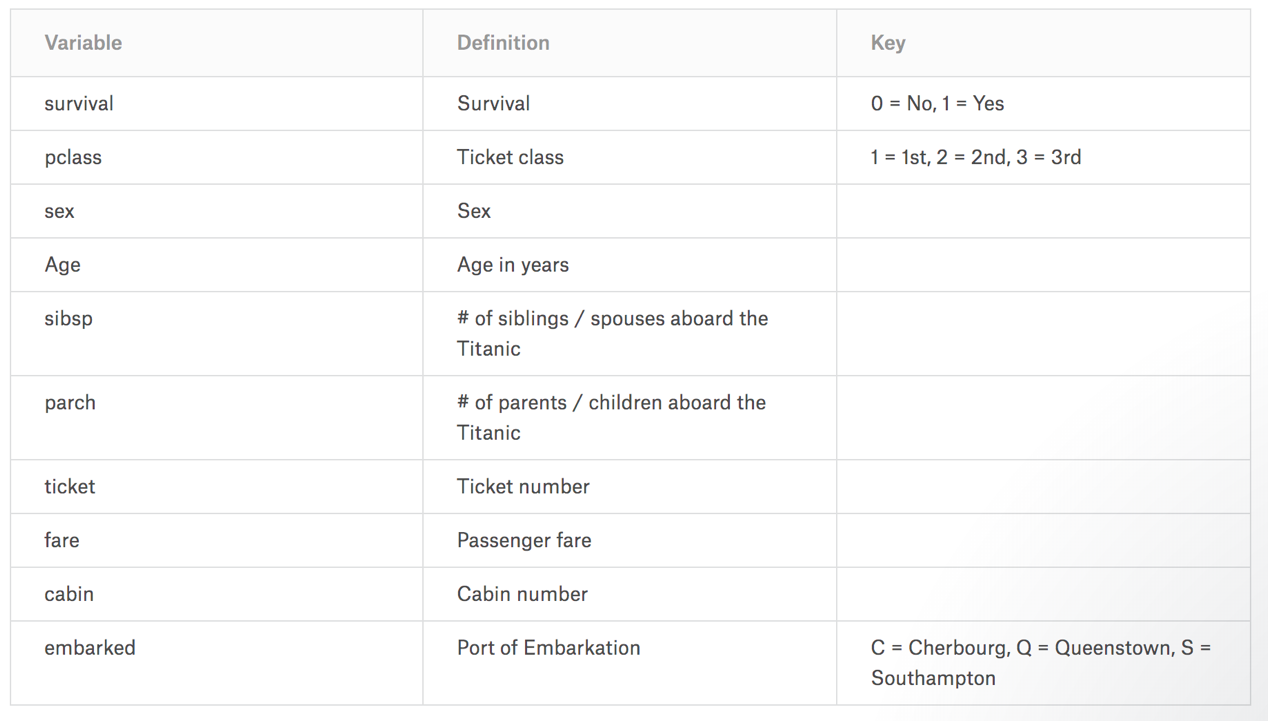
数据挖掘过程报告

# 数据集

