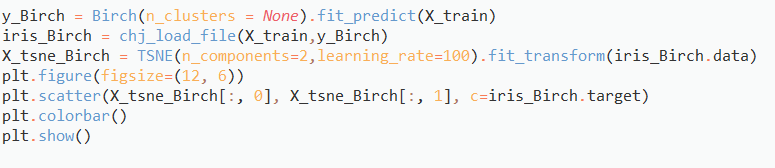
3.2 Birch

（1）选取数据中用于Birch聚类的数据字段，用Sklearn中Birch对数据进行Birch聚类：





（2）利用TSNE进行降维处理并可视化聚类结果



可视化效果如下

