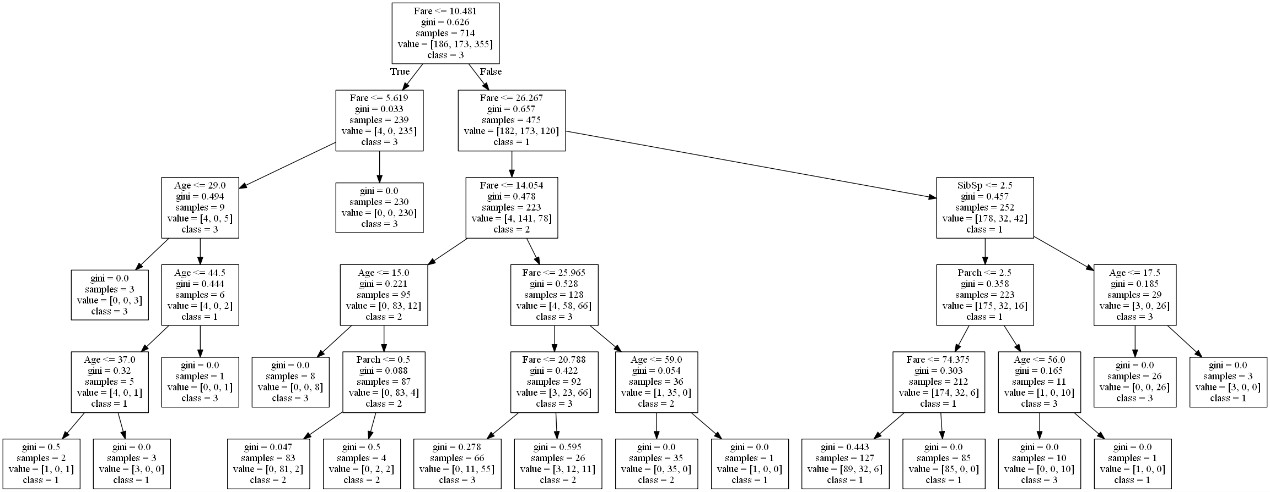
# 分类

采用了决策树模型，将数据集的"Age", "SibSp", "Parch", "Fare"列作为特征列，“Pclass”列作为目标列，分类前剔除了无年龄的行（约100行左右），决策树生成算法为CART，高度限制为5层，获得的决策树以.dot文件的形势输出，再利用Graphviz中的dot工具转化为可视图片，决策树如下。（另生成了一份.pdf文件版本，作为附件tree.pdf一并上传）

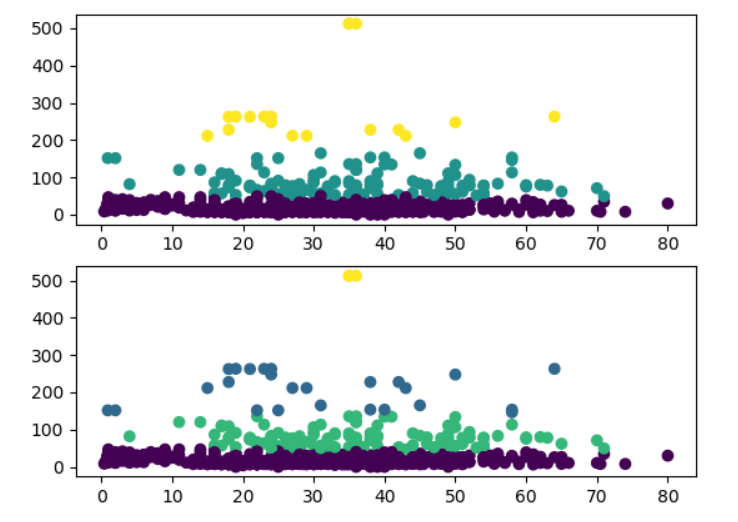


图中可见5层决策树的分类效果已经较好，多数叶节点的基尼系数已经归零，且可看出票价这一参量对于分类结果的影响最大，票价值在5.6与10.481之间的乘客全部坐在了三等舱，票价5.6以下出现了4个例外情况，分别为三名免费乘客和一名仅花了5英镑乘船的乘客。或许是抽奖抽中了船票吧。而票价在26.267以上的乘客则大部分集中在一等舱，少数三等舱乘客票价较为昂贵的，其随行亲属（"SibSp", "Parch"）较多。这与现实情况基本吻合。