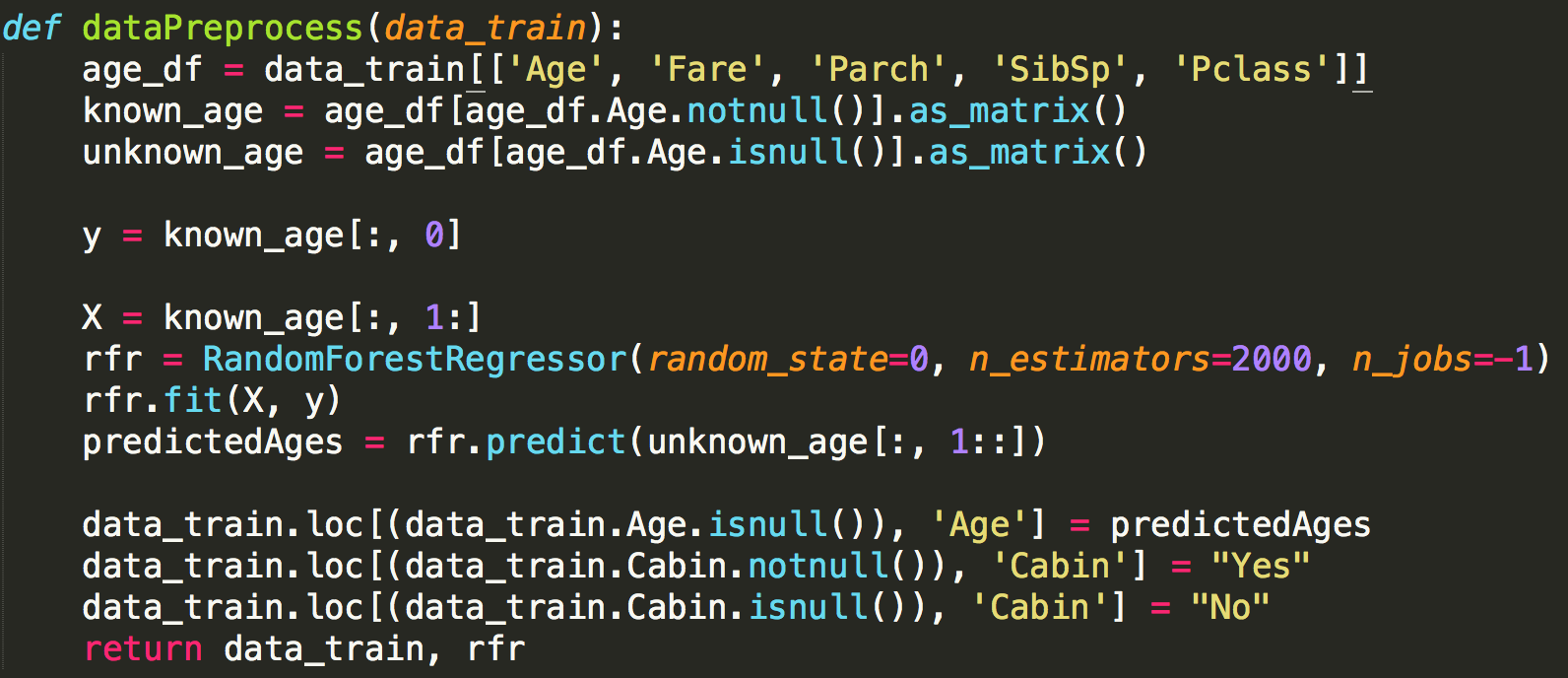
本文中选择了数据集titanic数据集进行数据挖掘，数据集格式为csv文件。该数据集的属性如下图所示：



# 数据预处理

为了便于进行后面的分类过程，需要对数据集进行处理，对缺失数据进行填充。

首先我们用RandomForest拟合字段的缺失数据，并将‘Cabin’属性处理成yes or no，代码如下图所示



除此之外，我们将标称数据转化为数值数据并对类目型的特征因子化，代码如下图所示

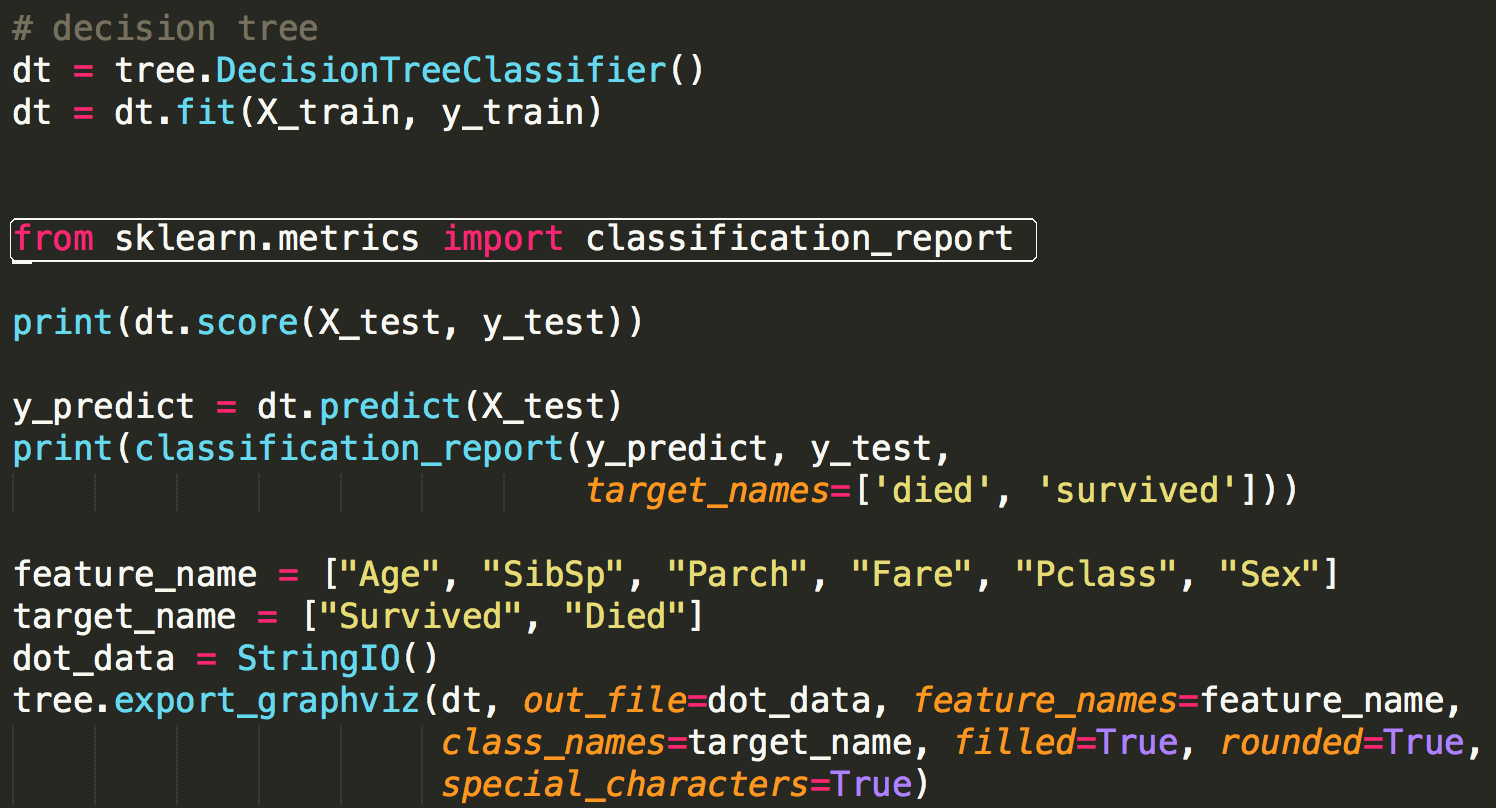


# 数据分类

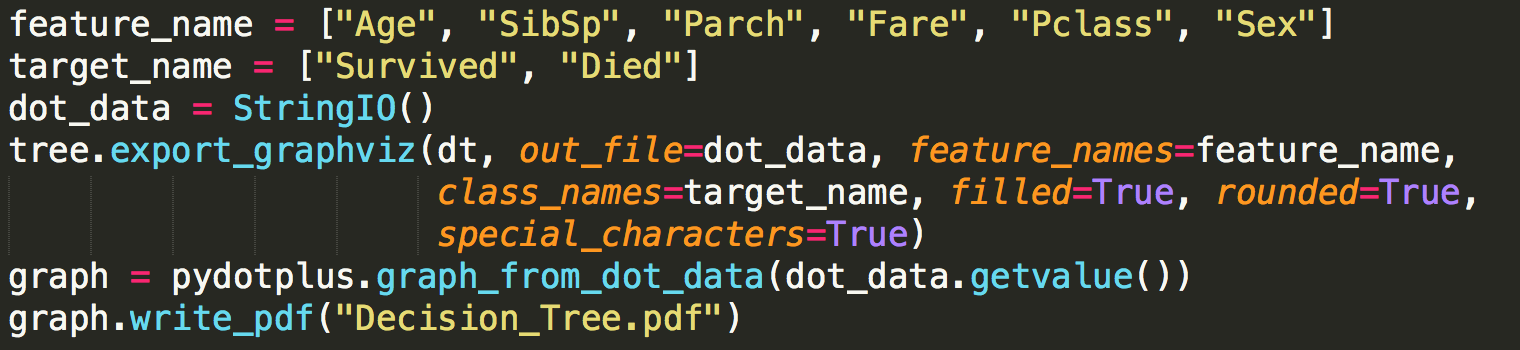
我们使用了两种分类算法，分别是决策树和SVM分类。决策树是一种十分常用的分类方法，是一种监管学习，所谓监管学习就是给定一堆样本，每个样本都有一组属性和一个类别，这些类别是事先确定的，那么通过学习得到一个分类器，这个分类器能够对新出现的对象给出正确的分类。SVM(Support Vector Machine)指的是支持向量机，是常见的一种判别方法。在机器学习领域，是一个有监督的学习模型，通常用来进行模式识别、分类以及回归分析。