# 聚类

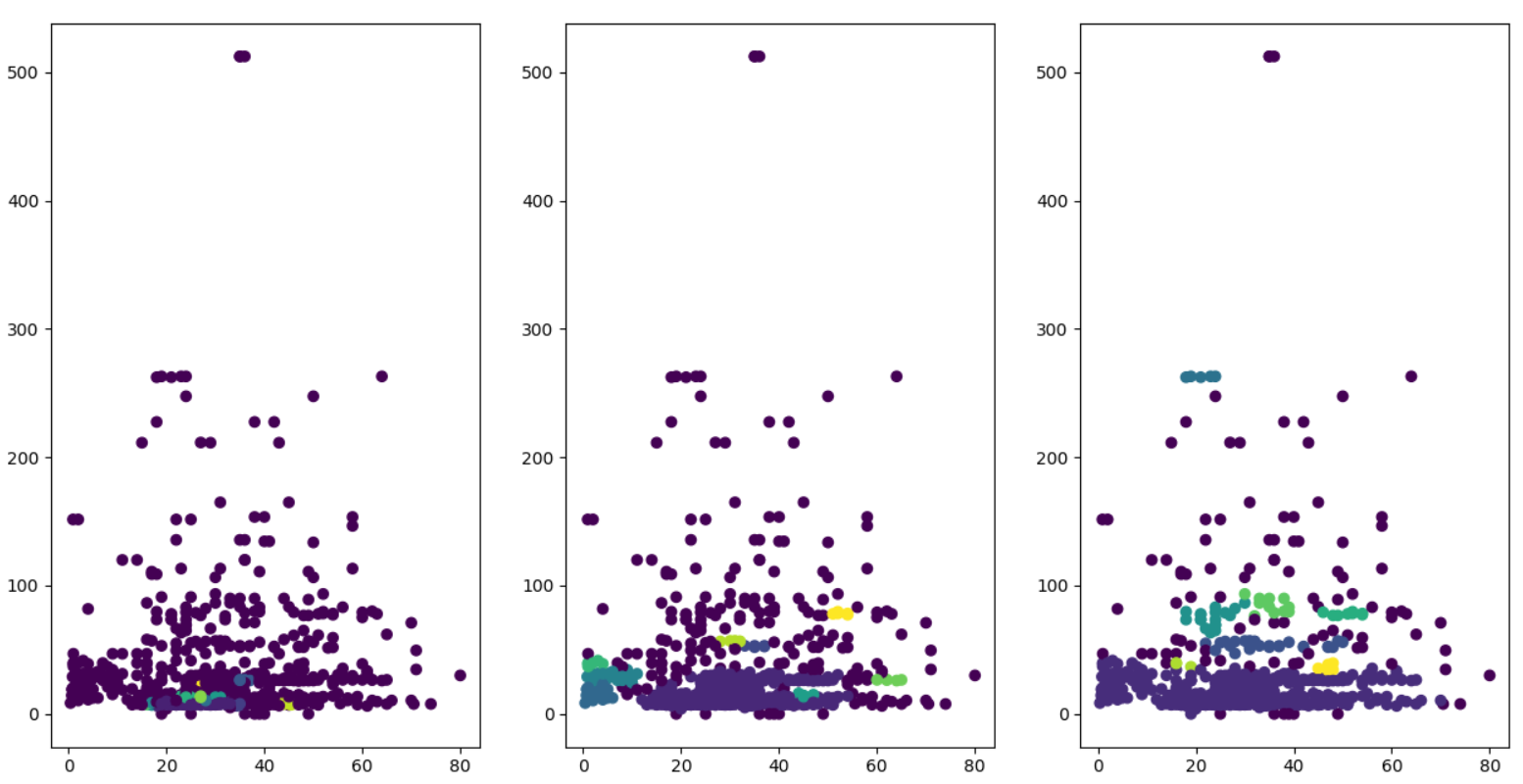
由于此例中只有Fare和Age两个数据为数值属性，所以试探性地对此两数据进行聚类，探究其关系。聚类模型选择了K-Means和DBSCAN两种。聚类前同样剔除了无年龄的行。

## K-Means



上图为k=3和k=4的K-Means聚类结果，可见票价主要分为0-50、50-150和150-300共3个档次，500以上的是极特殊情况。但在k=3聚类时，这两个异常值极大地影响了聚类结果，将第二类和第三类的边界提高了约50，可见K-Means算法对异常值十分敏感。

## DBSCAN



上图为邻域距离阈值eps=1、3、5，周围样本点数量阈值min\_samples=5，距离采用欧氏距离的DBSCAN算法聚类结果，图1中几乎所有的点都是离群点，因为数据的横坐标为年龄，大多数人的年龄都是整数，即两点间距离几乎必然大于1，所以此结果是毫不奇怪的。但是数据集中确实有非整数年龄，正是它们产生了图中显示出的极小的簇。

图2中0-10岁年龄段出现了以票价约20和35为界限的3个簇，而年龄10岁-50岁，票价在35以下的点几乎聚成了一个簇，60岁以上者仅形成了一个小簇，可见乘船者基本以中青年为主，老年人较少。

图3中票价50元以下的点几乎聚集成了一个簇，而更高处的点仅形成了少量小簇，可见大多数乘船者购买的是低价船票，考虑到当时还没有能横跨大西洋飞行的航空器，船只是英美间唯一的交通工具，此结果并不出人意料。