**（1）决策树**

决策树的算法实现相关代码如下所示

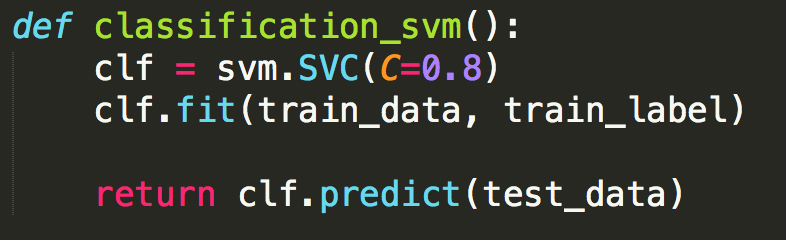


决策树的可视化代码如下图所示



**（2）SVM**

SVM的算法实现相关代码如下图所示

****

SVM分类后的可视化代码如下图所示

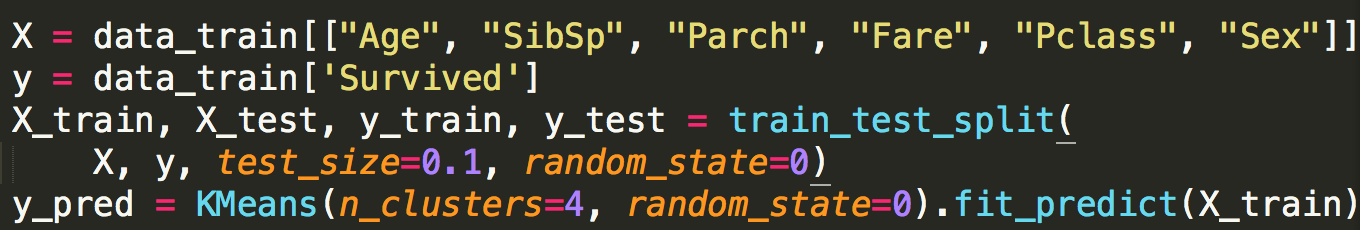
****

# 聚类

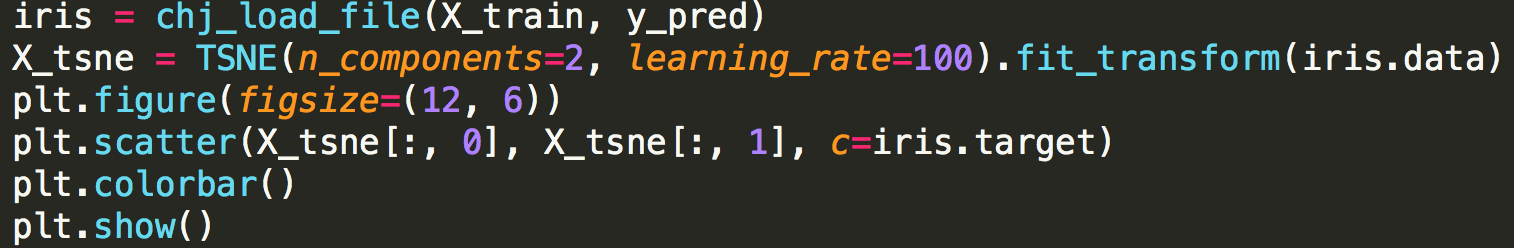
我们使用了两种聚类方法，第一种是K-means算法，另一种是Birch算法。K-means算法以欧式距离作为相似度测度，它是求对应某一初始聚类中心向量V最优分类，使得评价指标J最小。算法采用误差平方和准则函数作为聚类准则函数。BIRCH(Balanced Iterative Reducing and Clustering using Hierarchies)是一个综合的层次聚类算法。它用到了聚类特征(Clustering Feature, CF)和聚类特征树(CF Tree)两个概念，用于概括聚类描述。聚类特征树概括了聚类的有用信息，并且占用空间较元数据集合小得多，可以存放在内存中，从而可以提高算法在大型数据集合上的聚类速度及可伸缩性

**（1）K-means**

K-means的算法实现代码如下图所示

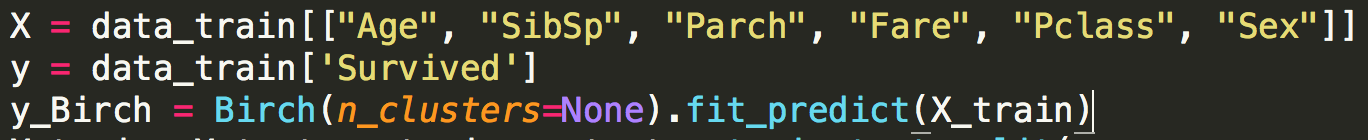


相应的可视化代码如下：

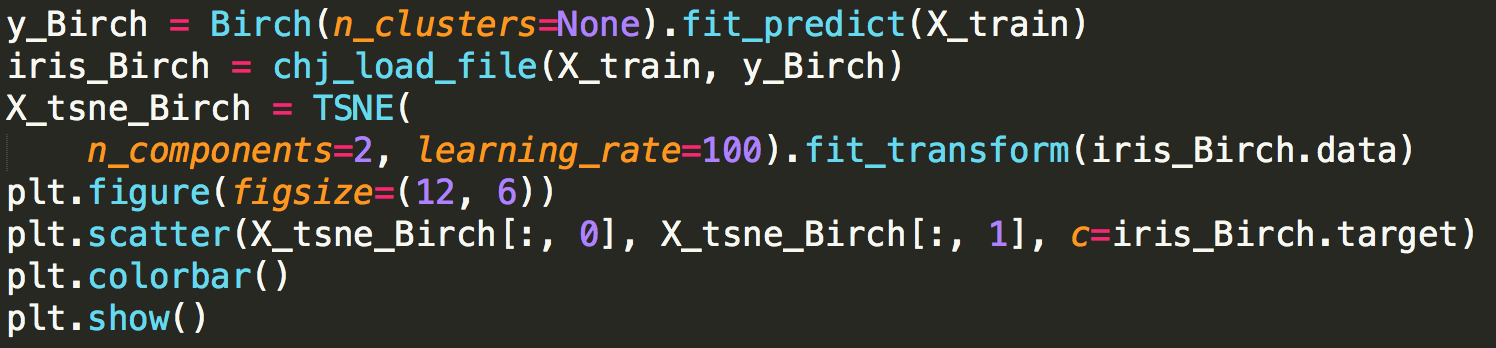


（2）BRICH

BRICH的算法实现代码如下图所示



相应的可视化代码如下：



对应的数据分析见“数据分析.pdf”